



Using Large Scale Language Models for Enabling NLP Abilities Within Earth Sciences

12 December 2022

**Muthukumaran Ramasubramanian*, Iksha Gurung, Raghu Ganti, Bishwaranjan
Bhattacharya, Davis Wertheimer, Marquita Ellis, Manil Maskey, Rahul
Ramachandran**

***presenter**

UAH / NASA MSFC / IBM Watson AI

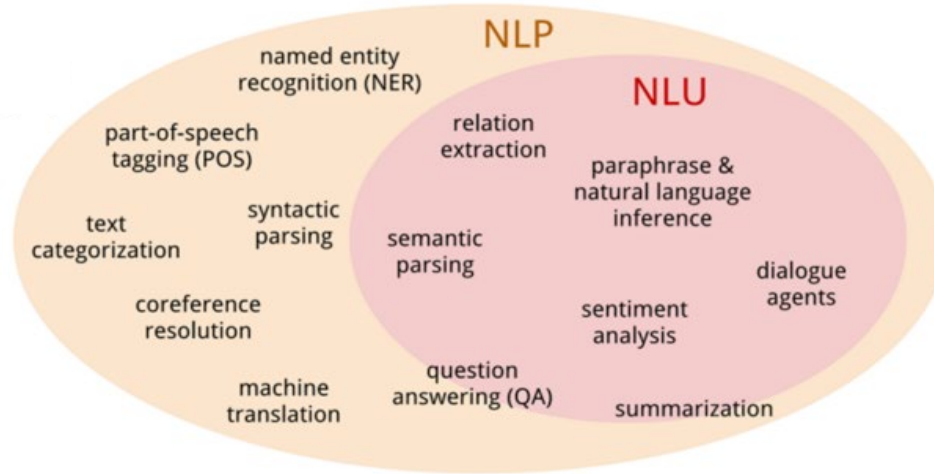




- Overview
- Earth Science Domain Adaptation
- Our Approach in Training LLM
- Current Status
- Evaluating LMs
- Future Steps and Conclusion

