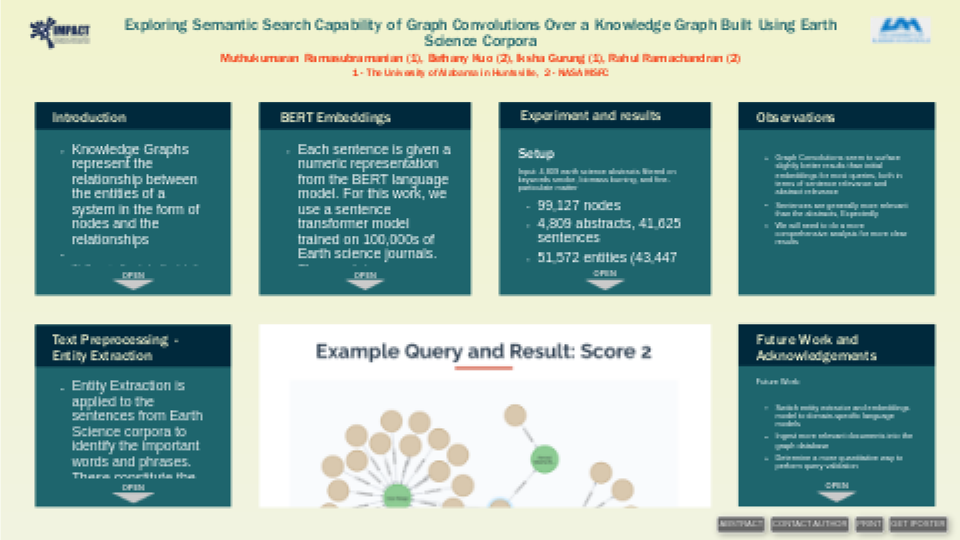
Exploring Semantic Search Capability of Graph Convolutions Over a Knowledge Graph Built Using Earth Science Corpora



Exploring Semantic Search Capability of Graph Convolutions Over a Knowledge Graph Built Using Earth Science Corpora

Muthukumaran Ramasubramanian (1), Bethany Kuo (2), Iksha Gurung (1), Rahul Ramachandran (2)

1 - The University of Alabama in Huntsville, 2 - NASA MSFC

Introduction

- Knowledge Graphs represent the relationship between the entities of a system in the form of nodes and the relationships

BERT Embeddings

- Each sentence is given a numeric representation from the BERT language model. For this work, we use a sentence transformer model trained on 100,000s of Earth science journals.

Experiment and results

Setup

Input: 4,809 earth science abstracts filtered on keywords smoke, biomass burning, and fine-particulate matter

- 99,127 nodes
- 4,809 abstracts, 41,625 sentences
- 51,572 entities (43,447

Observations

- Graph Convolutions seem to surface slightly better results than initial embeddings for most queries, both in terms of sentence relevance and abstract relevance
- Sentences are generally more relevant than the abstracts. Especially
- We will need to do a more comprehensive analysis for more clear results.

Text Preprocessing - Entity Extraction

- Entity Extraction is applied to the sentences from Earth Science corpora to identify the important words and phrases.

Example Query and Result: Score 2

Future Work and Acknowledgements

Future Work:

- Switch entity extractor and embeddings model to domain-specific language models
- Ingest more relevant documents into the graph database
- Determine a more quantitative way to perform query validation

Muthukumaran Ramasubramanian (1), Bethany Kuo (2), Iksha Gurung (1), Rahul Ramachandran (2)

1 - The University of Alabama in Huntsville, 2 - NASA MSFC

THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

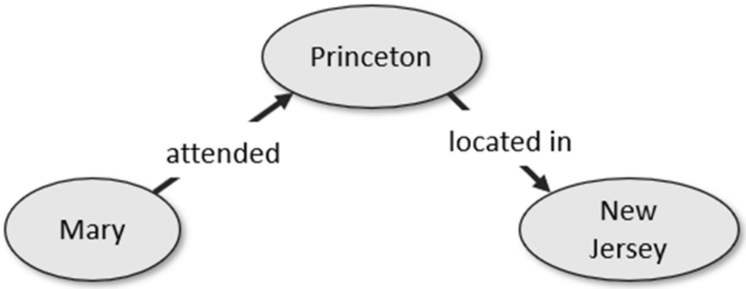
PRESENTED AT:

