

LANGUAGE MODEL FOR EARTH SCIENCE: EXPLORING POTENTIAL DOWNSTREAM APPLICATIONS AS WELL AS CURRENT CHALLENGES

R. Ramachandran¹, M. Ramasubramanian², P. Koirala², I. Gurung², and M. Maskey¹

¹NASA/MSFC

²The University of Alabama in Huntsville

ABSTRACT

The use of deep learning techniques to build transformer language models such as SciBERT and GPT3 have transformed the natural language technology (NLT) landscape. These new NLTs are being used in speech to text and vice versa, automated text classification, sentiment analysis, topic modeling, text summarization, and cognitive assistants. While Earth science has no shortage of unstructured data such as journal and conference papers, little efforts have focused on harnessing NLTs for knowledge extraction and supporting the scientific process.

This paper surveys the use of language models in different science. BERT-E, a new Earth science-specific language model, is presented. BERT-E is generated using a transfer learning solution. A language model that has already been trained for general Science (SciBERT) is fine-tuned using abstracts and full text extracted from various Earth science-related articles. A downstream keywords classification application is used for evaluation, and the use of BERT-E shows improved performance. The need to develop a robust set of benchmarks in evaluating the language model such as BERT-E is discussed. Finally, example applications are presented to inspire additional ideas for applications using domain-specific language models.

Index Terms— language model, natural language, deep learning

1. INTRODUCTION

Unstructured (mainly text) data remains an untapped resource with tremendous potential insights. Recent advances in technology have transformed the Natural Language Technology (NLT) landscape, specifically, the use of deep learning techniques to build transformer language models such as SciBERT [1] and GPT3 [2]. These new NLTs are being adopted in the private industry to improve operations and services [3]. Most of the uses focus on improving customer services. Industry applications focus on the operational use of these technologies include translations from speech to text and vice versa, automated text classification, sentiment analysis for feedback, comments and financial reports, topic modeling,

text summarization, personalization and cognitive assistants (contextual chatbots). These technologies also serve as the basis for insight engines. Insight engines combine search capabilities with artificial intelligence to deliver actionable insights from the full spectrum of content and data sourced within and external to an enterprise [3]. Earth science has no shortage of unstructured data. Almost all knowledge within Earth science gets published as journal or conference papers. However, limited efforts have focused on harnessing this potential resource for knowledge extraction and supporting the scientific process. This paper examines four aspects related to language models in Earth science. First, it surveys the use of Language models in different science areas to provide context. Second, it describes BERT-E, an Earth science-specific language model. Third, the challenges of developing robust benchmarks in evaluating the language model such as BERT-E are presented. Finally, the paper explores the use of BERT-E in some ongoing prototyping efforts and future applications to support the scientific process.

There are minimal efforts in Earth science focused on building and using a language model. This paper hopes to encourage others in the informatics community to explore language models and their applications to support data and knowledge life cycles.

2. BACKGROUND

Since language models are relatively new [1][2] their applications in earth science are not as yet widespread. Examples of some applications in science are presented in this section. While the list is not comprehensive, the objective of this section is to provide an overview of the spectrum of existing downstream applications leveraging these models. Mishra and Sharma [8] describe the use of word embeddings to detect ambiguous words in science and engineering domains. They propose the feasibility of using this technique in evaluating large-scale requirements specifications to identify words in the requirements that may cause implementation teams ambiguity. Katsurai [6] surveyed word embedding applications in Library and Information science. They classified these applications into three categories. The first application category focused on knowledge extraction and visualization. Typi-

cal steps in these applications involve topic extraction from documents and then clustered based on the t-SNE algorithm. The second category focused on the classification of both extracted entities and topics. Their third category was catchall with various applications ranging from sentiment analysis, opinion classification, new word detection, and query expansion. Tshitoyan et al. [12] paper exemplifies machine-assisted scientific breakthroughs using scientific literature. Authors show knowledge present in the material science scientific literature can be encoded as word embeddings. These embeddings can then be used to recommend materials for functional applications several years before their discovery. Their paper shows that a ranking of thermoelectric materials can be produced using cosine similarities of material embeddings. Highly ranked materials that have not been studied for thermoelectric application can be considered as potential predictions for future tests. Tshitoyan et al. [12] observe that the quality and domain specificity of the scientific literature as input to the embedding matters more than just the size of the corpus. Ramachandran et al. [9] explained a data-driven approach to build an embedding from a large corpus of Earth science journal publications and using the embedding to provide semantic support. The paper describes two such applications within the data and information system. The first application provides a semantic mapping service to bridge the gap between a user’s science application need and the appropriate instruments or datasets required to study that need. The second application uses embedding to augment a keyword recommender classifier to make the data set tagging process efficient for the data stewards creating metadata to ensure keyword consistency within a data catalog. More recently, Tassopoulou et al. [11] developed a pipeline for automatic text generation to produce informative and accurate descriptions of natural hazards and phenomena. The authors used a transformer-based model to construct a brief narrative describing the natural events based on event triggers and remote sensing imagery, natural event metadata, and other ancillary descriptive material.

