Proceedings of the SMART–2019, IEEE Conference ID: 46866
8th International Conference on System Modeling & Advancement in Research Trends, 22nd–23rd November, 2019
College of Computing Sciences & Information Technology, Teerthanker Mahaveer University, Moradabad, India

Vector Representation of Bengali Word Using Various Word Embedding Model

Ashik Ahamed Aman Rafat[1], Mushfiqus Salehin[2], Fazle Rabby Khan[3],
Syed Akhter Hossain[4] and Sheikh Abujar[5]
Dept. of Computer Science and Engineering,
Daffodil International University, Dhaka, Bangladesh
E-mail: [1]aman15-6858@diu.edu.bd, [2]mushfique15-7056 @diu.edu.bd, [3]rabby15-6727 @diu.edu.bd
[4]aktarhossain@daffodilvarsity.edu.bd, [5]sheikh.cse@diu.edu.bd

***Abstract*—To transfer human understanding of language to a machine we need word embedding. Skipgram, CBOW, and fastText is a model which generate word embedding. But finding pretrained word embedding model for the Bengali language is difficult for researchers. Also, training word embedding is time-consuming. In this paper, we discussed different word embedding models. To train those models, we have collected around 500000 Bengali articles from various sources on the internet. Among them, we randomly chose 105000 articles. Those articles have 32 million words. We trained them on SkipGram and CBOW model of Word2Vec, fastText. We also trained those words in Glove model. Among the all result fastText (Word2Vec) gave us a satisfactory result.**
Keywords— Bengali Words, Skip Gram, CBOW, Word2Vec, FastText, Glove, Word Embedding

I. Introduction

We have human can understand words by their context or surrounding words. In communication, we share thoughts, ideas with each other through language. We can produce an infinite number of sentences with a finite number of words. As we can produce an infinite number of sentences that told us words can have separate meaning based on the context used. But the computer doesn't understand words or its context. Here distributed representation of words plays a big role.

In distributed representation word describe as 50-300-dimensional vector. Word embeddings know as word representation that transfers human interpretation of language to the machine. Many NLP problems can be solved through word embeddings. There are so many neural network-based algorithms coming in the natural language processing field. In RNN we give input as a sequence of words. Many researchers showed that if we give these neural network distributed representation words, they perform better for various NLP tasks.

Among all of the word embeddings models Word2Vec is most popular which is proposed by Mikolov and Dean [1].

Glove [2] is another word embedding model which is presented by Pennington et al. Facebook developed another word embedding model which is known as fastText[3]. Here we will analyze these three words embedding models for Bengali words.

II. Literature Review

Recently There are so many works have been done for Bangla word embedding. In 2016 Abhishek et al. proposed a neural lemmatizer [5] for Bengali word embedding which used Word2Vec model. Adnan Ahmad and Mohammad Ruhul Amin [4] collected large dataset. That dataset has 2,185,701 articles which have 51,920,010 sentences. For the Word2Vec model, they took words that were occurred a minimum of 5 times. They released a Word2Vec word embedding model which has dictionary size over 200k of unique words. Nowshad et al [6] made a multi-label sentence classification model. Where they used LibSVM and Scikit-learn in 5000, 7500 and 10000 sentences corpus.

In 2018 Ritu [7] et al. analysis most used embedding models for Bengali words. They used SUMono [10] dataset as well as their own dataset to train their model. They showed differences between Word2vec and fastText models. Same year Sumit et al. from Socian [8] Ltd made a Word2Vec model which was trained on 623,510,478. They used CBOW and Skip-Gram with embedding vector 150 and 1,245,974 words.

Islam, Md Saiful [9] first time in Bengali embeddings used ANN, CNN, RNN. They trained CBOW, Skip-Gram of Word2Vec and fastText as well as Glove in ANN, CNN, RNN model. From what we saw that fastText (Skip-Gram) has the highest accuracy and Glove has the lowest accuracy in ANN, RNN, CNN models.