AGU FALL MEETING
New Orleans, LA & Online Everywhere
13–17 December 2021
Poster Gallery brought to you by
WILEY

INTRODUCTION

- Knowledge Graphs represent the relationship between the entities of a system in the form of nodes and the relationships
- We Present a Knowledge Graph built using Earth Science corpora
- We use BERT language model and graph convolutions to search the KG. A qualitative analysis of results is also presented

Knowledge Graphs(KG) are powerful tools that represent the relationship between the entities of a system in the form of nodes and the relationships (connections) between them. Building a KG can be a challenging task as they tend to be too generic, and often perform poorly on complex scientific queries. In this project, we aim to explore the effectiveness of combining a knowledge graph generated from earth science corpora with a language model and graph convolutions for the purpose of surfacing latent and related sentences given a natural language query.



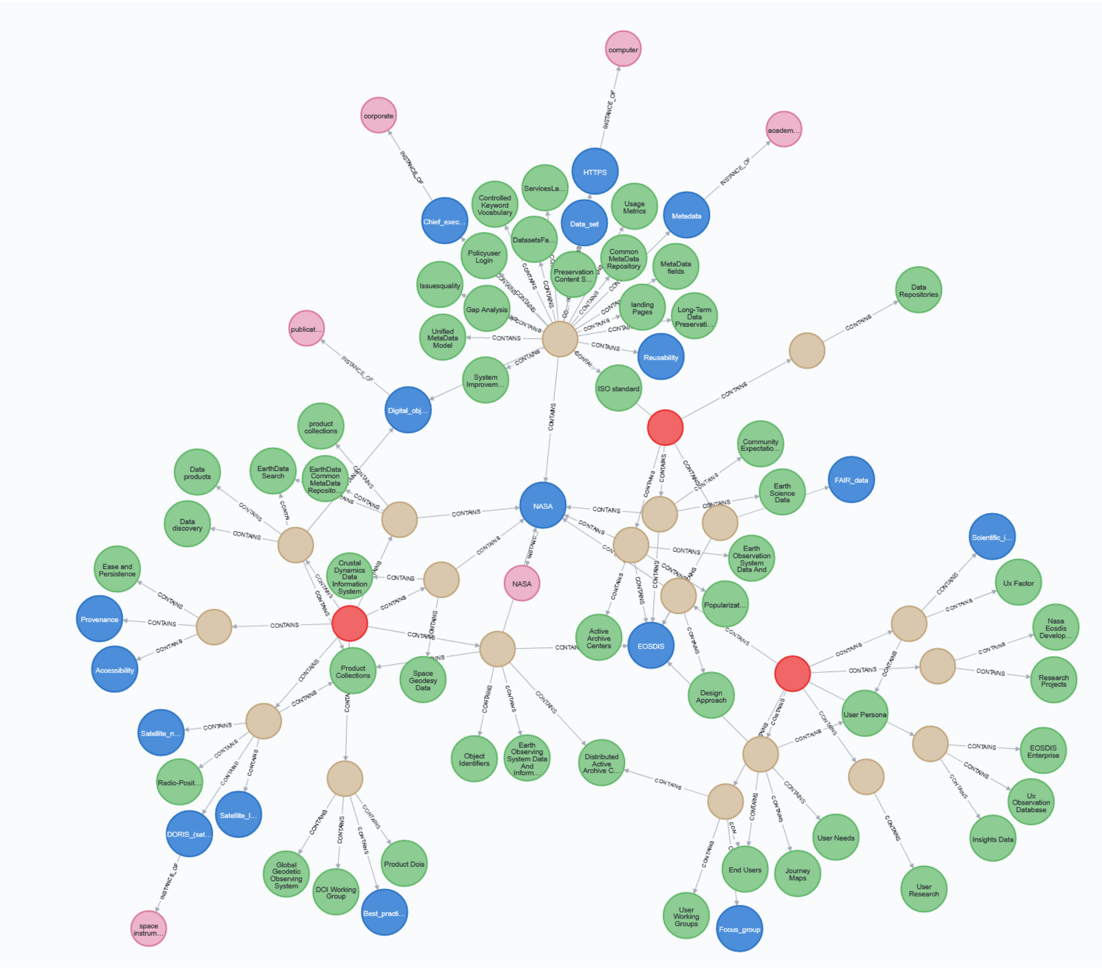
In this approach, sentences are conceptualized in the graph as nodes that are connected through entities—words of interest found in the text—extracted using Google Cloud’s entity extraction model. The language model we used for this is Bidirectional Encoder Representations (BERT). The sentences are given a numeric representation by the BERT model; graph convolutions are then applied to sentence embeddings in order to obtain a vector representation of the sentence as well as the adjacency information. The graph can then be queried with natural language queries by generating an embedding for the query then comparing it to the embeddings in the graph. The sentences with the most similar embeddings are returned as results. We find that when query embeddings are compared with convolved embeddings, topics of resulting sentences and the abstracts they come from are slightly more relevant than when the query embedding is compared with initial BERT embeddings, potentially demonstrating an improved ability to surface relevant, latent information based on the subject of the input query.

