EGU22-3131
https://doi.org/10.5194/egusphere-egu22-3131
EGU General Assembly 2022

Language model for Earth science for semantic search

Rahul Ramachandran[1], Muthukumaran Muthukumaran Ramasubramanian[2], Prasanna Koirala[2], Iksha Gurung[2], and Manil Maskey[1]
[1]NASA/MSFC, Huntsville, United States of America (rahul.ramachandran@nasa.gov)
[2]University of Alabama in Huntsville

Recent advances in technology have transformed the Natural Language Technology (NLT) landscape, specifically, the use of transformers to build language models such as BERT and GPT3. Furthermore, it has been shown that the quality and the domain-specificity of input corpus to language models can improve downstream application results. However, Earth science research has minimal efforts focused on building and using a domain-specific language model.

We utilize a transfer learning solution that uses an existing language model trained for general science (SciBERT) and fine-tune it using abstracts and full text extracted from various Earth science journals to create BERT-E (BERT for Earth Science). The training process utilized the input of 270k+ Earth science articles with almost 6 million paragraphs. We used Masked Language Modeling (MLM) to train the transformer model. MLM works by masking random words in the paragraph and optimizing the model for predicting the right masked word. BERT-E was evaluated by performing a downstream keyword classification task, and the performance was compared against classification results using the original SciBERT Language Model. The SciBERT-based model attained an accuracy of 89.99, whereas the BERT-E-based model attained an accuracy of 92.18, showing an improvement in overall performance.

We investigate employing language models to provide new semantic search capabilities for unstructured text such as papers. This search capability requires utilizing a knowledge graph generated from Earth science corpora with a language model and convolutions to surface latent and related sentences for a natural language query. The sentences in the papers are modeled in the graph as nodes, and these nodes are connected through entities. The language model is used to give sentences a numeric representation. Graph convolutions are then applied to sentence embeddings to obtain a vector representation of the sentence along with combined representation of the surrounding graph structure. This approach utilizes both the power of adjacency inherently encoded in graph structures and latent knowledge captured in the language model. Our initial proof of concept prototype used SIMCSE training algorithm (and the tinyBERT architecture) as the embedding model. This framework has demonstrated an improved ability to surface relevant, latent information based on the input query. We plan to show new results using the domain-specific BERT-E model.