Exploring word embedding models through Word2vec and GloVe

Technology Review – CS 410 Text Information Systems Fall 2021 – **sangjin2@illinois.edu**

Introduction

As a part of natural language processing (NLP), word comprehension, semantic structure, and language meaning amongst other language structures all play a part in defining text analysis. As covered in the course (CS410), one of the major downsides to text searching is the inability to distinguish word representations when in no specific order. The process of vector modeling can alleviate some of this confusion, providing a view of words as vectors in three-dimension space in a process known as word embedding. Of the models used in this process, two primary ones to explore are Word2vec and GloVe.

Word2Vec

An architectural predictive modeling for embedding words, Word2Vec utilizes vector per word, specialized in using two functions to carry out its process -viz. encoding and decoding. In a simplified case, the encoder allows for the word to be embedded. The decoder allows embedding following some contextual cues. As a more complex definition, Word2Vec embeds using training design over a specified corpus. This training is done incrementally, allowing training through use of a neutral network.

One interesting specification of Word2Vec comes from the Skip-gram model, a modeled vector representation focusing on specific relationships between semantic and syntactic structures.

GloVe

In the case of GloVe, it is a variant of the Word2Vec model. Similarly, it utilizes vectors per word, except in its process it works to model occurrences of words on a bigger scale using matrix representation of a corpus. Since the matrix is combinatorial, the matrix representation of context tends to be quite large.

Conclusion

Modeling structures following GloVe and Word2Vec provide interesting contextual representations of word structures that allow them to be represented in vector space. Being able to specify context for appearance is an important regular for Word2Vec while GloVe use of corpus for linear structures in vector space can provide viewpoints not found in the other. By utilizing these processes, a deeper understanding of not only NLP can be obtained but also the underlying structure of their representation using a three-dimensional model that can carry another perspective on language and semantic structures.

References

https://www.cs.toronto.edu/~lczhang/360/lec/w06/w2v.html

https://medium.com/analytics-vidhya/word-embeddings-in-nlp-word2vec-glove-fasttext-24d4d4286a73

https://nlp.stanford.edu/pubs/glove.pdf

https://proceedings.neurips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf