Requirements Discovery and Alignment Using Artificial Intelligence

Kevin Payumo - Final Report

Abstract

Deep learning based high-recall information retrieval (HRIR) includes novel language modeling methods such as Transformers and has shown promise in document review and identification. Graph Neural Networks (GNNs) are further shown to effectively learn graph data representations through a message-passing mechanism that informs how nodes are related. Previous works combined these architectures and demonstrated strong performance baselines on product graphs, paper citation graphs, and Wikipedia entity graphs. We propose using the combined GNN-Transformer architecture to automate document-personnel alignment on a small graph database including scientists' work organization and research history, publication documents, and project documents. We further explore how augmenting features through phrase extraction impact GNN performance and can be used for validation of document-personnel recommendations.

1 Introduction

Graph Neural Networks (GNNs) have found success in impressive applications such as recommendation systems, identifying cancer preventing molecules in foods, and image analysis [1][2][3]. Our group is interested in leveraging graphs to analyze scientific text corpuses from multiple sources. Recent works have shown that learned text semantics from large-scale pretrained (Transformer) models can benefit learning for text classification tasks [4].

In our dataset, nodes of a graph represent entities backed by descriptive features; scientists & their work experience, papers & their abstracts, projects & their descriptions. The graph is constructed around the scientist to model the relationship between them and their works (projects and papers). The choice of a GNN is justified in that text content will be collected from varying sources that differ in semantics. Furthermore, Transformer help encode semantics to be processed the by GNN. Then, model decisions can incorporate importance of neighboring nodes in addition to the target node. GNNs further offer multiple forms of prediction: node prediction, edge prediction, and subgraph prediction which can each contribute to a decision.

In this work we build upon Transformer-GNN architectures to learn scientists' experience, the descriptions of projects, papers, and other information in order to map government solicitation requirements to JPL skillsets. To supplement the main edge-prediction task, we use subgraph-prediction to resolve the difficulty of effective negative sampling in unsupervised contrastive learning [5][6]. Subgraph prediction further encourages data to be reorganized in a fashion that supports simple explainability of model decisions. We further implement an ensemble GNN approach with voting mechanism to counter the variability that comes from traditional negative sampling.

Our experiments demonstrate that a government solicitation can be treated as a research paper in respect to the graph dataset. Our model was able to recommend scientists that project managers also recommended for a solicitation about measuring the atmospheric wind profiles of 3D winds. For out-of-domain solicitations such as a DRACO propulsion solicitation, the subgraph-prediction side of the model was able to reject all recommendations. These results show that our model can match textual information with high resolution and we expect that this approach is applicable to many other high-recall information retrieval tasks.

1 Methods

1.1 Data

Scientist data was collected from JPL's Science People catalog [7] and loaded into a graph database with Neo4J. Fortunately, the data was quite clean except for the variety of formats that publications were cited on scientists' pages. Consistent formatting is important in order to accurately extract the titles of each paper. Then relationships between scientists can be encoded by the works they collaborated on. If two paper nodes are indeed the same, but one had a misspelled title, relevant relationships are lost in our graph. Our current dataset holds scientist data across 10 different categories within the parent domains of *Earth Science*, *Astrophysics and Space Sciences*, and *Exoplanetary Sciences,* resulting in 2124 nodes and 3035 relationships. Maintaining the quality of our dataset will be a challenge as we look to ingest the other 13 categories.

1.2 Transformers

We then reviewed the effectiveness of out-of-the-box transformer models, such as BERT [8], for preprocessing content. Content can be thought of as, the abstract to a paper, the biography to a person, the description of a project, amongst others. One question to consider was whether the semantics in the data used to train available BERT models, can be used to extract meaning in our data. For example, the language in climate change related Wikipedia articles may differ from language in climate change related research papers. Measuring this capability helps describe the potential for a GNN to accurately predict edges between nodes. To analyze BERT's potential for our use-case, we opt for a simple cosine similarity test between climate change and food related statements:

a. Climate change is having a drastic effect on our planet.
b. We need more sustainable ways to live.
c. Spaghetti and meatballs are my favorite food.

