Clustering Analogous Words in Myanmar Language using Word Embedding Model

Aye Myat Mon, Khin Mar Soe
Natural Language Processing Lab , University of Computer Studies, Yangon, Myanmar
ayemyatmon.ptn@ucsy.edu.mm, khinmarsoe@ucsy.edu.mm

Abstract

Word embedding represents the words in terms of vectors. It is influenced on different NLP research areas such as document classification, author identification, sentiment analysis, etc. One of the most popular embedding models is Word2Vec model. It provides efficient representations of words by using Continuous Bag of Words model (CBOW) and Skip Gram model. In English language, word embedding model can be applied for data preprocessing well but there is a very little amount of work done in Myanmar language. Text preprocessing is important part to build embedding model and it is a significantly effect on final results. This paper tries to extract the analogous words between Myanmar news articles focus on the bag of words (CBOW) model using different features vector sizes. By analyzing word embedding model are obtained the better results with a high dimensional vectors than a low dimensional vectors to cluster the words based on its relatedness.

Keywords: *Word2Vec, Continuous Bag of Words Model (CBOW), Myanmar Language, Word Embedding*

1. Introduction

Word embedding is one of the interesting trends in natural language processing areas. The main advantages of word embedding is that it offers a more expressive and efficient representation by maintaining the contextual similarity of words with a low dimensional vectors. There are two different kinds of word embedding model. They are frequency-based embedding and prediction based embedding. One of the most popular prediction based embedding models is Word2Vec implemented by Mikolov [2,3]. Word2Vec combines with two techniques: Continuous Bag of Words (CBOW) and Skip-Gram model. CBOW predicts the probability of a word by given context word in which single or group of words. Skip gram predicts the context of word by given word.

Word Analogy is finding the relationship of words between two situations. This paper aims to cluster the analogous words in Myanmar news articles based on their semantics relationship with Continuous Bag of Words (CBOW) model that can capture similar word vectors together in vector space. For example in English, vector [king – man + woman] is close to vector [queen].

The remaining parts of the paper are organized as follows: related works have been described in section 2 . In section 3, nature of Myanmar language has been presented. Section 4 describes system overview and the works concerning data preparation and segmentation process expressed in section 5 and section 6. Section 7 explains about word embedding especially focus on Continuous Bag of Words Model and cosine similarity. Section 8 describes the sample results of clustering analogous words and performance analysis in section 9. Finally, the paper has concluded with the future research in section 10.

2. Related Works

Clustering of analogical words has been an essential problem in text mining, question answering, text summarization and information retrieval. Most of the methods have been applied to represent linguistic items in vector spaces. However, very few researches have been carried out on Myanmar text. This section describes previous history of word embedding.

In paper [1], the authors predicted the quality of topic segmentation by using word embedding model depends on latent semantic analysis (LSA), Word2Vec and Glove. They identified which method is more effective to construct word vector representations to provide the semantic meaning of words in English and Arabic languages. Although, the authors found that Word2Vec with CBOW is better than Skip-Gram for frequent words, Skip Gram is more efficient for infrequent words. Based on these results, they compared Word2Vec to LSA and GloVe. They showed that Word2Vec and GloVe are more effective than LSA for both languages. Word2Vec presents the best word vector representations with a small dimensional semantic space compared to GloVe.

Moreover, a comparison of two-word embedding models to cluster semantic similarity words in Tamil language is described in [6]. The authors implemented Continuous Bag of Words and Skip Gram Model using Word2Vec toolkit. In this paper, they used the different feature vector sizes to compare content based word embedding and context-based word embedding for the same word to comment on the accuracy of the models for semantic similarity. They collected the India political news articles from various newspapers in Tamil language. Their data set is huge and it has around 2.7 lakh sentences which contains 50 lakh words. The result showed content- based word embedding model produces better results based on the semantic regularity, whereas contextual based word embedding model produces better results based on the syntactic regularity.

