# Myanmar news summarization using different word representations

**Soe Soe Lwin[1], Khin Thandar Nwet[2]**
[1]Natural Language Processing Lab, University of Computer Studies, Myanmar
[2]University of Information Technology, Myanmar

**Article Info**

***Article history:***

Received Feb 25, 2020
Revised Jul 29, 2020
Accepted Aug 9, 2020

***Keywords:***

Bag-of-words
Centroid
Myanmar news
Text summarization
Word embedding

**ABSTRACT**

There is enormous amount information available in different forms of sources and genres. In order to extract useful information from a massive amount of data, automatic mechanism is required. The text summarization systems assist with content reduction keeping the important information and filtering the non-important parts of the text. Good document representation is really important in text summarization to get relevant information. Bag-of-words cannot give word similarity on syntactic and semantic relationship. Word embedding can give good document representation to capture and encode the semantic relation between words. Therefore, centroid based on word embedding representation is employed in this paper. Myanmar news summarization based on different word embedding is proposed. In this paper, Myanmar local and international news are summarized using centroid-based word embedding summarizer using the effectiveness of word representation approach, word embedding. Experiments were done on Myanmar local and international news dataset using different word embedding models and the results are compared with performance of bag-of-words summarization. Centroid summarization using word embedding performs comprehensively better than centroid summarization using bag-of-words.

***Corresponding Author:***

Soe Soe Lwin
Natural Language Processing Lab
University of Computer Studies, Yangon
No.4 Main Road, Shwe Pyi Thar Township, Yangon, Myanmar
Email: soesoelwin@ucsy.edu.mm

## 1. INTRODUCTION

Nowadays, information is increasing gradually on the internet and it is necessary to compress different types of data. Summarization made by human is very time consuming and tedious. Therefore, automatic text summarization is essential to overcome the problem. Text summarization is a technique for extracting essential information from original text document as a shortened form.

Automatic text summarization systems can be generally classified into two different types. These two main approaches are: extractive and abstractive [1]. Extractive summarization is formed by extracting phrases or sentences from the document to form summary. The main goal of extractive summarization is to produce summary without redundancy and give important point of souce document. Abstractive summarization uses new words to form the summary to describe the main content. Extractive summarization method generally has three steps:

- Intermediate representation model
- Scoring the sentences based on the representation and

- Selection of a summary comprising of a number of sentences

For the first step, there are two types to represent the input text: topic representation approaches (centroid based method, latent semantic analysis, Discourse based method, Bayesian topic models) [2-4] and indicator representation approaches (graph-based method, machine learning) [5-6]. When the intermediate representation is generated, an importance score is assigned to each sentence in second step. Finally, the summarizer system selects the most important sentences to produce a summary.

Previous Myanmar text summarization system uses machine learning approach, CRF, takes information extraction as sequence labeling task. Machine learning approach needs large training dataset, so that the system performance is better. Kyaw [7] used CRF (conditional random field) for word segmentation and information extraction. Seven types of Myanmar natural disasters news (flood, landslide, earthquake, forest fire, storm, volcanic eruption, tornadoa) are summarized by using template driven text summarization approach. That system does not consider other features such as POS that make information extraction task better. That system does not consider anaphora resolution. In order to improve performance, more data should be collected to get greater training corpus. Kyaw [8] described query-focused multi-document summarization of Earthquake news in Myanmar. That system only describes word level summary of Myanmar news by using forward-backward algorithm. Longest matching approach is used to reduce redundant information. Our system is aimed to produce sentence level summary instead of word level summary.

Text summarization using semantic role labeling was proposed by Naing [9]. In that paper, anaphora resolution and semantic role labeling algorithm for Myanmar language was proposed. And then Myanmar verb frame resource based on PropBank semantic resources was built using Myanmar-English Computational lexicon and Lexique Pro lexicon. Finally, both entity and its reference in the sentence are chosen for summary generation. Verbs in spoken sentences are not identified.