We then encode each sentence using BERT and apply a cosine similarity metric to describe the similarity between each pair of sentences,

Table 1. Similarity between each pair of sentences. A symmetric matrix, so one only needs to review an upper or lower triangle below the diagonal.

	a	b	c
a	1	0.7335	0.5500
b	0.7335	1	0.5966
c	0.5500	0.5966	1

Two sustainability related sentences *a-b* share a higher similarity than any other pairs, matching our expectations. Since these sentences are semantically simple, we further review BERT by comparing abstract A, an Earth Science related research paper, with B, an Astrophysics research paper [9][10].

For this metric, we compare each sentence from abstract A, to each other sentence in abstract A. Similarly, each sentence from abstract B, to each other sentence in abstract B. We expect a higher average similarity between sentences that come from the same abstract. We also compare each sentence from abstract A, to each sentence in abstract B. We expect a lower average similarly between sentences that comes from different abstracts.

Table 2. Average similarity between each pair of sentences within a given pair of abstracts.

	A	B
A	0.81 +/- 0.05	0.78 +/- 0.05
B	0.78 +/- 0.05	0.87 +/- 0.02

Sentences coming from the same abstract (A-A or B-B) average a higher similarity than when comparing sentences from two different abstracts (A-B). However, the A-A and A-B pair averages are within range of each other, and it's important to note that there are more A-B pairs than A-A, which can potentially drive averages lower and falsely validate our expectations. Further review is needed to involve more abstracts, but for the purposes of moving towards the main goal, these results show some evidence for BERT potential in the requirements alignment task.

1.3 Graph Neural Networks

Now that we have a means for encoding textual data, we jumped to learning about GNNs. *Figure 1* reveals a graphical structure for our database, where *nodes* connect with each other through *edges.* A GNN utilizes a message passage mechanism, essentially encoding how one node is connected to another. Note that direction of a connection is also important.

We opt for the Graph SAmpling and AggreGatE (SAGE) algorithm [11], more specifically its convolutional implementation [12], for GNNs as it is an inductive training approach, allowing us to inject test nodes for edge prediction. It works as stated in its name, sampling nodes and aggregating neighbors. Aggregation allows information of neighboring nodes to drive the encoding of the sampled node. With this model in mind, there is also a preprocessing step before we can process every node through the graph neural network.

Note that a *Scientist* node has three properties: biography, research interests, and professional experience. However, a *Paper* node only has an abstract property in our dataset. In order to use these properties to support GNN training, we needed to bring them in a standard format. To do so, we create synthetic properties for nodes with less than 3 properties (decided by the number of properties of the *Scientist* node). These synthetic properties are just zero vectors meant to get the dimensions of all nodes the same. The transformer outputs an encoding vector of length 768 for each property meaning that each node has a total length of 2304.

1.4 Multi-layer Perceptron (MLP)

This model was designed after finding that our GNN's loss was not converging during training, which we hypothesized was due to the curse of dimensionality. Thus, to support reduction of these dimensions, we deploy a simple MLP to transform each node's features into a vector of size 32 prior to being passed to the GNN. However, this decision needs to be reconsidered later in the project. The MLP still has parameters that are learned while training the GNN, which makes our models vulnerable to dimensionality issues.

1.5 AutoEncoders as Surrogate Models

We propose an alternative filtering mechanism that leverages surrogate autoencoders to reconstruct document embeddings. The reconstruction errors are indicative of whether a document's subject matter is represented in the original training set.

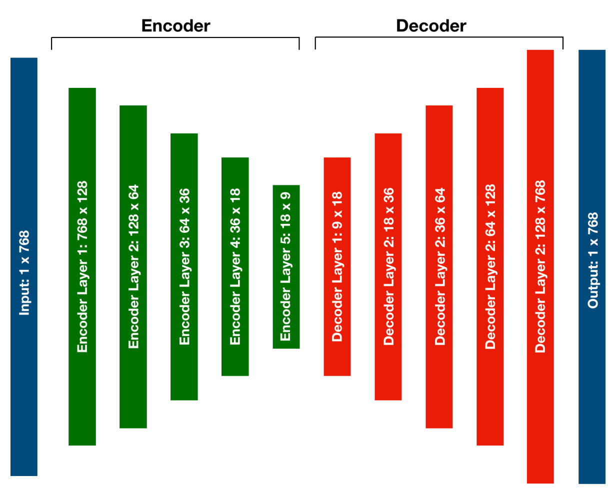

Figure 6. Autoencoder architecture, ReLU nonlinearity included between each layer. Input is downsampled steadily to a 1x9 feature vector before being upsampled back up to a 1x768 output vector.