III. Methodology

A. Data Collection and Pre-processing

In natural language processing, we do need a large amount of data. We made a web parser which collected around 500000 articles from the various website. For legal reasons and copyright issue, we cannot disclose the website's names. From 500000 articles we have selected 105000 articles for our word embedding and named it corpus105k. Then we trimmed the whole dataset with a minimum of a single word occurring 25 times. If a random word enumeration is less than 25 then we discarded that word. Table 1 demonstrate total words in the corpus105k dataset with unique words.

TABLE 1: DATASET DEMOGRAPHY

Total words before trimming	32112604
Unique words before trimming	505383
Total words after trimming	30731553
Unique words after trimming	53473

B. Word Embedding

When we communicate, we connect words according to their meanings. We say/write a word in the context of previous words or its surrounded words. But the computer doesn't understand this kind of things. So, researchers have published many word embedding models which can help a machine to understand the context of sentences

1. Skip-gram: This model predicts the nearby words when a target word is given. Consider an example "আমি বাংলায় গান গাইতে ভালবাসি". If we take the middle word 'গান' as a target word, Skip-gram model will predict the possibility of 'আমি', 'বাংলায়', 'গাইতে', 'ভালবাসি' as a surrounding word.

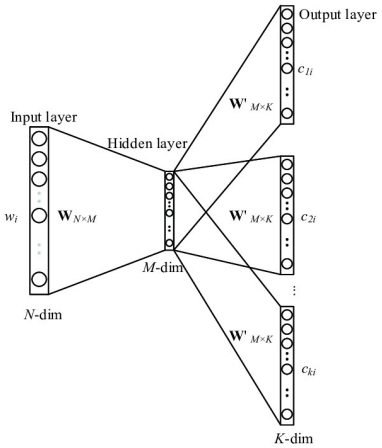

Fig. 1: Skip-gram Model Architecture

2. Continuous Bag-Of-Words (CBOW): cbow model works the exact opposite of skip-gram model. This model predicts middle words when nearby words are given. From the previous example if you take surrounding words 'আমি', ' বাংলায়', 'গাইতে', 'ভালবাসি' than cbow model will predict the possibility of 'গান' as middle word.

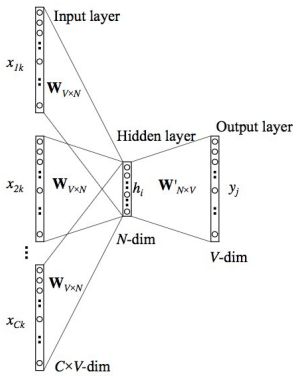

Fig. 2: CBOW Model Architecture

3. Global Vectors (GloVe): Glove build a big matrix which is co-occurrence of data, containing data on how often each word occurs. Afterward, glove minimizes this matrix into a lower-dimensional matrix using reconstruction loss.
4. FastText: This model is the addition of the Word2Vec model. fastText make every word as n-gram of characters. For example, if we take 'খবর' as a word with n=2 than fastText model represent it as <খ, খব, বর, র>. Here angular brackets specify the start and end of the word.

C. Experiment

Word embedding has so many models. Among them, we choose to work with five models which are prediction-based models.

- Word2Vec (Skip-Gram)
- Word2Vec (CBOW)
- FastText (Skip-Gram)
- FastText (CBOW)
- Glove Word Embeddings

In our dataset, we have 32 million of words where we discarded words that have been occurred less than 20 times. Thus, this makes 53000 words of vocabulary size. We used gensim library for Word2Vec models, fastText Facebook code's for fastText model and glove-python library for Glove model. Each model trained on 50 epochs. Parameters which are used in the model are shown in Table 2.

TABLE2: PARAMETERS(EMBEDDINGDIMENSION, WINDOWSIZE, LEARNINGRATE, MINIMUM WORD COUNT, NEGATIVE SAMPLING)

Model	dim	win	alpha	count	neg
Word2Vec (Skip-Gram)	300	5	0.03	20	20
Word2Vec (CBOW)	300	5	0.03	20	20
Glove	300	5	0.05	-	-
fastText(skip-gram)	300	5	0.03	20	20
fastText (CBOW)	300	5	0.03	20	20

IV. PERFORMANCE & DISCUSSION

To check our models, we validated the models with two experiments.