According of previous Myanmar text summarization research, word level summary is only produced. The main goal of this paper is to produce sentence level summary. Although there are other language text summarizers like Thai, Korean, Chinese, Myanmar language summarizer are rare now. According to these problems, Myanmar text summarization system is definitely necessary to develop in the area of Myanmar NLP research.

For representing text, bag of words models are used. Although it is easy to implement, it has some limitations such as it ignores the semantic of words in the document. For calculating scores for sentences and ranking sentences for generating important sentence, bag of words representation is mostly used. But there are limitations in bag of words model. Vector representation in bag of words model cannot capture semantic relationships of words if they have no words in common. To overcome the limitation of bag of words representation model, word embedding is used in this paper. Word embedding provides a better vector feature on most of natural language processing problem. Word embedding is a dense vector representation and can capture syntactic and semantic information of a word. This paper describes the comparison of centroid summarizer by taking advantages of different word embedding model and baseline bag-of-words model [10].

## 2. MYANMAR LANGUAGE

The Myanmar language is a Sino-Tibetan language that is spoken in Myanmar and it is an official language. Burmese script is adapted from Mon script, which is derived from Pali, the language of Theravada Buddhism. It is tonal language. It is written horizontally from left to right and Myanmar language consists of 33 consonants and 14 vowels. The word order of Myanmar language is Subject-Object-Verb.

Myanmar language is an under resourced language and there are no text summarization datasets like other languages. There are Thai, Japanese, and English text summarization systems. However, Myanmar text summarization systems are very few and rare. Myanmar text summarization using conditional random fields (CRF) and Myanmar verb frame based summarization system have been proposed. Nevertheless, Myanmar summarization system using the effectiveness of word embedding have not been proposed yet. Therefore, Text summarization using efficient of word embedding is proposed in this paper.

## 3. WORD EMBEDDING

Word embedding is an NLP technique, which can capture the meaning of a word in a document, semantic and syntactic similarity, also the relation with other words. These are vector representations of a particular word. The main purpose of Word2vec is to group the vectors of similar words together in vectorspace [11]. There are two architectures in Word2vec. These are continuous bag of words (CBOW) and skip-gram. CBOW predicts a target word based on its surrounding words. Skip-Gram predicts the surrounding words based on target word [12].

There are many word representation models: Word2Vec and Glove. But Word2vec and GloVe fail to provide any vector representation of words that are not in the corpus. FastText works well with rare words even if a word wasn't seen during training. Therefore, fastText and BPEmb embedding are used in this paper. FastText is a free library for learning word representation and sentence classification [13]. There are publicly available pretrained word vectors for many languages. BPEmb (byte-pair embeddings) is a collection of pre-trained subword embeddings based on byte-pair encoding (BPE) and trained on Wikipedia [14]. It is used as input for neural models in natural language processing. Subwords allow guessing the meaning of unknown words. Byte-Pair encoding gives a segmentation without requiring tokenization or morphological analysis. The vocabulary size is the sum of the number of BPE merge operations and the number of characters in the training data.

## 4. DESIGN OF PROPOSED SYSTEM

Sentences and documents are represented in some feature vector space. Centroid can be defined as the whole document's vector [15]. Summary sentences are selected by selecting sentences which have vectors similar to centroid vector. It can be done by using different representations. Original centroid approach uses bag of words as representation model for sentence scoring and selection. Bags of word model cannot grasp semantic relationship between sentences. Therefore, Rossiello [16] proposed centroid summarization by utilizing word embedding. They used word embedding to represent sentence and words. The centroid embedding is calculated as the sum of word embeddings of most important words, and sentence embeddings are calculated as sum of word embeddings they contain. Figure 1 depicts the design of the proposed system. The proposed system has four main steps;

- Preprocessing step
- Computing centroid embedding
- Computing sentence embedding
- Producing of summary sentence

Firstly, input news documents are segmented into words and stopwords are removed in the preprocessing step. And then, Centroid embedding and sentence embedding are computed by using lookup table. Cosine similarity scores between the sentence and centroid embedding are computed and sentences are ranked by descending order according to similarity scores. Finally, final summary is produced. This paper is proposed to utilize the different types of word embedding for representing sentences. Quality of representation sentences can affect the performance of centroid based summarization. Therefore, centroid based on different type of word embedding is proposed in this paper.

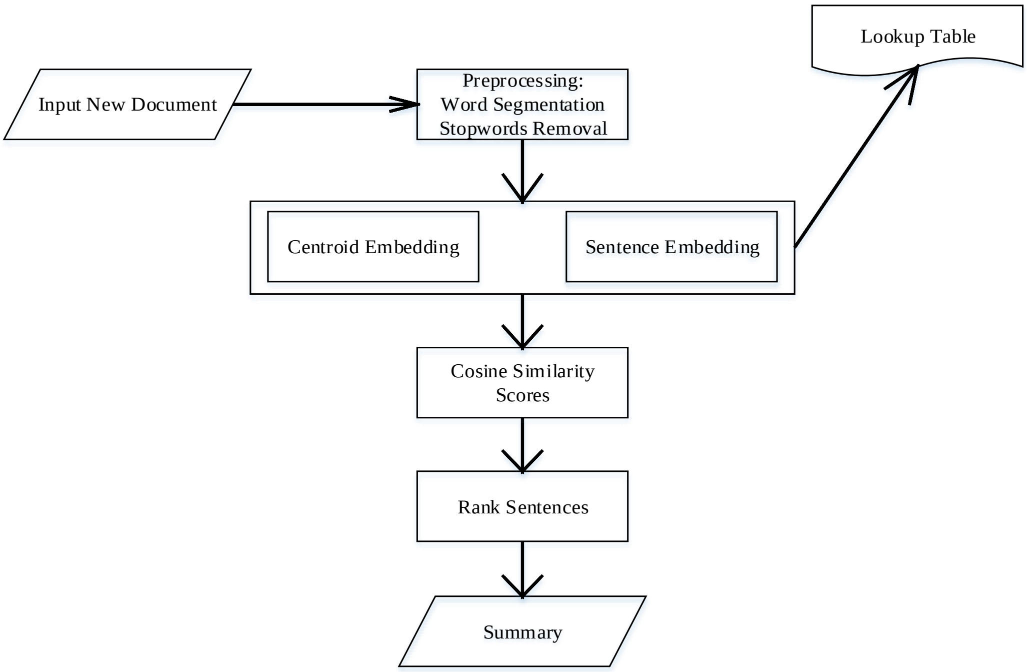

Figure 1. Proposed system for Myanmar summarization

### 4.1. Preprocessing

In English, word boundaries can be easily determined because there are spaces between words. In Myanmar language, words are written without space and sometimes spaces are used between phrases but there are no defined rules for using space. Therefore, word segmentation is a challenging task in natural langue processing. In this paper, Myanmar news are segmented by Myanmar word segmenter by [17]. A combined model, bigram and word juncture is used in that system. This segmenter uses longest matching and bigram method with a pre-segmented corpus of 50,000 words which are collected manually from Myanmar text books, newspapers, and journals. The corpus is in unicode encoding. In text summarization, stopwords are defined as irrelevant information. Therefore, Myanmar stopwords (prepositions and conjunctions) are removed from news document.

### 4.2. Centroid embedding

Words that have term frequency-inverse document frequency (TF-IDF) weight greater than a topic threshold are selected to compute centroid embedding. The vectors of the top ranked words in the document using lookup table a are summed to get centroid embedding. A corpus has documents [D1, D2, D3…] and it has vocabulary V with size N=|V|. Let define a matrix XϵRN,K, lookup table, where i-th row is a word embedding of size k,k<<N, of the i-th word in V. The value of matrix X are learned by using, word embedding model like FastText and BPEMP. Centroid embedding and sentence embedding are computed by using lookup table. Centroid embedding can be computed by (1):

$$C = \sum_{w\epsilon D, tfidf(w)>t} X[idx(w)] \tag{1}$$

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

### 4.3. Sentence embedding

Vectors of each word in sentence in lookup table X are summing up to get sentence embedding. Sentence scores are calculated by (2):

$$S_j = \sum_{w\epsilon S_j} X[idx(w)] \tag{2}$$

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 centroid C of the document D is computed as (3) to get sentence score.

$$Sim(C, S_j) = \frac{C^T.S_j}{||C||.||S_j||} \tag{3}$$

As shown in (3) shows cosine similarity which measures how close two vectors C and $S_j$ are based on their angle.