BERT EMBEDDINGS

- Each sentence is given a numeric representation from the BERT language model. For this work, we use a sentence transformer model trained on 100,000s of Earth science journals.
- The graph is implemented in a highly performant graph library called Neo4J

Scaling up with Neo4J graph database

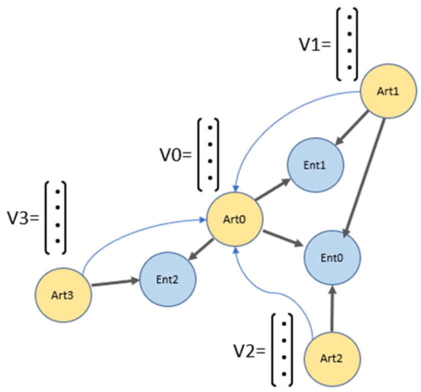
The Nodes, links (Relationships), and embeddings are ingested into a graph database for efficiency and scaling purposes. In this case, we use the Neo4J graph database library.



- Abstracts ●
  - Title
  - Source file
  - Embeddings
- Sentences ●
  - Title of source paper
  - Sentence content
  - Embeddings
- Entities ● ●
  - Entity name
  - WikiData ID
  - Wikipedia URL
- Instances ●
  - Instance name
  - WikiData ID

Graph Convolutions

Each layer of convolution aggregates neighboring node embeddings and adds them to the original embedding. Aggregation is a basic average of neighboring nodes.



$$X = \lambda * V0 + (1 - \lambda) * (V1 + V2 + V3)/3$$

$$V'0 = X/|X|$$

EXPERIMENT AND RESULTS

Setup

Input: 4,809 earth science abstracts filtered on keywords smoke, biomass burning, and fine-particulate matter

- 99,127 nodes
- 4,809 abstracts, 41,625 sentences
- 51,572 entities (43,447 normal entities and 8,125 wiki entities)
- 1,121 instances
- 210,422 relationships

Results

Score frequency distribution

Score	Initial embeddings		3rd layer of convolutions	
	Sentence	Abstract	Sentence	Abstract
0	12	21	10	17
1	17	14	15	17
2	16	10	20	11
Total	45	45	45	45

