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 constituents 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 Word-to-Vector (Word2vec) and Global Vectors (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 (Mikolov). The usage of neutral networks specifically for Skip-gram allows for faster training versus regular methods as it doesn't use matrix manipulation. This makes it quite a bit more efficient than regular methods of modeling, using pattern recognition and linear motions movements at its disposal. A vector comprised of New York and United States with Quebec would quickly pick up Canada as the closest vector mark in linear modeling. Another example can be seen in Figure 1(thinkinfi), where the left set can be used to find a complementary set to a word target (on the right). As a wider representation, Skip-gram excels at analogical tendencies,

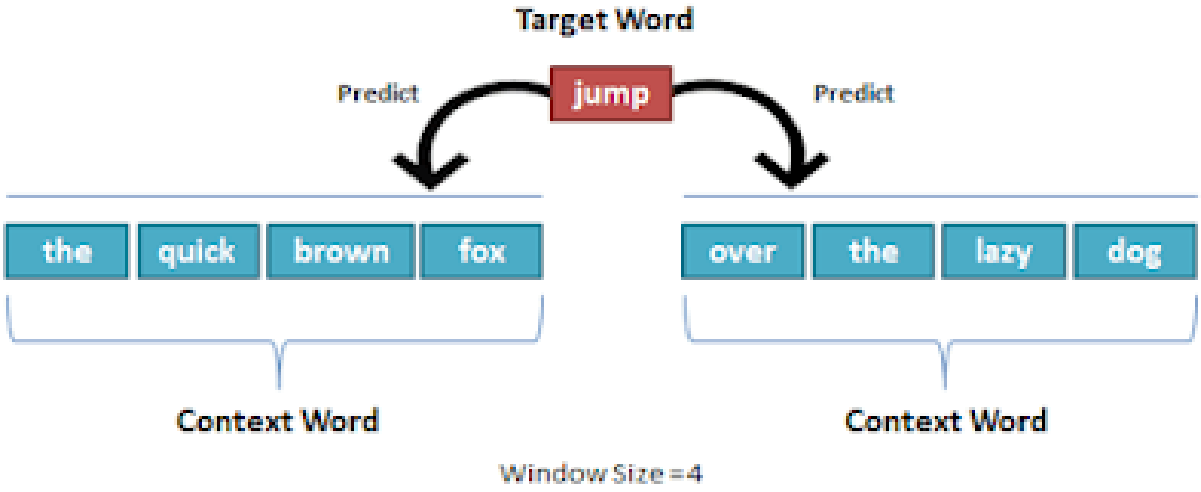

Figure 1

being able to pair or predict word representations. By utilizing this method alongside matrix-vector operations from other models, it can complement existing approaches to word embedding (Mikolov).

GloVe

In the case of GloVe, it is a variant of the Word2Vec model. Similarly, it utilizes vectors per word and recognition of corpus, 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. What allows GloVe to separate from Word2Vec in terms of result processing comes from negative sampling, where a weight change to a smaller sample compared to a whole in processing. It is like analogies, where terms aren't taken as two sets but rather the probability of their occurrence. Through Word2Vec, the process of sampling occurs slower over that of GloVe, which not only maintains speed of calculation and faster result over a sample (Pennington) but does so overall regardless of the speed of the process.

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 speed and accuracy on data points not found in the former. By utilizing these processes, a deeper understanding of not only NLP can be obtained but also the underlying structure of their representation using vectors and matrix modeling that can carry another perspective on language and semantic structures.

References

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

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Distributed Representations of Words and Phrases and their Compositionality*.

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

Pennington, J., Socher, R., & Manning, C. (2014). *GloVe: Global Vectors for Word Representation*.

https://thinkinfi.com/word2vec-skip-gram-explained/