### 4.4. Sentence selection

The sentences are sorted in descending order according to their similarity scores. The top score sentences are selected as summary until the predefined limit (number of words). In this paper, the number of words for generating summary is 150 words.

## 5. EXPERIMENTAL RESULTS AND ANALYSIS

### 5.1. Data set

Unlike the other language, there are no datasets for text summarization in Myanmar language. Therefore, daily update news from Myanmar news websites [18-21] are manually collected to build Myanmar news dataset. The data is unicode encoding. In Myanmar, unicode font is not very familiar and Myanmar people mostly use Zawgyi font. But unicode encoding is used as standard encoding in Myanmar natural language processing tasks. Now, these manually collected dataset consists of 2k news articles. The average length of each article is 5 sentences. There is no gold (reference) summary for datasets like English summarization dataset. Therefore, gold summary is generated by 10 human evaluators. The centroid summarizer is experimentally evaluated on Myanmar local and international news datasets. Centroid based on two different word embedding model (BPEmb, FastText) are used for sentence representation. BPEmb pretrained embedding for Burmese (Myanmar) language have been published recently [22]. In this paper, BPEmb, pretrained embeddings for Myanmar language with vocabulary sizes (200000) and dimension (300) are used. FastText, pre-trained word vectors learned on Myanmar Wikipedia articles are used in system. This model was trained using skip-gram with dimension 300 [23, 24]. In this system, these word embedding models are applied for sentence representation.

Table 1 shows the most relevance sentences of sample Myanmar news article by using centroid approach combined with word embedding approach. The first column shows sentence ID, and the second column shows sentences. The cosine similarity scores are shown in third column. Centroid words are selected by using TF-IDF value greater than a topic threshold. In the following article, centroid words are ကျန်းမာရေး(health) ကျောင်းသား(students) စောင့်ကြပ် (monitor) ဆေးရုံ (hospital) ရောဂါ (disease). The most important and relevant sentence ID 3 which contains many words that are closed to centroid vector. The centroid words in the sentences are marked in bold. The summary is provided until the limit (number of words) is reached.

Table 1. Sentences score by computing cosine similarity between centroid and sentence embedding

