Istanbul Technical University - Science and Letters Faculty

Mathematics Engineering Program

Graduation Project

Student Name: Arif Çakır

Student Number: 090190355

Course: MAT4091E

Advisor: Prof. Dr. Atabey Kaygun

Submission Date: May 21, 2023

Contents

1 Introduction

Language is a fundamental aspect of human communication and one of the most common ways to pass on information and culture throughout history. Due to its nature of encapsulating information, language is studied by linguists throughout the years. Thanks to this scientific foundation, machine learning, and computer science are also advanced upon this topic. Natural Language Processing (NLP) is the subfield of machine learning that studies the processing of human language by computers. One of the most popular NLP techniques in recent years is Word Embedding, which represents words as vectors in semantic space that allows applying mathematical operations on them to analyse. Word Embedding can be used on various subjects such as text classification, translation, and sentiment analysis with promising results. It is aimed in this paper to study several Word Embedding techniques and their applications. The aim is to create a model by applying various Word Embedding models and with the help of this model, describing the relationships and similarities of languages.

The rest of the paper is organised as follows: In Section 2, information about Word Embedding in general and several Word Embedding techniques are provided. Word Embedding techniques touched on in the paper are Term Frequency (TF), Inverse Document Frequency (IDF), Word2Vec, Global Vectors for Word Representation (GloVe), Bidirectional Encoder Representations from Transformers (BERT). This is followed by Section 3 which introduces the dataset Divine Comedy by Dante Aligheri. After that, several Word Embedding models are applied to the dataset in Section 4, and the outputs of those models are analysed in Section 5. Finally paper ends with some discussion and a conclusion in Section 6.

2 Word Embedding Techniques

Word Embedding is one of the most popular natural language processing techniques due to its vector representation of the words. As Agarwal stated, capturing semantic meaning of the words in a vector of text is the ambition of the word embedding techniques (Agarwal,

2022). With respect to that, some of the most popular word embedding techniques will be studied in this section. Those techniques are TF, which counts rarity of words; IDF, which counts rarity of words; Word2Vec, which uses cosine similarity; GloVe, which captures co-occurrence of words; and BERT, family of masked-language models introduced by Google.

2.1 Term Frequency - Inverse Document Frequency

Term Frequency (TF) is a word embedding technique which counts the occurrence of the words in a document. TF can be shown as following:

$$TF(i) = \frac{\log(Frequency(i,j))}{\log(TotalNumber(j))}$$

where $Frequency(i,j)$ is the frequency of a word that occurred in a j word document and $TotalNumber(j)$ is the total number of the words in the document.

On the other hand, Inverse Document Frequency (IDF) is practically the opposite of the TF method. In this method, algorithm relies on the information that gained from the words which are rarely used. IDF can be shown as following:

$$IDF(i) = \log(\frac{TotalNumber(j)}{Frequency(j,i)})$$

where $Frequency(i,j)$ is the frequency of a word that occurred in a j word document and $TotalNumber(j)$ is the total number of the words in the document.

TF-IDF mainly shows the degree of relevancy of word i in the document j. As Kınık and Güran stated, TF-IDF does not capture semantic relationship between words, and accepts them as independent values (Kınık and Güran,2021). Due to TF-IDF's lack of capturing semantic relationship of the words, TF-IDF mainly used to detect stop words.

2.2 Word2Vec

Word2Vec is a word embedding model that was published by researchers led by Tomáš Mikolov at Google in 2013. As stated in Gensim documentations (n.d.), words are embeded in lower-dimensional vector space by Work2Vec. In this vector space, vectors which have

similarity of context between them are close to each other while words that have different meanings are distant to each other. After that, Word2Vec uses cosine similarity metric to measure similarity of the words. If cosine value of two words is 0, then words do not hold similarity. If cosine value of two words is 1, then the words are overlapping. There are two versions of Word2Vec: Skip - Gram and Continuous Bag of Words.

2.2.1 Skip-Gram Algorithm

One of the Word2Vec model architectures is Skip-Gram Algorithm. Dive Into Deep Learning (n.d) states that a word can be used for generating its surrounding words in skip-gram model. For example, if the sentence "the man loves his son" is taken and "loves" is chosen as the center word, then skip-gram model considers the conditional probability for generating the words. Due to this approach, each word has a two dimensional vector representations in skip-gram model. According to Dive Into Deep Learning, for any word with index i in the directory, $\mathbf{v}_i \in \mathbb{R}^d$ and $\mathbf{u}_i \in \mathbb{R}^d$ are its two vector representations, where $\mathbf{v}_i$ is when the word is used as the "center word" and $\mathbf{u}_i$ is when the word is used as the "context word". If w_o is any context word and w_c is given center word, then conditional probability of generating w_o can be modelled as following:

$$P(w_o|w_c) = \frac{exp({\mathbf{u}_o}^\top \mathbf{v}_c)}{\sum_{i \in \mathcal{V}} exp({\mathbf{u}_i}^\top \mathbf{v}_c)}$$

where $\mathcal{V}$ is the vocabulary index set. If a text sequence of length T is given and word at time step t is denoted as w^t, then likelihood function of skip-gram model for context widow size m is as following:

$$\prod_{t=1}^{T} \prod_{-m \leq j \leq m, j \neq 0} P(w^{t+j}|w^t)$$

2.2.2 Continuous Bag of Words

The algorithm that does exact opposite of the Skip-Gram Algorithm is called Continuous Bag of Words (CBOW)...

2.3 Global Vectors for Word Representation

Global Vectors for Word Representation (GloVe)

2.4 Bidirectional Encoder Representations from Transformers

Created by researchers at Google in 2018, Bidirectional Encoder Representations from Transformers (BERT) is a family of masked-language models...

3 Dataset

3.1 Information About the Dataset

3.2 Preparation of the Dataset

4 Applying Models to the Dataset

4.1 TF-IDF

4.2 Skip-Gram Algorithm

4.3 Continuous Bag of Words

5 Analysis and Visualization

6 Discussion