In recent years, Mikolov et al.,[2][3] implemented Word2Vec model that is efficient than the previous embedding models. They observed large improvements in accuracy at much lower computational cost with Word2Vec. It only takes less than a day to learn high quality word vectors from a 1.6 billion words data set.

The advantages of Word2Vec is that it can convert high dimensional vector into low dimensional vector and it can maintain word context. Word2Vec utilize Continuous Bag of Words and Skip Gram model to produce distributed representation of words. Nowadays, many researchers investigate and experiment with Word2Vec and similar techniques to find the relatedness between two conditions described in [4][5]. This paper has focused on Continuous Bag of Words (CBOW) model because it is faster on frequent words than Skip Gram model.

3. Myanmar Language Nature

Myanmar language is the official language of the Union of Myanmar. It is a very rich morphologically language and also a low resource language. Myanmar language has 34 consonants. Myanmar grammar structure is composed of nine parts of speech such as noun, pronoun, adjective, adverb, verb, post-positional marker, particle, conjunction and interjection. A Myanmar syllable has base characters: pre-base character, post-base character, above-base character and below-based character. Each syllable boundary is written from left to right and start with base consonant. It has no delimiter between syllables and words. Myanmar language structure is constructed with subject, object and verb.

4. System Overview

The overview of the system shows in figure1. Firstly, Myanmar sentences are collected to pre-process through the process of segmentation. Besides, stop words list is also prepared to remove the unnecessary words from the document. Finally extracted words are fed into continuous bag of words model in order to cluster analogous words by converting numerical vectors.

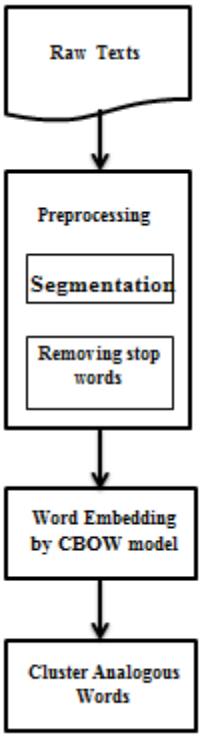

Figure 1. System Overview

5. Data Preparation

Myanmar text corpus which consists of local news articles and blogs from Myanmar websites was created for training model [8][9][10]. Then 16612 sentences are collected from 7Days Daily News, 1337 sentences from Moemaka Blog and 43807 sentences from Burma Irrawaddy Blog to construct word embedding model. These sentences include different types of articles: health, crime, sport and general knowledge. We converted the text in Unicode [12]. These sentences are saved as (.txt) format for training process. Each sentence contains 50 words as average. More data will be collected in the future. But, many standard data sets can be easily used for

English language. Table (1) shows the collected data from Myanmar news and blogs.

Table 1. News and Blogs Data Set

	#Words	#Sentences
7Days Daily News	5,667,280	16,612
Moemaka Blog	431,007	1,337
Burma Irrawaddy Blog	16,949,598	43,807

6. Preprocessing

Pre-processing plays an important role in many natural language processing research areas because Myanmar language has no space like Japanese, Chinese, Thai and India. They are written from left to right continuously. In order to extract words from the collection of Myanmar text, it is firstly needed to segment text into separate meaningful words. For English text, we can easily use the existing libraries and tools. In this paper, Myanmar text data are segmented by 'Myanmar Word Segmentor' from UCSY-NLP lab [11]. After that, stop words, punctuations and special characters are removed from the collected documents. Some Myanmar stop words are related with date, time, numbers and conjunction. Currently, we manually checked the spelling errors for the segmented text. Table (2) shows the sample segmented and spelling corrected sentences of the preprocessing task.