| Sentence ID | Sentence | Score |
|---|---|---|
| 3 | " ကျောင်းသား ကျောင်း သူတွေ ကို နေ့စဉ် သာ မာန် အားဖြင့် အပူချိန် တိုင်း တဲ့ လုပ်ငန်း တစ်ခု ကို တိတိကျကျ လုပ်ဆောင် ပေး ပါတယ် ။ ဒီ လုပ်ငန်း ကို သူနာပြု ဆရာ မက PPE ဝတ်စုံ ကို အပြည့်အဝ ဝတ်ဆင် ပြီးမှ ဝင်ရောက် တိုင်းတာ ပါတယ် ။ သူတို့ မှာ ချောင်းဆိုးတာ ၊ ကိုယ် ပူတာ ၊ ဖျား တာ တစ်ခုခု မ ဖြစ် မ ချင်း ဘာမှ မလုပ်ပါ ဘူး ။ အကယ်၍ ဒီ ကျောင်းသားတွေ ထဲမှာ သံသယ လူနာ ရှိ လာ မယ် ဆိုရင် ကျောင်းသားတွေ ကို ဝူဟန် ကို သွား ခေါ် တုန်းက လိုက်ပါ သွားပြီး ဒီ ဆေးရုံ မှာ ကိုယ်တိုင် ကျန်းမာရေး စောင့်ကြပ် ကြည့်ရှု မှု ခံယူ နေတဲ့ သမားတော် တစ်ယောက် ရှိပါတယ် ။ ဒီ သမားတော် နဲ့ ညှိနှိုင်း ပြီး သူ က အရင် ဆုံး စစ်ဆေး မှာပါ " ဟု ဒေါက်တာ ခင်ကျော် က ပြောကြားသည် ။<br>“The students have showed no symptoms of the coronavirus infection yet. If one of them shows symptoms such as fever with high temperature, the patient will be checked by a physician who also participated in the mission to bring them back from Wuhan to Myanmar,” said Dr. Khin Kyaw. | 0.96891 |
| 1 | " ဒီ ကန်တော်နဒီ ဆေးရုံ မှာ ဝူဟန် မြို့က ပြန်လည် ခေါ်ဆောင် လာ တဲ့ ကျောင်းသား ၁၆ ဦး ၊ ကျောင်းသူ ၄၂ ဦး နဲ့ သုံးနှစ် အရွယ် မိန်းကလေး တစ်ဦး စုစုပေါင်း ၅၉ ဦး ရှိပါတယ် ။ သူတို့ အားလုံး ရဲ့ ကျန်းမာရေး အခြေအနေ က ဒီနေ့ အထိ ကောင်းမွန် ပါတယ် ။ တစ်ယောက် မှ အဖျား တတ် ခြင်း မ ရှိတဲ့ အပြင် အခြား အဆုတ် လမ်းကြောင်း ဆိုင်ရာ ရောဂါ လက္ခဏာ တွေ လည်း မ တွေ့ ရှိ ရ ပါ ဘူး ။ ဒါ့အပြင် ဒီ ကျောင်းသားတွေ ကို မြို့ က သွား ခေါ် တဲ့အခါမှာ လိုက်ပါ သွားခဲ့ တဲ့ ကျန်းမာရေး ဝန်ထမ်း သုံးဦး ကိုလည်း ဒီ ဆေးရုံ မှာပဲ ကျောင်းသားတွေ နဲ့အတူ ထား ရှိပြီး ကျန်းမာရေး စောင့်ကြပ် ကြည့်ရှု မှု တွေ လုပ် ပေး နေ ပါတယ် ။ သူတို့ သုံးယောက် ရဲ့ ကျန်းမာရေး အခြေအနေ ကလည်း ကောင်းမွန် ပါတယ် ။ ယေဘူယျ အားဖြင့် တော့ သူတို့ကို ပြန် ခေါ် လာ တဲ့အခါ သူတို့ နဲ့အတူ ရောဂါပိုး ပါ လာနိုင် တဲ့အတွက် သီးသန့် ခွဲခြား ထား တာ ပဲ ရှိပါတယ် ။ သာမာန် အားဖြင့် သူတို့ကို ရောဂါ ရှိတယ် လို့ သတ်မှတ် လို့ မရပါဘူး " ဟု ၎င်း က ပြောသည် ။<br>“A total of 59 people were evacuated from Wuhan-16 male students, 42 female students, and a three-year-old girl and they all are well on the third day of quarantine, with no fever and no respiratory symptoms,” and "the condition of the three health workers who took part in the mission to bring the students back is also being monitored, and they are also in good health, and quarantine does not mean that they have been infected with the coronavirus and it is being imposed to prevent the possibility of an infection,” he added. | 0.96891 |
| 0 | မန္တလေးမြို့ ကန်တော်နဒီ ဆေးရုံ တွင် သီးသန့် ထား ရှိသည့် တရုတ်နိုင်ငံ ဝူဟန် မြို့ မှ ပြန်လည် ခေါ်ဆောင် လာ သည့် ကျောင်းသား ကျောင်းသူ ၅၉ ဦး ၏ ကျန်းမာရေး အခြေအနေ မှာ ယနေ့ အထိ စောင့်ကြပ် ကြည့်ရှုစစ်ဆေး မှု များ အရ ကောင်းမွန် လျက် ရှိကြောင်း အဆိုပါ ကျောင်းသားကျောင်းသူများ ၏ ကျန်းမာရေး စောင့်ကြပ် ကြည့်ရှု ခြင်း လုပ်ငန်းများ ကို တာဝန် ယူ ဆောင်ရွက်ပေးနေ သည့် မန္တလေး အထွေထွေ ရောဂါ ကု ဆေးရုံကြီး မှ ဆေးရုံအုပ် ဒုတိယ ညွှန်ကြားရေးမှူး ဒေါက်တာ ခင်ကျော် က ပြောကြားသည် ။ The 59 Myanmar students brought back from Wuhan were in good health on the third day of their 14-day quarantine yesterday at the Kandawnadi Hospital in Mandalay, according to Deputy director Dr Khin Kyaw, Medical Superintendent of the Mandalay General Hospital, who is currently responsible for monitoring the students’ health. | 0.96731 |
| 4 | " မန္တလေး ပြည်သူလူထု အနေနဲ့ အများကြီး မစိုးရိမ် ကြ ဖို့ ပြော ချင် ပါတယ် ။ စိုးရိမ် တာ ကို နားလည် ပေ မဲ့ ပိုပြီး အ စိုးရိမ် လွန်ကဲ ပြ ရင် အများပြည်သူ အထိတ်တလန့် ဖြစ် မှာ ကို မ လိုလား ပါ ဘူး ။" ဟု ၎င်း က ဆက်လက် ပြောကြားသည် ။<br>“I would like to tell the people of Mandalay not to be very anxious about the disease. Too much anxiety can cause panic among people,” he added. | 0.95220 |