OBSERVATIONS

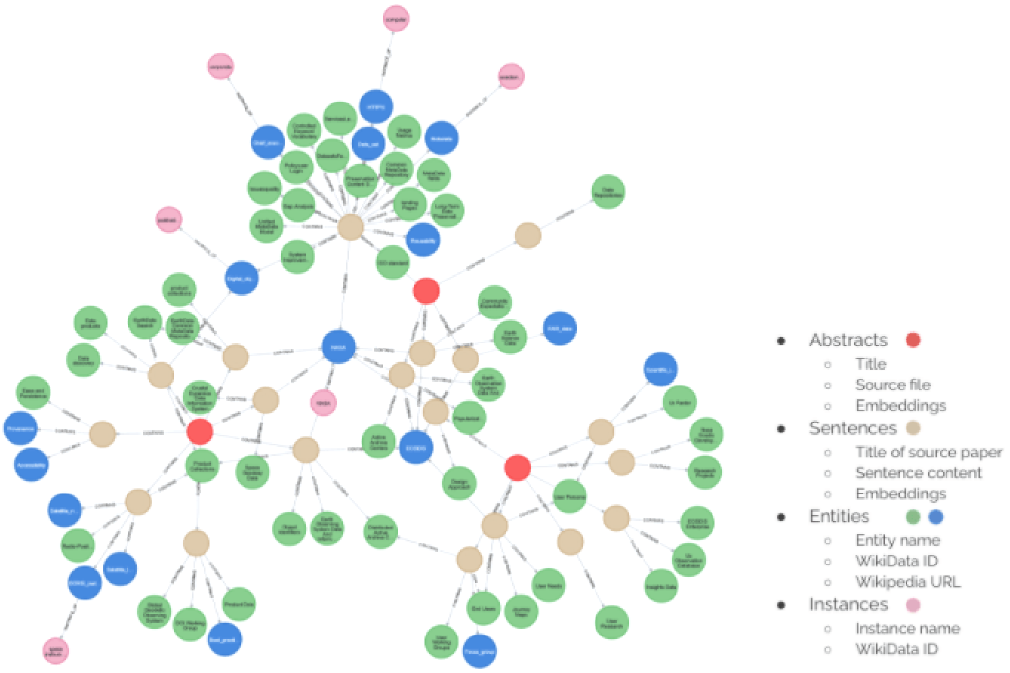
- Graph Convolutions seem to surface slightly better results than initial embeddings for most queries, both in terms of sentence relevance and abstract relevance

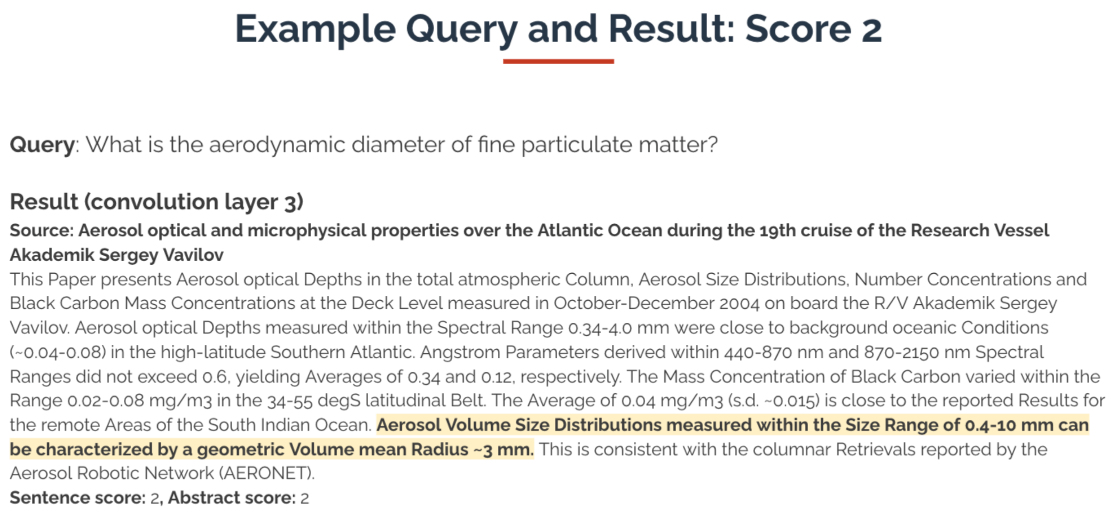
- Sentences are generally more relevant than the abstracts, Expectedly

- We will need to do a more comprehensive analysis for more clear results

TEXT PREPROCESSING - ENTITY EXTRACTION

- Entity Extraction is applied to the sentences from Earth Science corpora to identify the important words and phrases.
- These constitute the nodes of the graph. The sentences also are assigned as the sentence node.
- If the Extracted Entities are from the same sentence, they are connected with a link.
- Consequently, the sentences in the abstract are connected as well.





