Catherine Parker
cph7@illinois.edu
CS 410
November 7, 2021

Technology Review: Word2vec

Introduction

This tech review will review the Word2vec natural language processing technique. Word2vec was originally developed by a team led by Tomas Mikolov at Google, with two papers describing it published in 2013. [1] [2] At its heart, Word2vec is a technique for transforming words in a body of text (typically large) into vectors in a vector space. The word vectors can then be used to analyze or predict text.

Explanation and Analysis

At a high level, Word2vec is a technique for transforming words in natural language into word embeddings, which are representation of words that capture the similarity to and relationships with other words. These vectors can then be mathematically manipulated for a variety of uses. A famous example of using math to analyze words is the equation KING – MAN + WOMAN = QUEEN[3]. We can represent these four words with vectors. As an example, if we have a vector that gives values for [gender power aristocracy height], we could assign the words the following values:

KING: [1 0.9 0.9 0.85]
MAN: [1 0.1 0.05 0.8]
WOMAN: [-1 0.09 0.05 0.65]
QUEEN: [-1 0.85 0.9 0.65]

Then the equation KING - MAN + WOMAN = [1 0.9 0.9 0.85] – [1 0.1 0.05 0.8] + [-1 0.09 0.05 0.65] = [-1 0.89 0.9 0.7], which is very close to the value of QUEEN ([-1 0.85 0.9 0.65] vs [-1 0.89 0.9 0.7]), and, in our vocabulary of four words, QUEEN is the closest word to this value.

Word2vec creates similar embeddings as vectors but uses many more words and much larger vectors. It creates the word embeddings by using training data (generally, a large corpus of text), and generally trains by looking at words close to other words in the text. There are two different models used for training the words. Both use neighboring words within a certain window around the target word. For instance, consider the sentence “Dorothy lived in the midst of the great Kansas prairies, with

---

[1] Mikolov, Tomas; et al. (2013). "Efficient Estimation of Word Representations in Vector Space". arXiv:1301.3781

[2] Mikolov, Tomas (2013). "Distributed representations of words and phrases and their compositionality". *Advances in Neural Information Processing Systems*. arXiv:1310.4546

[3] https://jalammar.github.io/illustrated-word2vec/

Uncle Harry, who was a farmer, and Aunt Em, who was the farmer's wife."[4] In this sentence, the technique would look at words in a "window". In this case, windows such as [Dorothy lived in], [lived in the], [in the midst], etc. would be created.

One model used for training is the Continuous Bag of Words (CBOW), which works by guessing the target word by using the words around the target, e.g. using the words "Dorothy" and "in" to guess the word "lived". The other model is the Skip-gram, which works in the opposite way, by using the target word to guess the words around it, e.g. using the word "lived" to guess the words "Dorothy" and "in". In both cases, the data is used in a neural network to create an output vector which is then used to calculate the similarities between words, usually using the cosine similarity of the vectors[5]. In practice, negative sampling must also be used to supply more accurate training data; these samples can be generated by giving a low value to the target word and a randomly selected word from the vocabulary. Generally, the skip-gram model is better for less frequent words, but is slower, whereas the CBOW model is faster and better for more frequent words, according to Mikolov[6].

Applications and Uses

Word2vec was originally developed by a team at Google to improve the language model and to account for multiple degrees of similarity between words.[7] Since then, word2vec has been used for a variety of tasks. One major use of Word2vec has been to use it not on text, but on other types of data. For instance, Word2vec was used by Airbnb to train on the user clicks made in conversive sessions (the final session in which a user makes a booking) to make a model that personalizes Search Ranking and Similar Listing Recommendations to drive conversions[8]. The website Alibaba did a similar thing, using Node2vec, an extension of Word2vec, to improve product recommendations[9].

Word2vec has also been used for knowledge discovery. One example of this is the use of abstracts from the paper *Nature* to determine properties of chemical compounds simply by using word associations between different compounds, and to predict the properties of certain compounds even before these properties had been discovered scientifically. For example, the thermoelectric properties of the compound $CuGaTe_2$ would have been able to be predicted from word association information four years before a paper about these properties was published in 2012[10].

Conclusion

---

[4] https://www.gutenberg.org/cache/epub/55/pg55-images.html

[5] https://towardsdatascience.com/introduction-to-word-embedding-and-word2vec-652d0c2060fa

[6] https://code.google.com/archive/p/word2vec/

[7] Mikolov, Tomas; et al. (2013). "Efficient Estimation of Word Representations in Vector Space". arXiv:1301.3781

[8] M. Grbovic & H. Cheng, *Real-time personalization using embeddings for search ranking at Airbnb, 2018.* https://dl.acm.org/doi/10.1145/3219819.3219885

[9] J. Wang et. al., *Billion-scale commodity embedding for e-commerce recommendation in Alibaba, 2018.* https://arxiv.org/abs/1803.02349

[10] V. Tshitoyan et. al., *Unsupervised word embeddings capture latent knowledge from materials science literature, 2019.* https://www.nature.com/articles/s41586-019-1335-8

In conclusion, the Word2vec technique has been used very effectively for word embedding, and has been extended to embed not only words but also other data. It is a powerful implement that can be used to explore word relations, and from there, these associations can be used to perform many types of analysis. I believe that knowledge and understanding of Word2vec is an important part of a toolkit of anyone who is interested in doing text analysis.