A. Nearby words

We gave our models a word to show us ten nearby words. Each of the models gave us a different result. But diversity in the models is not huge. Table 3, 4, 5, 6, 7 shows different models result. In nearby words column, the left-most word is the best match for the context word and right-most word is the least match for the context word.

TABLE 3: WORD2VEC (SKIP-GRAM)

Word	Nearest Word
প্রধানমন্ত্রী	'হাসিনা', 'হাসিনার', 'প্রধানমন্ত্রীর', 'শেখ', 'মন্ত্রী', 'নরেন্দ্র', 'রাষ্ট্রপতি','হাসিনাকে', 'তিনি', 'বিরোধীদলীয়'
গ্রীষ্মকালীন	শীতকালীন', 'ঈদ-উল-ফিতর', 'মৌসুমে', 'এপ্রিল-মে', 'পীরগঞ্জে', 'শাকসবজি', 'এবছর', 'আন্তঃবিভাগ', 'জম্মু-কাশ্মীরে', 'চলতি'
বাংলা	'ট্রিবিউনকে', 'বলেন', 'আলোকে', 'বিডিনিউজ', 'টোয়েন্টিফোর','প্রসঙ্গে', '‘', 'ডটকমকে', 'জানান', 'ট্রিবিউন'
শনিবার	'শুক্রবার', 'মঙ্গলবার', 'বুধবার', 'বৃহস্পতিবার', 'সোমবার','রবিবার', 'রোববার', 'এপ্রিল', 'গতকাল', 'সকালে'
দূষণমুক্ত	'দূষণ', 'নদীগুলোকে', 'পরিবেশদূষণ', 'বর্জ্যমিশ্রিত', 'তীরভূমি', 'দূষণের', 'পরিবেশ', 'যানজটমুক্ত', 'দখলমুক্ত', 'দূষিত'

TABLE 4: WORD2VEC (CBOW)

Word	Nearest Word
প্রধানমন্ত্রী	'হাসিনা', 'প্রধানমন্ত্রীর', '‘প্রধানমন্ত্রী', 'রাষ্ট্রপতি', 'শেখ', 'মন্ত্রী', 'বঙ্গবন্ধুকন্যা', 'পররাষ্ট্রমন্ত্রী', 'সরকারপ্রধান', 'অর্থমন্ত্রী'
গ্রীষ্মকালীন	'শীতকালীন', 'ঈদ-উল-ফিতর', 'গ্রীষ্মের', 'বড়দিনের', 'আশুরার', 'গ্রীষ্মে', 'মাসকে', 'আযহা', 'পবিত্র', 'ঈদুল'
বাংলা	'বাংলা', 'বলেন', 'ট্রিবিউনকে', 'চৌধুরী', 'আলোকে', 'বিডিনিউজ', 'মো', 'খান', 'টোয়েন্টিফোর', 'ইংরেজী'
শনিবার	'শুক্রবার', 'সোমবার', 'বুধবার', 'বৃহস্পতিবার', 'রোববার', 'মঙ্গলবার','রবিবার', 'সোমবার', 'গতকাল', 'এপ্রিল'
দূষণমুক্ত	'দূষণ', 'নদীগুলোকে', 'দূষিত', 'দখলমুক্ত', 'যানজটমুক্ত', 'খালগুলো','দূষণের', 'প্রবাহমান', 'পরিশোধনের', 'পরিচ্ছন্ন'