### 5.2. Evaluation and analysis

One of the metrics for evaluation of text summarization is recall-oriented understudy for gisty evaluation (ROUGE) [25]. It works as comparing human summary (one or several reference summaries) and system summary based on n-grams. Precision in ROUGE means that how much of the system summary was relevant? Precision can be computed as (4):

$$\frac{number\_of\_overlapping\ words}{total\_words\_in\_system\_summary} \tag{4}$$

Recall is computed as (5). Recall in ROUGE simply means how much of the reference summary is the system summary recovering or capturing.

$$\frac{number\_of\_overlapping\ words}{total\_words\_in\_reference\_summary} \tag{5}$$

To find the best parameter value, different parameter configurations are tested with original centroid by using bag-of-words representation and centroid by taking advantages of different word embedding models. Parameter setting are topic threshold (topic-t) in [0, 0.5] and similarity threshold (sim-t) in [0.5, 1] respectively. Two pre-trained word vectors, (FastText trained on Wikipedia, and BPEmb trained on Wikipedia) are used for centroid and sentence embedding. Figures 2-4 show the evaluation results of original centroid by using bag of words representation and centroid with different word embedding models. In Figure 2, topic threshold 0.4 and similarity threshold 0.98 is set and rouge 2 scores of centroids based on different word embedding models performed better than the baseline original centroid based on bag-of-words summarizer.

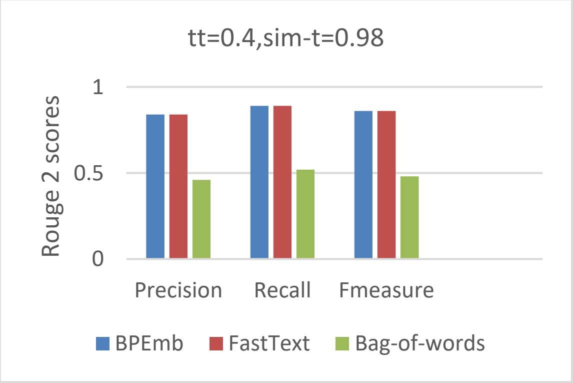

Figure 2. Rouge2 scores of Centroid summarizer based embedding models (*BPEmb*, and *FastText*) and bag-of-words model with parameter values [tt=0.4, sim-t=0.98]