3. BERT-E: EARTH SCIENCE SPECIFIC LANGUAGE MODEL

Yu Gu et al.[4] and Tshitoyan et al. [12] both argue quality and the domain-specificity of input corpus to language model improve the result. With that as the driver, we utilized a transfer learning solution that uses a language model that has already been trained for general science (Sci-BERT) as our basis. We fine-tuned Sci-BERT using abstracts and full text extracted from various Earth science-related articles to create BERT-E [7] (see Figure 1). The training process utilized the input dataset of 270k+ Earth science articles with almost 6 million paragraphs. The train eval split chosen was 95% and 5%, respectively. The dataset was parsed into separate paragraphs and the Sci-BERT tokenizer was utilized to

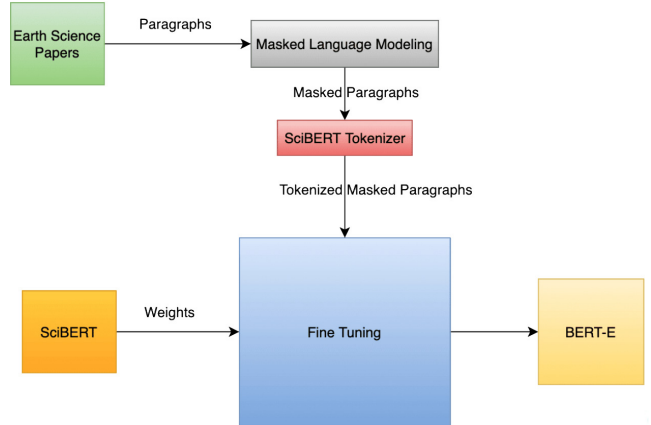


Fig. 1. BERT-E Model Development Flow

form encodings. Masked Language Modeling (MLM) was used to mask random words in the paragraph. The model trains by predicting the word and validating it against the actual masked word. BERT-E was evaluated by performing a highly domain specific document keywords classification task. GCMD keywords are a hierarchical set of controlled Earth science vocabularies used to annotate NASA data sets consistently and comprehensively to allow precise searches. A fully connected multi-class multi-label network was trained that uses BERT-E embeddings for the text as input and sigmoid activation as the output layer. The BERT-E-based keyword classification was compared against classification just using SciBERT embedding. The SciBERT based model attained an accuracy of 89.99, whereas the BERT-E based model attained an accuracy of 92.18, showing improvement in overall performance. This result is consistent with Yu Gu et al. [4] and Tshitoyan et al. [12] findings.

4. BENCHMARKING CHALLENGES AND POTENTIAL SOLUTIONS

One of the significant challenges to using these language models for downstream applications lies in the lack of a large suite of benchmarks to evaluate the robustness of the model. Our quantitative evaluation has focused on only one type of task—document keyword classification. A more comprehensive suite of shared evaluation benchmark datasets is needed to develop robust and usable models. Yu Gu et al. [4] provide a framework for evaluating the language model using benchmarks in the biomedical domain. The authors list a set of benchmark datasets covering application categories ranging from named entity recognition (NER), relationship extraction, sentence similarity matching, document classification, and question answering. Benchmark datasets covering these categories are needed in Earth science for models like BERT-E to be used confidently for future downstream applications.

We have started creating some of these benchmark sets

and have ideas to develop and share others with the Earth science community. Two of NASA’s ongoing data stewardship efforts will generate benchmark data to test relationship extraction between entities. The first one focuses on the data citation problem and manually maps a given paper to the data set and instruments used in the study [5].¹ The second is the result of our previous application to bridge the gap between a user’s science application need and the appropriate instruments or datasets required to study that need.² We have curated a dataset from different catalogs that map different phenomena of interest to instruments and platforms.³

To develop a NER benchmark, we plan to scrape and use existing glossaries in Earth science such as one provided by AMS⁴ and Oxford Reference⁵. These glossaries provide descriptive definitions of scientific terms, and these descriptions can be masked and then used as test input to the language model to predict the term. We also plan to create a question and answer benchmark dataset similar to <https://pubmedqa.github.io/> by leveraging the Earth science community’s assistance. We plan to create a website where researchers can enter a DOI for a paper, a science question, and the sentence or paragraph in that paper that addresses their question. In addition to our keyword classification, these benchmark datasets can be used to evaluate BERT-E and any future Earth science language models for their usefulness in different downstream applications.