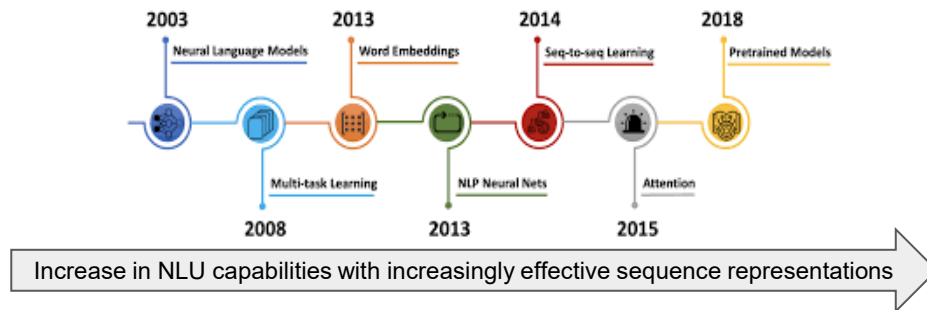
Outline

Overview - Natural Language Processing

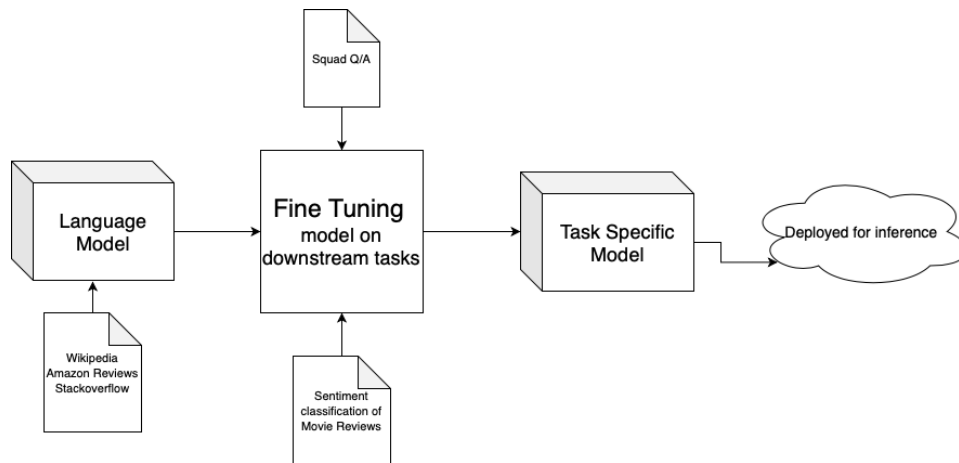


- **Natural Language Processing:** Algorithms that can process human language and perform classification, generation, semantic analysis tasks.
- **Natural Language Understanding:** Understanding the context and semantic relationships within texts.
- **Natural Language Generation:** Methodology that strings together meaningful sentences, based on the task and internal representation of the domain stochastically.

Overview - Language Models



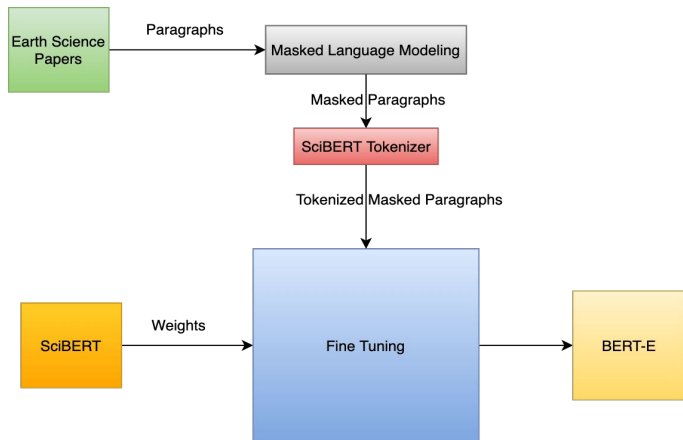
Pretraining language models



Need for LMs in Earth Science

- Availability of unstructured data
 - Exploit sizeable Earth Science Corpus available for creating LMs
- Lack of abundant labeled data in
 - Data / Information discovery activities
 - Data Stewardship
- LLMs are ***few shot learners***
 - Can we create an Earth Science LLM with strong associations between domain concepts, making it possible to train with limited label data?

Our Initial Approach



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

Why Sci-BERT for transfer learning?

- SciBERT offers good starting point for BERT-E due to the nature of training documents (Semantic Scholar)
- Training a novel tokenizer involves training a model from scratch—involves high costs for a prototype
- Offers a good indication of domain specificity improving the tasks

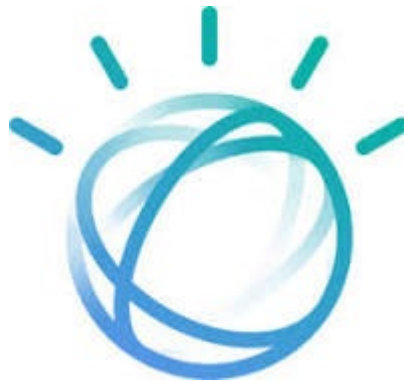
training parameters:

- Abstracts from ~ **150K ES journal papers**
- Optimized with **MLM loss** (masked language model)

Current Status

Collaboration with IBM

- Train LLMs using IBM's Foundational Model Tech Stack.
- Augment Earth Science corpus with general scientific corpus.
- Large Scale Training
 - 64 NVIDIA A100 GPUs
 - 10 hrs of training
 - 67k words/GPU/sec throughput