As shown in Figure 6, a simple autoencoder architecture by GeeksforGeeks is used to test the potential of this filtering approach [13][14]. The input vector is compared with the reconstructed output through a Mean Squared Error Loss, generating the results in Table 5.

Table 5. Reconstruction error of solicitation documents

Solicitation	Reconstruction Error
Measuring the Atmospheric Wind Profile (3D Winds)	83.94
Demonstrating the Hyperspectral Microwave Sensor (HyMS)	101.98
Digital Twin for Earth Observations Using Artificial Intelligence	90.26
Astrophysics Probe Program Concept Spacecraft	63.81
Demonstration Rocket for Agile Cislunar Operations (DRACO)	217.26

Table 5 reveals the DRACO solicitation generates a significantly higher reconstruction error than other solicitations, matching our expectations, as propulsion systems required by this solicitation are not represented in our dataset. While our solicitation collection is too limited to guarantee using this model for filtering, we aim to further validate its purpose upon further data ingestion.

2 Results

We used our trained model to analyze example government solicitations and predict scientists that would be a good fit for responding to those solicitations.

Table 1. Top model recommendations (Score>0.50) for Measuring the Atmospheric Wind Profile (3D Winds). Scores represents the model confidence that scientists' data is relevant to the solicitation's requirements.

	names	score	revised_scores
0	Mathias Schreier	0.850	0.849
1	Bjorn Lambrigtsen	0.840	0.840
2	Brian Kahn	0.851	0.822
3	Hui Su	0.844	0.820
4	Evan Fishbein	0.809	0.806
5	Derek Posselt	0.799	0.795
6	Jinbo Wang	0.801	0.791
7	Eric Fetzer	0.820	0.775
8	Lucien Froidevaux	0.767	0.764
9	Nathaniel Livesey	0.789	0.764
10	William Read	0.735	0.727

Derek Posselt and Hui Su are known personnel who are working on the 3D Winds solicitation. Other names highlighted in green are known recommendations to have been made by a program manager. The names not highlighted in green are known collaborators with some of the ground truth recommended scientists. Only one recommendation was an outlier which we found to be caused by similar name initials in our dataset. Then 9 out of 10 scientists can be deemed valid recommendations for the solicitation, with 6 of those recommendations being our target recommendations. These results serve as strong evidence for the potential of Transformer-GNNs for information retrieval.

One might then ask, what does the model predict if provided an irrelevant solicitation? Irrelevant solicitations are described as documents with subjects not represented in the training dataset.

Table 2. Top model recommendations (Score>0.50) for Demonstration Rocket for Agile Cislunar Operations (DRACO).

	names	score	Graph Validated Score
0	Bruce Bills	0.896	6.257500e-06
1	Renyu Hu	0.658	1.508800e-06
2	Nathaniel Livesey	0.251	4.913000e-07
3	Jinbo Wang	0.213	4.725000e-07
4	James Ling	0.476	4.215000e-07
5	Michelle Santee	0.157	1.539000e-07
6	Baijun Tian	0.107	1.174000e-07
7	Eric Fetzer	0.098	6.100000e-08
8	Brian Kahn	0.131	5.190000e-08
9	Kevin Baines	0.792	3.970000e-08
10	Glenn Orton	0.716	3.160000e-08

Our dataset does not contain personnel with strong experience in rockets and propulsion systems, so we expect the model to return no recommendations. Recall that our model is composed of an Ensemble model side that does edge-prediction and a Graph Validation side that does subgraph-prediction. The Ensemble side predicted some high scores for personnel like Bruce Bills despite those scientists not having any propulsion experience. We hypothesize this to be due to how neural networks construct a representation space for training data which makes it difficult to construct a space for all out-of-domain topics. This is where the strength of the graph validation model comes in as it can encode combinations of subgraphs to be valid. Out of domain subjects will not have significant phrase nodes that match our dataset and thus the model will be able to reject recommendations that cannot construct valid subgraphs. This is seen by the Graph Validated Score which penalizes the Ensemble score down to a magnitude less than 1e-06.

6 Conclusion and Future work

Our team managed to experience the whole system pipeline from ingesting data, encoding data, training a GNN, and inferring recommendations for new solicitations. We uncovered interesting edge cases which we solved by devising new training strategies which lead to more stable model training and higher test accuracy. We were able to retrain the language model as a backup method to adjust to applications with completely unique semantics.