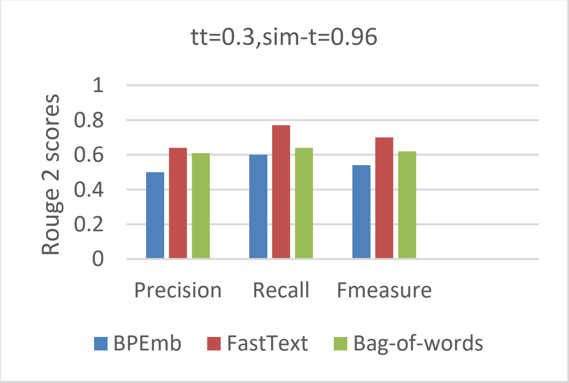

Figure 3. Rouge2 scores of Centroid summarizer based embedding models (BPEmb, and FastText) and bag-of-words model with parameter values [tt=0.3, sim-t=0.96]

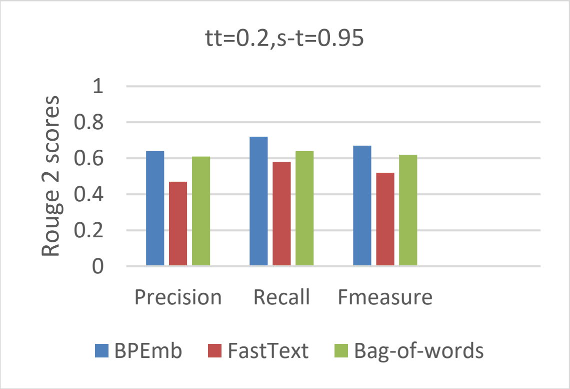

Figure 4. Rouge2 scores of Centroid summarizer based embedding models (BPEmb, and FastText) and bag-of-words model with parameter values [tt=0.2, sim-t=0.95]

Where tt is topic threshold and sim-t is similarity threshold. *BPEmb* is pre-trained word vectors learned on Wikipedia sources, and *FastText* is FastText pretrained word vectors learned on Wikipedia sources.

In Figure 3, topic threshold 0.3 and similarity threshold 0.96 is set and the results assert that rouge 2 score of centroid summarizer based on FastText pretrained word vectors trained on Wikipedia performed better than all other representation schemes. Centroid summarizer based on BPEmb, pre-trained word vectors trained on Wikipedia is slightly less than the baseline bag-of-words models. As shown in Figure 4, topic threshold 0.2 and similarity threshold 0.95 is set and rouge 2 scores of centroid summarizer based BPEmb embedding model achieve the best rouge scores. Bag of word model gets higher rouges scores than *FastText*. A possible reason is that word embedding needs higher threshold because it requires accurate choice of meaningful words to compose the centroid vector.

## 6. CONCLUSION

In this paper, the relevance sentences are extracted by choosing sentences which have vectors similar to centroid vector. Word embedding is utilized to capture contexts of the words with dense representations in the form of numeric vectors. In this paper, word embedding is applied to sum the word vectors from the trained word embeddings to form sentence and document embeddings. The experimental results on Myanmar news data show that centroid summarizer based embedding model improve the performance of bag of words model. Word embedding which is better on what words is similar on syntactic and semantic relationship rather than bag-of-words (BOW). This centroid method will be applied in more complex summarization tasks such as multi-document summarization, query focused summarization in future work.

## REFERENCES

[1] N. Munot, and S. Govilkar, “Comparative Study of Text Summarization Methods,” *International Journal of Computer Applications*, vol. 102, no. 12, pp. 33-37, 2014.

[2] D. G. Ghalandari, “Revisiting the Centroid-based Method: A Strong Baseline for Multi-Document Summarization,” *Proceedings of the Workshop on New Frontiers in Summarization*, 2017, pp. 85-90.

[3] S. S. Lwin and K. T. Nwet, “Extractive Summarization for Myanmar Language,” *2018 International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, Pattaya, Thailand, 2018, pp. 1-6.