TABLE 5: GLOVE

Word	Nearest Word
প্রধানমন্ত্রী	'মে-ই', 'হেকমতিয়ারকে', 'হাসিনা', 'এক্সপায়ারি', 'টেরেসা', 'নাওমিচি', '‘ইসরায়েলি', 'সরকার', 'পেনাংয়ে', 'আবাদিকে'
গ্রীষ্মকালীন	'শীতকালীন', 'গার্মেন্টসগুলোর', 'ছুটি', 'পিতৃত্বকালীন', 'মাতৃত্বকালীন','ক্রিসমাসের', 'মন্ত্রণালয়ে', 'অবকাশ', 'এনজ্যাক', ডাক্তারেরও
বাংলা	'করপাস', 'হিন্দি-উড়িয়া', 'আগ্রাবাদিয়ানদের', 'বাংলা-বাঙালি', '‘এপার', 'সমার্থ', 'ভাষাবাসী', 'পলাশী–পূর্ব', 'লংকা', 'ইউ-এস'
শনিবার	'মঙ্গলবার', 'বুধবার', 'বৃহস্পতিবার', 'তাসিস', 'সোমবার', 'রোববার','শুক্রবার', 'চেষ্টোয়', 'খায়ওনি', 'মুবারাক'
দূষণমুক্ত	চরিতার্থ', 'জীবাণুমুক্ত', 'পরিবেশকে', 'হকারমুক্ত', 'শান্তি-শৃঙ্খলা', 'চাঙ্গা', 'চাঙা', 'মেধাশূন্য', 'পয়োনিষ্কাশন', 'গোয়েন্দাবাহিনীও'

TABLE 6: FASTTEXT (SKIP-GRAM)

Word	Nearest Word
প্রধানমন্ত্রী	'হাসিনা', 'হাসিনার', 'প্রধানমন্ত্রীর', '‘প্রধানমন্ত্রী', 'প্রধানমন্ত্রীও', 'উপ-প্রধানমন্ত্রী', 'শেখ', 'প্রধানমন্ত্রীসহ', 'মন্ত্রী', 'উপপ্রধানমন্ত্রী'
গ্রীষ্মকালীন	গ্রীষ্মকাল', 'গ্রীষ্মকালে', 'শীতকালীন', 'গ্রীষ্মের', 'গ্রীষ্ম', 'গ্রীষ্মে', 'ঈদ-উল-ফিতর', 'এপ্রিল-মে', 'মৌসুমে', 'চলতি'
বাংলা	'ট্রিবিউনকে', 'বলেন', 'আলোকে', 'জানান', '‘বাংলা', '‘', 'প্রসঙ্গে', 'খান', 'বিডিনিউজ', 'মো'
শনিবার	'শুক্রবার', 'মঙ্গলবার', 'বুধবার', 'সোমবার', 'বৃহস্পতিবার', 'রোববার', 'রবিবার', 'গতকাল', 'এপ্রিল', 'মার্চ'
দূষণমুক্ত	'দূষণমুক্ত', 'দূষণ', 'দূষণে', 'শোষণমুক্ত', 'নদীগুলোকে', 'দূষণকারী', 'নদীগুলো', 'দূষণের', 'দূষিত', 'প্রবাহমান'

TABLE 7: FASTTEXT (CBOW)

Word	Nearest Word
প্রধানমন্ত্রী	'প্রধানমন্ত্রীও', 'উপপ্রধানমন্ত্রী', ''প্রধানমন্ত্রী', 'উপ-প্রধানমন্ত্রী', 'মন্ত্রীর', 'প্রধানমন্ত্রীসহ', ''প্রধানমন্ত্রীর', 'প্রধানমন্ত্রিত্ব', 'প্রধানমন্ত্রীকে', 'বিমানমন্ত্রী'
গ্রীষ্মকালীন	'গ্রীষ্মকাল', 'গ্রীষ্মকালে', 'গ্রীষ্ম', 'গ্রীষ্মে', 'শীতকালীন', 'গ্রীষ্মের', 'রাত্রিকালীন', 'সান্ধ্যকালীন', 'শীতকালে', 'স্বল্পকালীন'
বাংলা	'বাংলা', 'জয়বাংলা', 'বাংলাহিলি','বাংলা', 'পূর্ববাংলা', 'ডাচ-বাংলা', 'ডাচ্-বাংলা', 'ডাচ্-বাংলা', 'বাংলার', 'বাংলার'
শনিবার	'শুক্রবার', 'রোববার', 'বুধবার', 'সোমবার', 'মঙ্গলবার', 'রবিবার', 'বৃহস্পতিবার', 'বৃস্পতিবার', ''বৃহস্পতিবার', ''রবিবার'
দূষণমুক্ত	'দূষণমুক্ত', 'দূষণ', 'দূষণে', 'নিষ্কাশন', 'পয়ঃনিষ্কাশন', 'দূষিত', 'শোষণমুক্ত', 'দূষণের', 'পয়নিষ্কাশন', 'বর্জ্য'

B. *Cosine Similarity Words*

Cosine similarity in words measure the angle between them. To get similarity between two vectors we used cosine similarity. Table 8 and 9 showing cosine similarity between couple of words in angle.