We have developed an architecture that leverages large scale pretrained models and multiple GNNs to view data from multiple perspectives. Our model demonstrated strong ability to provide recommendations that matched expert performance for several solicitations. We demonstrated robustness against out-of-domain use cases and created complementary explainability methods to support user confidence in applications.

Future works involve reviewing versatility of model architecture. For requirements alignment, we found 85% accuracy on test cases for paper-scientist edges but only had two government solicitations to measure against for the true solicitation requirements scientist alignment task. We are interested in packaging this work into a software package that we can then start to test on other use-cases such as a JPL proposal knowledge base curated by JCC. Such transition may involve more fine-tuning of Transformer models which we are now familiar with. We'd also like to start incorporating user feedback which is a core component to improving recommendation systems and we'd also like to develop new explainability visualizations to support application use.

References

1. Shiwen Wu, Fei Sun, Wenta Zhang, Xu Xie, Bin Cui, "Graph Neural Networks in Recommender Systems: A Survey", (2022) https://arxiv.org/pdf/2011.02260.pdf
2. Kirill Veselkov, Guadalupe Gonzalez, Shahad Aljifri, Dieter Galea, Reza Mirnezami, Jozef Youssef, Michael Bronstein & Ivan Laponogov, "HyperFoods: Machine intelligent mapping of cancer-beating molecules in foods", (2019) https://www.nature.com/articles/%20s41598-019-45349-y
3. Varun Vasudevana, Maxime Bassenneb, Md Tauhidul Islamb, Lei Xing, "Image Classification using Graph Neural Network and Multiscale Wavelet Superpixels", (2022) https://arxiv.org/pdf/2201.12633.pdf
4. Yuxiao Lin, Yuxian Meng, Xiaofei Sun, Qinghong Han, Kun Kuang, Jiwei Li, Fei Wu, "BertGCN: Transductive Text Classification by Combining GCN and BERT", (2021) https://aclanthology.org/2021.findings-acl.126.pdf
5. Zhen Yang, Ming Ding, Chang Zhou, Hongxia, Yang, Jingren Zhou, Jie Tang, "Understanding Negative Sampling in Graph Representation Learning" (2020) https://keg.cs.tsinghua.edu.cn/jietang/publications/KDD20-Yang-et-al-Understanding_Negative_Sampling_in_GNN.pdf]
6. Shiyi Chen, Ziao Wang, Xinni Zhang, Xiaofeng Zhang, Dan Peng, "Probing Negative Sampling for Contrastive Learning to Learn Graph Representations", (2021) [https://2021.ecmlpkdd.org/wp-content/uploads/2021/07/sub_704.pdf
7. "Jet Propulsion Laboratory: People", https://science.jpl.nasa.gov/people/
8. Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" (2019), https://arxiv.org/abs/1810.04805
9. David Halpern, Dimitris Menemenlis, Xiaochun Wang, "Impact of Data Assimilation on ECCO2 Equatorial Undercurrent and North Equatorial Countercurrent in the Pacific Ocean" (2015) https://journals.ametsoc.org/view/journals/atot/32/1/jtech-d-14-00025_1.xml
10. Mark Lacy, Brian Mason, Craig Sarazin, Suchetana Chatterjee, Kristina Nyland, Amy Kimball, Graca Rocha, Barnaby Rowe, Jason Surace, "Direct Detection of Quasar Feedback Via the Sunyaev-Zeldovich Effect" (2019) https://arxiv.org/abs/1811.05023
11. William L. Hamilton, Rex Ying, Jure Leskovec, "Inductive Representation Learning on Large Graphs" (2018) https://arxiv.org/abs/1706.02216
12. Felix Wu, Tianyi Zhang, Amauri Holanda de Souza Jr., Christopher Fifty, Tao Yu, Kilian Q. Weinberger, "Simplifying Graph Convolutional Networks", (2019) https://arxiv.org/pdf/1902.07153.pdf
13. D.E. Rumelhart, G.E. Hinton, R.J. Williams, "Learning internal representations by error propagation" (1986), http://www.cs.toronto.edu/~hinton/absps/pdp8.pdf
14. GeeksforGeeks, "Implementing an Autoencoder in Pytorch" (2022), https://www.geeksforgeeks.org/implementing-an-autoencoder-in-pytorch/