5. POTENTIAL FUTURE APPLICATIONS

It is also important to share how we envision the use of BERT-E in different downstream Earth science applications. We list a few potential applications.

1. Semantic search for unstructured data such as publications. We have started exploring the effectiveness of combining a knowledge graph generated from earth science corpora with a language model and graph convolutions to surface latent and related sentences given a natural language query [10]. In this approach, sentences are modeled in the graph as nodes. These nodes are connected through entities—words and phrases of interest found in the text—extracted using Google Cloud’s entity extraction model. The SIMCSE (tiny-BERT) 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 and the surrounding graph structure, thereby leveraging the power of adjacency inherently encoded in graph structures.

¹https://github.com/NASA-IMPACT/snwg_demo/blob/feature-combined_instrument_platforms/data/instruments_combined.json

²https://github.com/NASA-IMPACT/snwg_demo/blob/feature-combined_instrument_platforms/data/missions_combined.json

³https://github.com/NASA-IMPACT/snwg_demo/blob/feature-combined_instrument_platforms/data/ml_data/combined_ins_attributes.json

⁴<https://glossary.ametsoc.org/wiki/Welcome>

⁵<https://www.oxfordreference.com/view/10.1093/acref/9780199211944.001.0001/acref-9780199211944>

Our initial proof of concept prototype has demonstrated an improved ability to surface relevant, latent information based on the subject of the input query. We plan to expand this prototype using the BERT-E model.

2. Automate taxonomy generation. The creation of taxonomies tends to be a manual, tedious task typically performed by a small group of individuals, even though these taxonomies represent broad domains. One approach to scale taxonomy generation is utilizing existing taxonomies as seeds and then using BERT-E to explore clusterings for terms from glossaries for different categories to identify new additions quickly.

3. Text summarization approaches for metadata generation. Language models such as BERT-E in conjunction with algorithm theoretical basis Documents (ATBDS) are used to generate summaries that can capture collection-level metadata for datasets. This application would assist metadata curators in metadata creation and improve the overall data stewardship efficiency.

The example applications are provided to inspire new and better application ideas from the Earth science community using BERT-E. BERT-E is openly available.⁶

6. REFERENCES

- [1] I. Beltagy, K. Lo, and A. Cohan, “Scibert: A pretrained language model for scientific text,” arXiv Prepr. arXiv, vol. 1903.10676, 2019.
- [2] T. B. Brown et al., “Language Models are Few-Shot Learners,” CoRR, vol. abs/2005.14165, 2020.
- [3] S. Emmott and A. Mullen, “Gartner: Magic Quadrant for Insight Engines,” no. March, pp. 1–24, 2021.
- [4] Y. Gu et al., “Domain-Specific Language Model Pre-training for Biomedical Natural Language Processing,” ACM Trans. Comput. Healthc., vol. 3, no. 1, pp. 1–23, 2022.
- [5] V. Inverso, C. Lynnes, A. Cramer, V. Chatman, and Elena Steponaitis, “Populating a Graph Database to Run a Usage-Based Discovery Tool,” in Fall American Geophysical Union Meeting, 2021.
- [6] M. Katsurui, “Using word embeddings for library and information science research,” ACM SIGWEB Newsl., no. Spring, pp. 1–7, 2020.
- [7] P. Koirala, M. Ramasubramanian, I. Gurung, M. Maskey, and Rahul Ramachandran, “BERT-E: An Earth Science Specific Language Model for Domain-Specific Downstream Task,” in Fall American Geophysical Union Meeting, 2021.

⁶<https://huggingface.co/nasa-impact/bert-e-base-mlm>

- [8] S. Mishra and A. Sharma, “On the use of word embeddings for identifying domain specific ambiguities in requirements,” Proc. - 2019 IEEE 27th Int. Requir. Eng. Conf. Work. REW 2019, pp. 234–240, 2019.
- [9] R. Ramachandran, M. Ramasubramanian, I. Gurung, C. Davis, and Derek Koehl, “AUGMENTING DATA SYSTEMS WITH PREDICTION BASED EMBEDDINGS,” in IEEE Geoscience and Remote Sensing Society, 2021.
- [10] M. Ramasubramanian, B. Kuo, R. Rahul, and I. Gurung, “Exploring Semantic Search Capability of Graph Convolutions Over a Knowledge Graph Built Using Earth Science Corpora,” in Fall American Geophysical Union Meeting, 2021.
- [11] V. Tassopoulou et al., “Generating informative and accurate descriptions of natural hazards and phenomena using large transformer-based models,” in Fall American Geophysical Union Meeting, 2021.
- [12] V. Tshitoyan et al., “Unsupervised word embeddings capture latent knowledge from materials science literature,” Nature, vol. 571, no. 7763, pp. 95–98, 2019.