Digital Receipt

This receipt acknowledges that Turnitin received your paper. Below you will find the receipt information regarding your submission.

The first page of your submissions is displayed below.

Submission author:	Ye Kyaw Thu
Assignment title:	NLP Paper
Submission title:	para-word2vec-final
File name:	word2vec.pdf
File size:	187.78K
Page count:	6
Word count:	4,278
Character count:	21,736
Submission date:	28-Jan-2021 12:12PM (UTC+0630)
Submission ID:	1496065398

Word Presentation For Paraphrase Myanmar Language Using Word2vec Model

1st Myint Myint Htay
University of Technology(Yatanarpon Cyber City)
myintmyinthtay@utycc.edu.mm

2nd Myat Nyein Chan
University of Technology(Yatanarpon Cyber City)
myatnyeinchan@utycc.edu.mm

3rd May Phyo Aung
University of Technology(Yatanarpon Cyber City)
mayphyoaung@utycc.edu.mm

4th Ei Phyu Phyu Mon
University of Technology(Yatanarpon Cyber City)
eiphyuphyumon@utycc.edu.mm

***Abstract*—Word2Vec methods are widely used to evaluate similarity scores and to classify text in a variety of language. Both skipgram and CBOW are also two of the most popular methods based on word2vec in state-of-art natural language processing (NLP). Skip-gram works well with small amount of data and is found to represent rare words well. When we anlaysed the two experiments with large amount of data, Continuous Bag-of-Word (CBOW) was faster and got better representations for more frequent words. This paper analyses the performance of two models with similarity scores and processing time for myanmar language (Burmese). Data are collected in domain with respect to the travelling and daily general conversations.**

***Index Terms*—Word Embedding, Word2vec, CBOW, skipgram**

I. Introduction

Natural Language Processing (NLP) is the field of artificial intelligence that studies the interactions between computers and human languages, in particular how to program computers to process and analyze large amounts of natural language data. NLP is often applied for classifying text data. Text classification is the problem of assigning categories to text data according to its content. There are different techniques to extract information from raw text data and use it to train a classification model. These techniques are Bag-of-Words (used with a simple machine learning algorithm such as TI-Idf), the popular word embedding model Word Embedding (used with a deep learning neural network such as Word2Vec), and the state-of-the art Language models (used with transfer learning from attention-based transformers such BERT) that have completely revolutionised the NLP landscape. All of these techniques, we applied this proposed system by using word embedding. There is a lot of progress being currently made in NLP using word embedding, it is a positive trend that can be used in a very broad range of practical NLP applications such as computing the similarities between words, using as features in text classification and different natural language tasks such as sentiment analysis. Word2Vec produces a vector space, typically of several hundred dimensions, with each unique word in the corpus such that words that share common contexts in the corpus are located close to one another in the space. That can be done using 2 different approaches: starting from a single word to predict its context (skipgram) or starting from the context to predict a word (Continuous Bag-of-Words). The word embedding can be useful to predict the news category. The word vectors can be used in a neural network as weights. First, the corpus is transformed into padded sequences of word ids to get a feature matrix. Then, create an embedding matrix so that the vector of the word with id N is located at the Nth row. Finally, build a neural network with an embedding layer that weighs every word in the sequences with the corresponding vector. In this proposed paper, we analyzed the performance of skipgram and CBOW models on both small amount of training data (only 100 senteces) and large amount of training data (242,327 sentences).

II. Related Work

In [1] showed that where embeddings factorise pointwise mutual information (PMI), it is paraphrasing that determines when a linear combination of embeddings equates to that of another word. They derived a probabilistically grounded definition of paraphrasing that they reinterpret as word transformation, a mathematical description of "w_x is to w_y". From these concepts they proved that the existence of linear relationships between W2V-type embeddings that underlie the analogical phenomenon, identifying explicit error terms. [2] addressed the OOV problem for low resource SMT by paraphrasing with word embeddings and semantic lexicons. They proposed using semantic lexicons including WordNet, FrameNet, and the Paraphrase Database (PPDB) for paraphrasing. In addition, they applied a method to combine these two types of paraphrases, which achieved further improvements in SMT. OOV paraphrasing that augments the translation model for the OOV words by using the translation knowledge of their paraphrases has been proposed to address the OOV problem. In this paper, authors proposed using word