GloVe

Global Vectors for Word Representation

Rudranil Deb Roy (rdebroy2@illinois.edu)

Introduction

Word embedding is a technique which finds extensive usage in in Natural Language Processing. Different types of word embeddings represent words as real-valued vectors in a predefined vector space and GloVe (Global Vectors) is one of the latest ones developed in that space in Stanford. Different aspects of GloVe and its usage is discussed in this paper.

How GloVe is different?

Word embeddings allow words with similar meaning to have a similar representation and this is the basis of impressive performance of deep learning methods on challenging natural language processing problems. Each word is mapped to one vector and the vector values are learned in a way that resembles a neural network. The distributed representation is learned based on the usage of words and words that are used in similar ways result in having similar representations. Word2Vec which is a de facto standard in this space is a statistical method for efficiently learning a standalone word embedding from a text corpus, whereas *GloVe can be viewed as an extension of Word2Vec collecting word representations training on aggregated global word-word co-occurrence statistics from a corpus.* GloVe combines the global statistics of matrix factorization techniques like LSA (Latent Semantic Analysis) with the local context-based learning in word2vec. There are other models like FastText representing each word as an n-gram of characters.

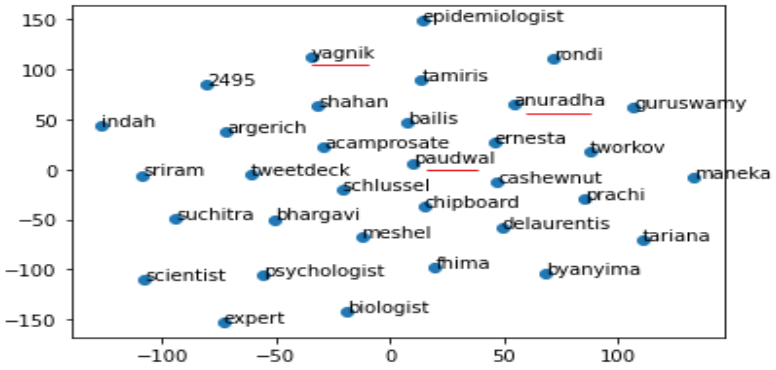
As explained above GloVe is a vector representations for words where training is performed on aggregated global word-word co-occurrence statistics from a corpus resulting in interesting linear substructures of the word vector space. From a higher level generalization this can be stated as word representations having similarity relations between words in a corpus. Nearest neighbors can be measured using the

Euclidean distance (or cosine similarity) depicting the linguistic or semantic similarity of the corresponding words. Evaluations of similarity metrics of the words involves producing a simplistic single scalar that quantifies the relatedness of two words. Simply put, *GloVe enables transformation and representation of words into a position in high-dimensional space from a corpus of text placing similar words closer to each other.*

GloVe – Closer Look

Pre-trained vectors GloVe can be downloaded from many sites which can be used and for convenience conversion libraries like *gensim glove2word2vec* are available. For higher level applications, training own GloVe vectors is recommended in the learning model itself. Effectivity of GloVe (tried by me on `glove.6B.50d.txt`) is simple and interesting as

- Closest words to "`researcher`" finds `['scientist', 'expert', 'psychologist', 'epidemiologist', 'biologist']`
- Closest words to "`paudwal`" finds `['anuradha', 'argerich', 'bailis', 'delaurentis', 'prachi']`. The first related word is `anuradha` and "`anuradha paudwal`" was a famous Bollywood singer.
- As words are vectors, simple math operation can be applied. Closest word search for "`anuradha`" + "`paudwal`" returns `['paudwal', 'anuradha', 'yagnik', 'fhima', 'maneka']`, where `yagnik` is the surname of another Bollywood singer during that time, which reveals an interesting relation between words.
- Embeddings can be easily visualized (t-SNE) by plotting them in a 2D dimensional space showing how words are placed in the space (simplified one below)



The observations above shows the efficacy of GloVe in word representation and its finding more and more extensive usage in Natural Language Processing and related applications.

GloVe – Key idea behind implementation

One key idea behind GloVe is that - *semantic relationships between words can be derived from the co-occurrence matrix*. If a given a corpus has N words, then the co-occurrence matrix X will be N x N matrix, where the ith row and jth column of X, Xij denotes how many times word i has co-occurred with word j. GloVe is a co-occurrence-based model which starts by going through the entire corpus and constructing a co-occurrence matrix as mentioned above. While Word2vec represent words by trying to predict context words from a center word or vice versa, GloVe learns by looking at pair of words at a time in the corpus and the co-occurrence of the pair. Below is the objective function which GloVe uses to train word vectors from the co-occurrence matrix.

$$J(\theta) = \frac{1}{2} \sum_{i,j=1}^{W} f(P_{ij})(u_i^T v_j - \log P_{ij})^2$$

(* Image from source: https://trailhead.salesforce.com/)

The function - given that i and j are two words that co-occur, one can optimize a word vector by minimizing the difference between the dot product of the word vectors for i and j, and the log of the number of times i and j co-occur, squared.

I think, GloVe vectors plays a major role in text mining viz. in topic modelling, sentiment analysis

GloVe – Use in Applications

GloVe is one of embedding method that is majorly used to define or extract the co-occurrence of two related data especially in unsupervised learnings based on word co-occurrence statistics in various applications like Health care which I got introduced to in DLH course of UIUC. The project was about a neural way to predict length of stay using MIMIC III database merging physiological data and clinical

notes. In this project we compared embeddings as in Word2Vec, Sent2Vec, BERT and GloVe while processing unstructured clinical notes from doctors, nurses and other sources. This helped me understand the basic concepts and application of word embeddings, the foundation on which I improved further for this paper on GloVe.

References

https://nlp.stanford.edu/projects/glove/

https://www.mygreatlearning.com/blog/word-embedding/#sh4

https://trailhead.salesforce.com/

https://towardsdatascience.com/light-on-math-ml-intuitive-guide-to-understanding-glove-embeddings-b13b4f19c010

https://medium.com/analytics-vidhya/basics-of-using-pre-trained-glove-vectors-in-python-d38905f356db