TABLE 8: COSINE SIMILARITY AND THE ANGLE BETWEEN BENGALI WORDS (ছলে - ময়ে)

Model	Cosine Similarity	Angle
Word2Vec (Skip-Gram)	0.79	36.99
Word2Vec (CBOW)	0.73	42.28
Glove	0.51	59.12
fastText (Skip-Gram))	0.80	36.38
fastText (CBOW)	0.73	42.61

TABLE 9: COSINE SIMILARITY AND THE ANGLE BETWEEN BENGALI WORDS (রাজা - রানি)

Model	Cosine Similarity	Angle
Word2Vec (Skip-Gram)	0.73	64.23
Word2Vec (CBOW)	0.43	67.34
Glove	0.73	43.11
fastText (Skip-Gram))	0.49	60.39
fastText (CBOW)	0.42	64.98

Five different models produced different word embeddings. Among them fastText Skip-Gram showing us some good result. It produced similar words as well as slightly moderate words. Glove gave us the worst result out of all five models. It produces some good result for specific words but sometimes generated random words. Second best result with coming from fastText CBOW model. It gives us a similar result as fastText Skip-Gram model but some random changes in certain words. Both of the Word2Vec model giving us a result which good but not great.

V. CONCLUSION AND FUTURE WORK

Natural language processing is not an easy task and Bangla language one of the complexes in the world. In our work, we give some intuition for Bengali word embedding. We have trained all of the models over 32 million of words although we have more than 1 billion words of the dataset. Due to computational limitation, we weren't able to train that dataset. But in 32 million of words fastText gave some satisfactory result. The main purpose of word embedding to learn its surrounding words. We try to give result of nearest word of every model. For future work, we will try to investigate how result differs in those models if we trained them 1 billion of Bengali words.

REFRENCES

[1] Mikolov, T., Chen, K., Carrado, G. and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. 1st ed. [ebook] Available at: http://arxiv.org/pdf/1301.3781.pdf [Accessed 20 Nov. 2015].
[2] GloVe: Global Vectors for Word Representation Jeffrey Pennington, Richard Socher, Christopher D. Manning.
[3] Armand Joulin et al. "Bag of tricks for efficient textclassication". In: arXiv preprint arXiv:1607.01759(2016).
[4] A. Ahmad and M. R. Amin, "Bengali word embeddings and its application in solving document classification problem," 2016 19th International Conference on Computer and Information Technology (ICCIT), Dhaka, 2016, pp. 425-430.
[5] Chakrabarty, A. & Garain, U. (2016). BenLem (A Bengali Lemmatizer) and Its Role in WSD. ACM Transactions on Asian and Low Resource Language Information Processing ACM Trans. Asian Low-Resour. Lang. Inf. Process., 15(3), 1-18. doi:10.1145/2835494.
[6] Nowshad Hasan, Md & Bhowmik, Sourav & Rahaman, Md. (2017). Multi-label sentence classification using Bengali word embedding model. 1-6. 10.1109/EICT.2017.8275207.
[7] Ritu, Zakia & Nowshin, Nafisa & Nahid, Md Mahadi & Ismail, Sabir. (2018). Performance Analysis of Different Word Embedding Models on Bangla Language. 1-5. 10.1109/ICBSLP.2018.8554681.
[8] Sumit, Sakhawat & Hossan, Md. Zakir & Muntasir, Tareq & Sourov, Tanvir. (2018). Exploring Word Embedding for Bangla Sentiment Analysis. 10.1109/ICBSLP.2018.8554443.
[9] Islam, Md Saiful. (2018). A Comparative Analysis of Word Embedding Representations in Authorship Attribution of Bengali Literature.
[10] M. A. Al Mumin, A. A. M. Shoeb, M. R. Selim, and M. Z. Iqbal, "Sumono: A representative modern bengali corpus.