Example Query and Result: Score 2

Query: What is the aerodynamic diameter of fine particulate matter?

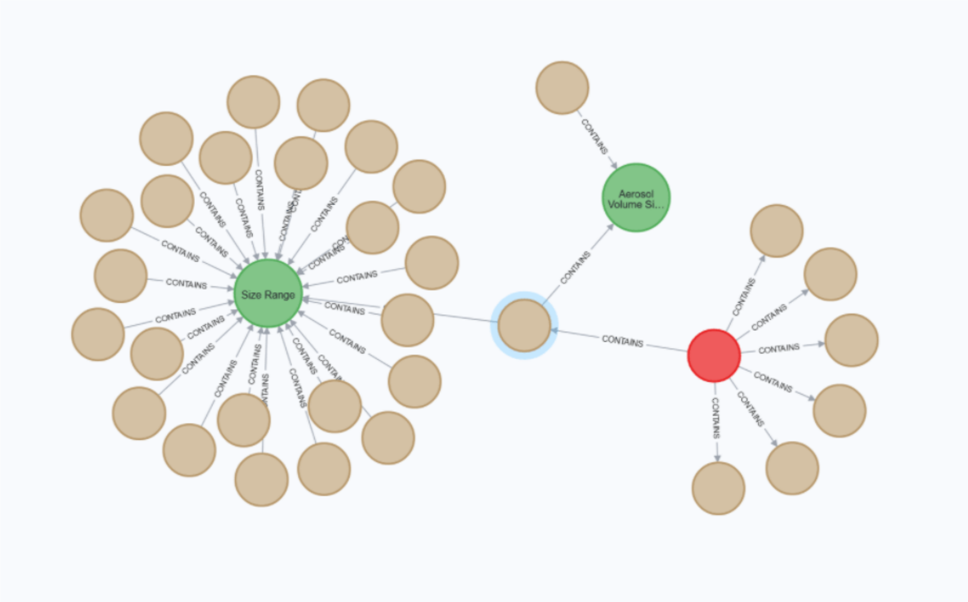
Result (convolution layer 3)

Source: Aerosol optical and microphysical properties over the Atlantic Ocean during the 19th cruise of the Research Vessel Akademik Sergey Vavilov

This Paper presents Aerosol optical Depths in the total atmospheric Column, Aerosol Size Distributions, Number Concentrations and Black Carbon Mass Concentrations at the Deck Level measured in October-December 2004 on board the R/V Akademik Sergey Vavilov. Aerosol optical Depths measured within the Spectral Range 0.34-4.0 mm were close to background oceanic Conditions (~0.04-0.08) in the high-latitude Southern Atlantic. Angstrom Parameters derived within 440-870 nm and 870-2150 nm Spectral Ranges did not exceed 0.6, yielding Averages of 0.34 and 0.12, respectively. The Mass Concentration of Black Carbon varied within the Range 0.02-0.08 mg/m3 in the 34-55 degS latitudinal Belt. The Average of 0.04 mg/m3 (s.d. ~0.015) is close to the reported Results for the remote Areas of the South Indian Ocean. **Aerosol Volume Size Distributions measured within the Size Range of 0.4-10 mm can be characterized by a geometric Volume mean Radius ~3 mm.** This is consistent with the columnar Retrievals reported by the Aerosol Robotic Network (AERONET).