Table 2. Sample Segmented Myanmar Word

1.	ပုသိမ်မြို့တွင်ကျင်းပတဲ့မီးပုံးပျံလွှတ်ပွဲတော်သို့တောင်ကြီးတန်ဆောင်တိုင်ပွဲဦးရောက်ယှဉ်ပြိုင်ခဲ့သည့်ရှမ်းတိုင်းရင်းသားမီးပုံးပျံပညာရှင်များကိုဖိတ်ခေါ်၍ကျင်းပခဲ့ခြင်းဖြစ်ပြီးအလှပြိုစိန်နားပန်မီးပုံးပျံ၂လုံးနဲ့ညမီးကြည်မီးပုံးပျံတစ်လုံးလွှတ်တင်ခဲ့ရာပုသိမ်မြို့ခံများထောင်နှင့်ချီ၍လာရောက်အားပေးခဲ့ကြပါတယ်။	ပုသိမ်/မြို့/တွင်/ကျင်းပ/တဲ့/မီးပုံးပျံ/လွှတ်/ပွဲတော်/သို့/တောင်ကြီး/တန်ဆောင်တိုင်/ပွဲ/ဦ/ဝင်ရောက်/ယှဉ်ပြိုင်/ခဲ့သည့်/ရှမ်း/တိုင်းရင်းသား/မီးပုံးပျံ/ပညာရှင်/များ/ကို/ဖိတ်ခေါ်/၍/ကျင်းပ/ခဲ့ခြင်း/ဖြစ်/ပြီး/အလှပြ/စိန်နားပန်/မီးပုံးပျံ/၂/လုံး/နဲ့/ညမီးကြည်/မီးပုံးပျံ/တစ်/လုံး/လွှတ်တင်/ခဲ့ရာ/ပုသိမ်/မြို့ခံ/များ/ထောင်/နှင့်ချီ၍/လာရောက်/အားပေး/ခဲ့ကြပါတယ်/။
2.	ရောဝတီတိုင်းဒေသကြီးမှာအဖွဲ့အစည်းမျိုးစုံကကြောင်းအမျိုးမျိုးပြသိမ်းဆည်းထားတဲ့မြေအများအပြားရှိနေတာပါ။	ရောဝတီတိုင်းဒေသကြီး/မှာ/အဖွဲ့အစည်း/မျိုးစုံ/က/အကြောင်း/အမျိုးမျိုး/ပြ/သိမ်းဆည်း/ထားတဲ့/မြေ/အများအပြား/ရှိ/နေတာပါ/။

7. Word Embedding

The main target of word embedding model is to convert word to the form of numeric vectors. We need to do word embedding because many machine learning algorithms and most of the deep learning architectures cannot process the raw form of strings or plain texts. There are several models to learn word embedding. They are count-vector, tf-idf vectorization , co-occurrence matrix and Word2Vec. Most popular model architectures in Word2Vec are Continuous Bag of words Model and Skip Gram Model. Skip Gram works on small amount of training data and it can represent for rare words or phrases. Continuous Bag of Words model is faster than the skip gram model and it can train on large amount of data and it is slightly better accuracy for the frequent words. In this paper, we work on Continuous Bag of Words Model.

7.1. Continuous Bag of Words Model (CBOW)

Continuous Bag of Words model is the neural network inspired model. In CBOW, the context vectors are summed and used to predict the target. Let there be a corpus, a sequence of words $w_1, w_2 \ldots, w_T$. The window is defined by parameter c, where c words at the right and left of the target are taken. The objective function of continuous bag of words model is in equation (1):

$$\frac{1}{T}\sum_{t=1}^{T}\log p(w_t|\sum_{-c\le j\le c, j\neq 0} w_{t+j}) \qquad \text{eq(1)}$$

Continuous Bag of Words Model with negative log likelihood function of a word given a set of context is shown in equation (2).

$$p(w_O|w_I) = \frac{\exp\left({v'_{w_O}}^{\top} v_{w_I}\right)}{\sum_{w=1}^{W}\exp\left({v'_w}^{\top} v_{w_I}\right)} \qquad \text{eq(2)}$$

where,

wo = output word
wi = context words

Architecture of CBOW model is shown in figure 2. In the following architecture, vocabulary size is V and hidden layer (projection layer) size is d (dimension). It is fully connected with their adjacent layers. The input is a one-hot encoded vector that given input context words, only one out of V units $\{x_1, x_2, \ldots x_V\}$ be 1 and all other units be 0. Hidden layer is represented by V x d matrix W. Each row of W is the d dimensional vector representation of related word w in input layer.