Evaluating LLM: Keyword Classification

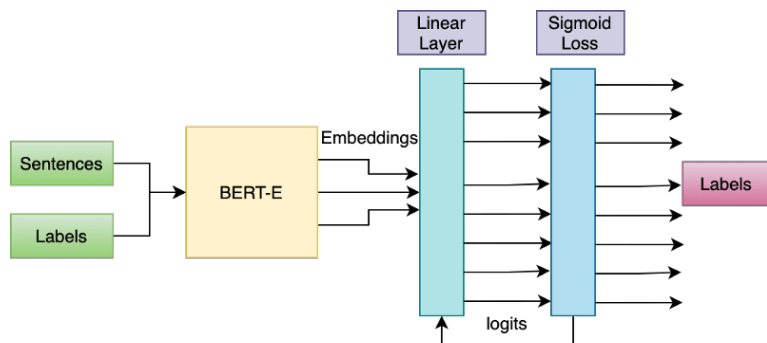
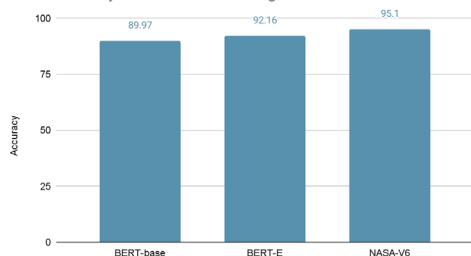


Fig. GCMR Keyword Recommender

GKR Accuracy vs. BERT Embeddings



GCMR Keyword Recommender

- GCMR → hierarchical set of Earth science keywords; used for improving discovery
- Automated tagging of keywords will help eliminate human error in tagging and help data curators make decisions based on consensus.
- Produces more semantically correct keywords not tagged by humans (indicates generalizability)

Evaluating LLM: Question Answering

PrimeQA

Question Answering

Find answers to question from retrieved evidence blocks

What would you like to know?

what does dust events arise from?

Ask

Question

what does dust events arise from?

Found 15 answers matching your question.

Answer

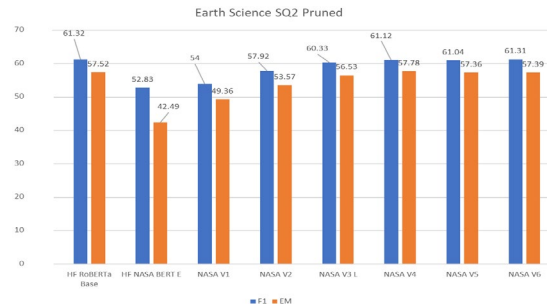
sand or silt grains are either dislodged from the surface by other windblown grains (saltation) or lifted by wind directly (direct entrainment).

Confidence: 50%
Was this answer useful?

Evidence

Dust, which affects weather and climate and can be hazardous to health, can be generated when sand or silt grains are either dislodged from the surface by other windblown grains (saltation) or lifted by wind directly (direct entrainment). Direct entrainment of silt has been thought to generate only minimal quantities of dust, and atmospheric models that include dust usually assume that dust generation occurs through saltation.

Fig. PrimeQA



Validation on SQ2 pruned benchmark (F1, EM scores)

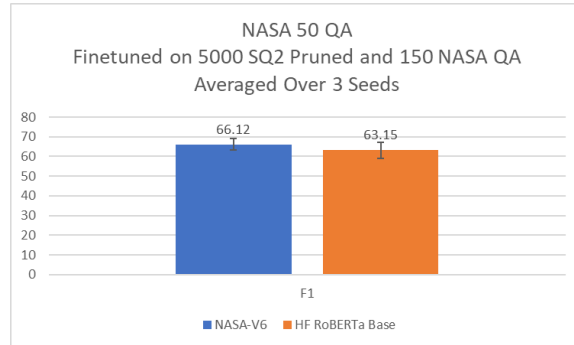


Fig. Earth Science QA

Future Steps / Conclusions

- Comprehensive Benchmarking
- Iterative Model improvement based on benchmarks

Thank you!

Contact Info:

mr0051@uah.edu

This work is supported by NASA Grant 80MSFC22M004