Sentence score: 2, **Abstract score:** 2

Example Query and Result: Score 2



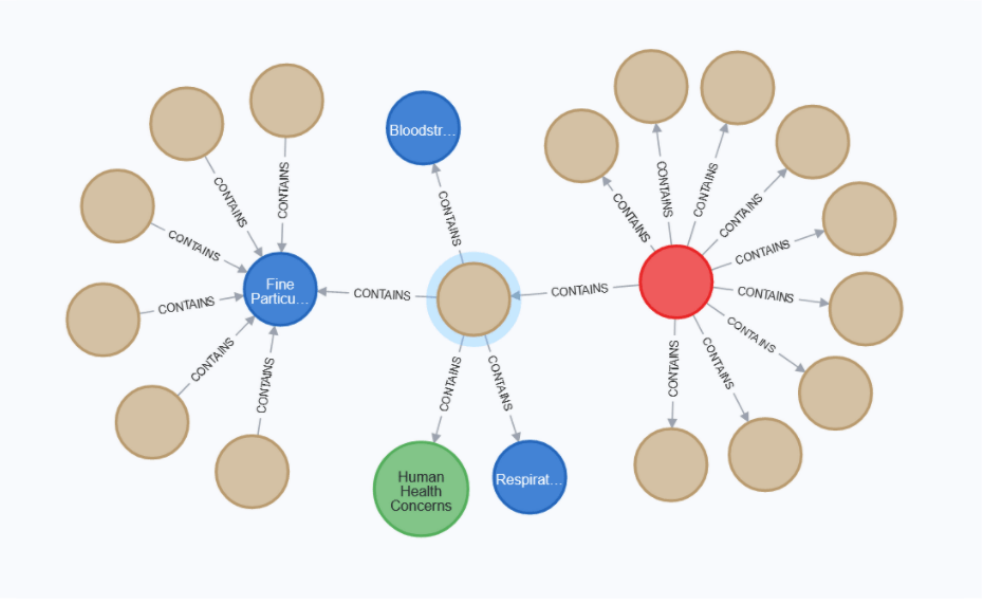
Example Query and Result: Score 1

Query: Does particulate matter lead to increased hospitalizations?

Result (convolution layer 3)

Source: Santa Ana Winds of Southern California Impact PM2.5 With and Without Smoke From Wildfires
Fine Particulate Matter (PM2.5) raises Human Health Concerns since it can deeply penetrate the Respiratory System and enter the Bloodstream, thus potentially impacting vital Organs. Strong Winds Transport and disperse PM2.5, which can travel over long Distances. Smoke from Wildfires is a major episodic and seasonal Hazard in Southern California (SoCal), where the Onset of Santa Ana Winds (SAWs) in early Fall before the first Rains of winter is associated with the region's Most damaging Wildfires. However, Saws also tend to improve Visibility as they sweep Haze Particles from highly polluted Areas far out to Sea. Previous Studies characterizing PM2.5 in the Region are limited in time span and spatial Extent, and have either addressed only a single Event in time or short time Series at a limited Set of Sites. Here we Study the space-time Relationship between daily Levels of PM2.5 in Socal and Saws spanning 1999-2012 and also further identify the Impact of Wildfire Smoke on this Relationship. We used a rolling Correlation Approach to characterize the spatial-temporal Variability of daily SAW and PM2.5. Saws tend tolower PM2.5 Levels, particularly along the Coast and in urban Areas, in the Absence of Wildfires upwind. On the other Hand, Saws markedly Increase PM2.5 in ZipCodes downwind of Wildfires. These empirical Relationships can be used to identify Windows of Vulnerability for public Health and orient preventive Measures.
Sentence score: 1, **Abstract score:** 1

Example Query and Result: Score 1



Example Query and Result: Score 0

Query: Is biomass burning intensity and frequency increasing?