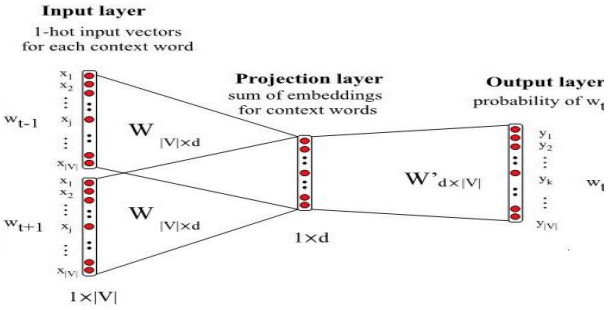

Figure2. Continuous Bag of Words Model Architecture

7.2. Cosine Similarity

Cosine similarity is a similarity measurement between vectors of an inner product space. The cosine value 0 is 1 and less than 1 for any other the angle; the lowest value of the cosine is -1. Basically, the cosine of the angle between two vectors determines whether two vectors are pointing in the same direction. Cosine similarity is widely used in text/document matching. The similarity between two documents A and B compute the cosine similarity of their vector representations and measure the cosine of the angle between vectors [7]. We return the documents ranked by the closeness of their vectors. The resulting similarity range from -1 means exactly opposite, to 1 means exactly the same and with 0 usually indicating independence, and in -between values indicating intermediate similarity or dissimilarity. The cosine similarity formula is shown in equation (3).

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}, \quad \text{eq(3)}$$

Where,

A_i=components vector of A
B_i =components vector of B

8. Clustering Analogous Words

This experiment shows the sample result of analogous words for မြန်မာ-Myanmar (country), ရန်ကုန်-Yangon (City) and ပုသိမ်-Pathein (Town) .We use the dataset that contains 7 Days Daily news articles, Moemaka Blog and Burma Irrawaddy Blog including various domains to train on Word2Vec model. We tested with different dimensional vectors 100, 200 and 300 context window size 2 but we can setup with context window size between 2 to 10. Table (3,4,5) show the sample clustered result of top 10 analogous words with dimensional vectors 300. The cosine similarity score 1 means that two vectors are equal and 0 means they have no relation to each other.

Table3.Top Ten Analogous words for မြန်မာ (Myanmar)

Word	Cosine Similarity [0-1]
လာအို	0.627298
ကနေဒါ	0.625642
ကိုးရီးယား	0.590485
ထိုင်း	0.588882
သြစတြေးလျ	0.583560
နော်ဝေး	0.573853
ဖင်လန်	0.572634
အိမ်နီးချင်း	0.570092
ဖိလစ်ပိုင်	0.568185
ဒိန်းမတ်	0.560263

For example in မြန်မာ (Myanmar) , the cosine similarity score between မြန်မာ (Myanmar) and လာအို (Laos) is 0.627298 means the most analogous word for မြန်မာ (Myanmar) is လာအို (Laos).