[4] G. Yang, "A Novel Contextual Topic Model for Query-Focused Multi-document Summarization," *2014 IEEE 26th International Conference on Tools with Artificial Intelligence*, Limassol, 2014, pp. 576-583.

[5] K. S. Thakkar, R. V. Dharaskar and M. B. Chandak, "Graph-Based Algorithms for Text Summarization," *2010 3rd International Conference on Emerging Trends in Engineering and Technology*, Goa, 2010, pp. 516-519.

[6] J. N. Neto, et al., “Automatic Text Summarization Using a Machine Learning Approach,” *Lecture Notes in Computer Science*, vol. 2507, pp. 205-215, 2002.

[7] W. T. Z. 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, 2014.

[8] 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)*, 2014.

[9] M. T. Naing and A. Thida, “Automatic Myanmar Text Summarization System with Semantic roles,” *Proceedings of the 12th International Conference on Computer Applications (ICCA2014),* Yangon, Myanmar, 2014, pp. 217-223.

[10] D. Radev, H. Jing, M. Stys, D. Tam, “Centroid-based summarization of multiple documents,” *Information Processing & Managemen*t, vol. 40, no. 6, pp. 919-938, 2004.

[11] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *CoRR*, vol. abs/1301.3781, 2013.

[12] Rouane O., Belhadef H., Bouakkaz M., “Word Embedding-Based Biomedical Text Summarization,” *International Conference of Reliable Information and Communication Technology*, vol. 1073, pp. 288-297, 2019

[13] P. Bojanowski, E. Grave, A. Joulin and T. M. ikolov, “Enriching Word Vectors with Subword Information,” *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135-146, 2017.

[14] B. Heinzerling and M. Strube, “BPEmb: Tokenization-free Pre-trained Subword Embeddings in 275 Languages,” *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[15] S. S. Lwin and K. T. Nwet, “Extractive Myanmar News Summarization Using Centroid Based Word Embedding,” *2019 International Conference on Advanced Information Technologies (ICAIT)*, Yangon, Myanmar, 2019, pp. 200-205.

[16] G. Rossiello, P. Basile, and G. Semeraro, “Centroid-based Text Summarization through Compositionality of Word Embeddings,” in *Proceedings of the MultiLing 2017 Workshop on Summarization and Summary Evaluation Across Source Types and Genres*, 2017, pp. 12-21.

[17] W. P. Pa, and N. L. Thein, “Myanmar Word Segmentation using a Combined Model,” *e-Case2009*, 2009

[18] “Mizima Myanmar News and insight,” [Online]. Available: http://www.mizzimaburmese.com/.

[19] “7day News,” [Online]. Available: https://7day.news/.

[20] “Burma Irrawaddy,” [Online]. Available: https://burma.irrawaddy.com/.

[21] “Ministry of information,” [Online]. Available: https://www.moi.gov.mm/.

[22] B. Heinzerling and M. Strube, "Burmese (my) subword embeddings," [Online]. Available: https://nlp.h-its.org/bpemb/my/.
[23] P. Bojanowski, et al., "Wiki word vectors," [Online]. Available: https://fasttext.cc/docs/en/pretrained-vectors.html.
[24] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, T. Mikolov, "Learning Word Vectors for 157 Languages," *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.
[25] C. Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, pp. 74-81, 2004.

## BIOGRAPHIES OF AUTHORS

**Soe Soe Lwin** got her B.C.Sc (Hons) from Computer university (Meiktila) in 2010, and got M.C.Sc in 2012, respectively. Currently she is doing her Ph.D research focusing on Myanmar text summarization in Natural Language Processing Lab, at the University of Computer Studies, Yangon. She is also interested in machine learning and artificial intelligent project.

**Khin Thandar Nwet** received Ph.D (IT) from University of Computer Studies (Yangon) in 2011. She is interested in naturl langage processing research like text summarization, text classification and sentiment analysis. She is currently working at the University of Information technology (UIT).