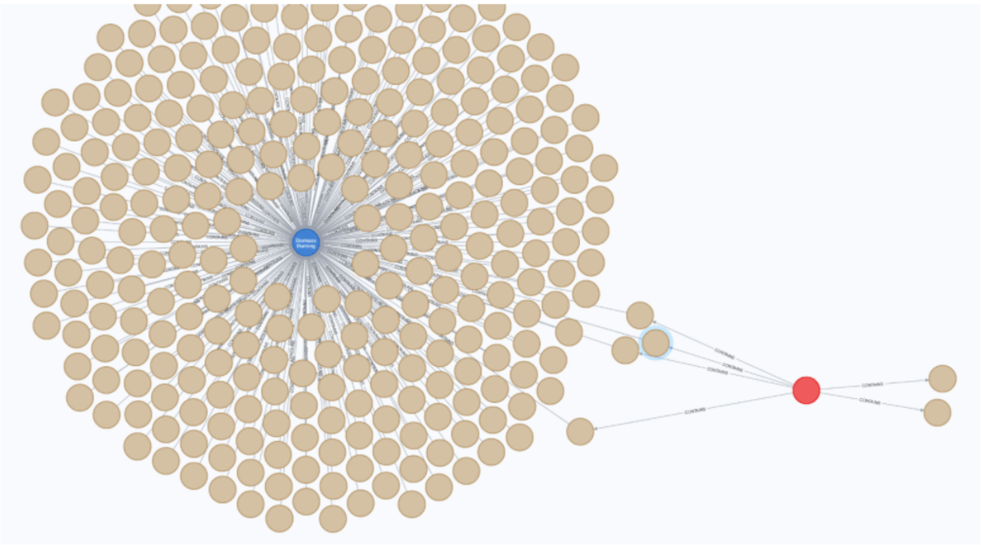
Result (convolution layer 3)

Source: Sensitivity of tropospheric oxidants to biomass burning emissions: implications for radiative forcing

Biomass Burning is One of the largest Sources of Trace Gases and Aerosols to the Atmosphere and has profound Influence on tropospheric Oxidants and Radiative forcing. Using a fully coupled chemistry-climate Model (GFDL AM3), we find that Co-Emission of Trace Gases and Aerosol from present-day Biomass Burning increases the global tropospheric Ozone Burden by 5.1% and Decreases global Mean Oh by 6.3%. Gas and Aerosol Emissions combine to Increase Ch4 Lifetime nonlinearly. Heterogeneous Processes are shown to contribute partly to the observed lower Do3/Dco Ratios in northern high Latitudes versus tropical Regions. **The Radiative forcing from Biomass Burning is shown to vary nonlinearly with Biomass Burning Strength**. At present-day Emission Levels, Biomass Burning produces a Net Radiative forcing of -0.19 W/m2 (-0.29 from short-lived Species, mostly Aerosol direct and indirect Effects, +0.10 from CH4- and CH4-induced Changes in O3 and stratospheric H2O) but increases Emissions to over 5 Times present Levels would Result in a Positive Net forcing.

Sentence score: 0, **Abstract score:** 0

Example Query and Result: Score 0



FUTURE WORK AND ACKNOWLEDGEMENTS

Future Work:

- Switch entity extractor and embeddings model to domain-specific language models
- Ingest more relevant documents into the graph database
- Determine a more quantitative way to perform query validation

This research is supported by NASA Grant NNM11AA01A as part of the Interagency Implementation and Advanced Concepts Team (IMPACT) program.

ABSTRACT

Traditional knowledge graphs tend to be too generic, and often perform poorly on complex scientific queries. Oftentimes, precedence is given to pop culture over scientific knowledge for queries. This is predominantly due to the use of internet sources for building the knowledge graph. With this work, we aim to explore the effectiveness of combining a knowledge graph generated from earth science corpora with a language model and graph convolutions for the purpose of surfacing latent and related sentences given a natural language query. In this model, sentences are conceptualized in the graph as nodes which are connected through entities—words and phrases of interest found in the text—extracted using Google Cloud's entity extraction model. The language model we used for this is Bidirectional Encoder Representations from Transformers(BERT). The sentences are given a numeric representation by the BERT model. Graph convolutions are then applied to sentence embeddings in order to obtain a vector representation of the sentence as well as the surrounding graph structure, thereby leveraging the power of adjacency inherently encoded in graph structures. With this presentation, we demonstrate the ability of graph convolutions and their improved ability to surface relevant, latent information based on the subject of the input query.