Table4. Top Ten Analogous words for ရန်ကုန် (Yangon)

Word	Cosine Similarity [0-1]
ကန်တော်ကြီး	0.499090
ပြည်လမ်း	0.497475
ရွှေတိဂုံ	0.478186
ဆူးလေ	0.464573
ပဲခူး	0.456181
အနော်ရထာလမ်း	0.455162
ဗဟန်း	0.441475
ရွှေဂုံတိုင်	0.440680
ပြည်သူ့ရင်ပြင်	0.423175
အင်းစိန်	0.412653

For ရန်ကုန် (Yangon) the cosine similarity score between ရန်ကုန် (Yangon) and ကန်တော်ကြီး (KanDawGyi Park) is 0.499090 means the most analogous word for ရန်ကုန် (Yangon) is ကန်တော်ကြီး (KanDawGyi Park)

Table5. Top Ten Analogous words for ပုသိမ် (Pathein)

Word	Cosine Similarity [0-1]
ဧရာဝတီတိုင်းဒေသကြီး	0.589040
သာပေါင်း	0.553002
ကျုံပျော်	0.545658
လက်ပွတ္တာ	0.537680
ကန်ကြီးထောင့်	0.530157
ကျောင်းကုန်း	0.520893
မြောင်းမြ	0.510949
ပန်းတနော်	0.502957
ကမ်းနားလမ်း	0.490398
ရေကြည်	0.486770

The cosine similarity score between ပုသိမ် (pathein) and ဧရာဝတီတိုင်းဒေသကြီး (Ayeyarwaddy division) is 0.589040 .

The most analogous word for ပုသိမ် (pathein) is ဧရာဝတီတိုင်းဒေသကြီး (Ayeyarwaddy division).

9. Evaluation

The performance can be measured by using intrinsic evaluation that directly check the semantic and syntactic relationships between words. We pre-collected 100 related words pairs in Myanmar language and manually checked to calculate the accuracy of Continuous Bag of Words model by using several dimensional vectors 100, 200 and 300. Most of the Myanmar words have many analogous words but different words have similar meaning. In this experiment, the higher accuracy is gained with higher dimensional vectors. Table (6) shows the analogous words list for "မြန်မာ"(Myanmar) and "ပုသိမ်" (Pathein) . The performances are also evaluated by precision, recall and f-measure in figure 3.

Table 6. Analogous Words List

Analogous Words List for မြန်မာ (Myanmar)	Analogous Words List for ပုသိမ် (Pathein)
မြန်မာ-လာအို	ပုသိမ်-ဧရာဝတီတိုင်းဒေသကြီး
မြန်မာ-ကနေဒါ	ပုသိမ်- သာပေါင်း
မြန်မာ-ကိုးရီးယား	ပုသိမ်- ကျုံပျော်
မြန်မာ-ထိုင်း	ပုသိမ်- လက်ပွတ္တာ
မြန်မာ-ဩစတြေးလျ	ပုသိမ်- ကန်ကြီးထောင့်
မြန်မာ-နော်ဝေး	ပုသိမ်- ကျောင်းကုန်း
မြန်မာ-ဖင်လန်	ပုသိမ်- မြောင်းမြ
မြန်မာ-အိမ်နီးချင်း	ပုသိမ်- ပန်းတနော်
မြန်မာ-ဖိလစ်ပိုင်	ပုသိမ်- ကမ်းနားလမ်း
မြန်မာ-ဒိန်းမတ်	ပုသိမ်- ရေကြည်
မြန်မာ-ဗီယက်နမ်	ပုသိမ်- ငပုတော
မြန်မာ-ဥရောပ	ပုသိမ်- ဘုတ်ပြင်း
မြန်မာ-ဆွီဒင်	ပုသိမ်- ဟင်္သာတ
မြန်မာ-ချက်သမ္မတ	ပုသိမ်- ရွှေတောင်
မြန်မာ-ကူဝိတ်	ပုသိမ်- ငွေဆောင်
မြန်မာ-ဆော်ဒီအာရေးဘီးယား	ပုသိမ်- ဆလိုင်းသန့်
မြန်မာ-အင်ဒိုနီးရှား	ပုသိမ်- အင်္ဂပူ
မြန်မာ-ကမ္ဘောဒီးယား	ပုသိမ်- ညောင်တုန်း
မြန်မာ-ဆွစ်ဇာလန်	ပုသိမ်- ချောင်းသာ
မြန်မာ-တရုတ်	ပုသိမ်- ဂေါရန်ဂျီ
မြန်မာ-ဂျာမဏီ	ပုသိမ်- ဟိုင်းကြီးကျွန်း
မြန်မာ-အမေရိက	ပုသိမ်- လေတပ်စခန်း
မြန်မာ-ဂျပန်	ပုသိမ်- ပြည်မြို့
မြန်မာ-ဘူတန်	ပုသိမ်- ရွှေမုဋ္ဌော
မြန်မာ- ဘင်္ဂလားဒေ့ရှ်	ပုသိမ်- ရေကြည်ဦး
မြန်မာ- အိန္ဒိယ	ပုသိမ်- အသည်ကြီး
မြန်မာ- အာရ္ဂတိုက်	ပုသိမ်- ရွှေမြင်းပျံ
မြန်မာ- အရှေ့တောင်အာရှ	ပုသိမ်- ဟာလဝါ

