EGU2020-9966
https://doi.org/10.5194/egusphere-egu2020-9966
EGU General Assembly 2020

Exploring Earth Science Applications using Word Embeddings

Derek Koehl[1], Carson Davis[2], **Rahul Ramachandran**[2], Udaysankar Nair[1], and Manil Maskey[2]
[1]University of Alabama in Huntsville, Earth System Science Center, United States of America
[2]National Aeronautics and Space Administration, Marshall Space Flight Center, United States of America

Word embedding are numeric representations of text which capture meanings and semantic relationships in text. Embeddings can be constructed using different methods such as One Hot encoding, Frequency-based or Prediction-based approaches. Prediction-based approaches such as Word2Vec, can be used to generate word embeddings that can capture the underlying semantics and word relationships in a corpus. Word2Vec embeddings generated from domain specific corpus have been shown in studies to both predict relationships and augment word vectors to improve classifications. We describe results from two different experiments utilizing word embeddings for Earth science constructed from a corpus of over 20,000 journal papers using Word2Vec.

The first experiment explores the analogy prediction performance of word embeddings built from the Earth science journal corpus and trained using domain-specific vocabulary. Our results demonstrate that the accuracy of domain-specific word embeddings in predicting Earth science analogy questions outperforms the ability of general corpus embedding to predict general analogy questions. While the results are as anticipated, the substantial increase in accuracy, particularly in the lexicographical domain was encouraging. The results point to the need for developing a comprehensive Earth science analogy test set that covers the full breadth of lexicographical and encyclopedic categories for validating word embeddings.

The second experiment utilizes the word embeddings to augment metadata keyword classifications. Metadata describing NASA datasets have science keywords that are manually assigned which can lead to errors and inconsistencies. These science keywords are controlled vocabulary and are used to aid data discovery via faceted search and relevancy ranking. Given the small size of the number of metadata records with proper description and keywords, word embeddings were used for augmentation. A fully connected neural network was trained to suggest keywords given a description text. This approach provided the best accuracy at ~76% as compared to other methods tested.