Precision= Correct number of words pairs / Total number of tested words pairs

Recall= Correct number of words pairs / Total number of system generated words pairs

F-measure=2* Precision * Recall / Precision + Recall

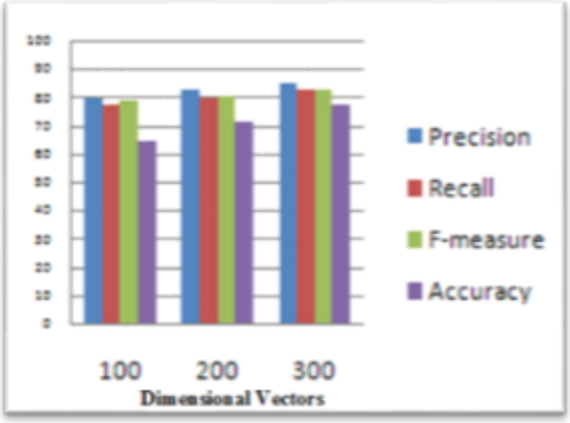

Figure 3. Evaluation measured by CBOW model

10. Conclusion

This study has applied word embedding model focus on Continuous Bag of Words representation to find analogous words in Myanmar language has been described. By doing this, document classification, information retrieval and other natural language processing areas can be increased their accuracies. As future research, this work can also be extended to find cross lingual semantic similarity. We can find words which are similar in meaning from different languages by applying word embedding model. It plays an important role in multilingual machine translation system.

References

[1]MarwaNaili*, AnjaHabachaChaibi, HendaHajjami Ben Ghezala , " Comparative study of word embedding methods in topic segmentation" ,International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES2017, 6-8 September 2017, Marseille, France, PP:340-349.

[2]Mikolov, T., Chen, K., Corrado, G., & Dean, J. "Efficient Estimation of Word Representations in Vector Space". In Proceedings of Workshop at ICLR, 2013. PP: 1301-3781.

[3]Mikolov, T ., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). "Distributed representations of words and phrases and their compositionality". In Advances in neural information processing systems. PP: 3111-3119.

[4] Pennington, J., Socher, R., & Manning, C. (2014). "Glove: Global vectors for word representation." In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).

[5]Soliman, A. B., Eissa, K., & El- Beltagy, S. R. (2017). "Aravec: A set of Arabic word embedding models for use in Arabic nlp," Procedia Computer Science, 117, 256-265.

[6]S. G. Ajay*, M. Srikanth, M. Anand Kumar and K. P. Soman, "Word Embedding Models for finding Semantic Relationship between Words in Tamil Language",Indian Journal of Science and Technology, Vol 9(45), DOI: 10.17485/ijst/2016/v9i45/106478, December 2016

[7]http://www.miislita.com/Cosine -Similarity.html

[8]https://blog.moemaka.com/

[9]http://www.7daydaily.com/

[10]https://burma.irrawaddy.com/

[11]http://www.nlpresearch-ucsy.edu.mm

[12]http://www.myanmarengineer.org/converter/fontconverter.html