Exploring the Potential of Transformer Models as a Novel Approach to Biological Knowledge Graph Construction

Aidan Lowrie
Boray Kasap
Eliott Mattei

Text Mining
Amsterdam University College
Spring 2023 Research Project

Abstract

This paper focuses on the use of transformer models to aid in the creation of knowledge graphs (KGs) from a corpus of PubMed extracts. The data processing protocol involves several stages, including data collection using Entrez-Utilities and Biopython, relationship extraction using REBEL, keyword extraction using KeyBERT and BioBERT, and entity resolution using Stanzas biomedical model. The resulting triples are visualized using NetworkX as KGs. The performance of the generated triples is evaluated using PyKEEN, a KG embedding library. The results show high performance in terms of Adjusted Geometric Mean Rank Index (AGMRI), but lower performance in terms of Adjusted Arithmetic Mean Rank Index (AAMRI) and Adjusted Hits at K (AH@K). Further analysis is needed to identify specific areas where the model struggles and to understand the implications for the KG and generated triples. The paper discusses the need for an association measure to capture the strength of relationships and suggests the combination of data to explore interactions between relationships.

Introduction

The biological sciences are associated with myriad interrelated terms and concepts. Molecular machines, genes and metabolites form an intricate network of dependencies that even experts cannot fully comprehend. To mitigate the risk of overlooking key information during the research process, significant effort has been made over the past several decades towards the organisation of biological data, and the development of tools that capture the semantic relationships between terms.

One important task in the development of such tools is Relationship Extraction (RE) a Natural Language Processing (NLP) task aimed towards extracting relationships between co-occurring terms. When carried out, RE returns ‘triples’ (co-occurring terms and their relationship), which may subsequently be represented by a knowledge graph (KG).

KGs have demonstrated value beyond just data visualisation, having been used to train a variety of deep learning models. One study successfully trained a model to use KG embeddings to generate ‘concept prerequisites’ - basic concepts required to understand more complicated ones (Manrique et al., 2019). Furthermore, a recent article by Google (Shakeri and Agarwal, 2021) demonstrated an approach allowing for the integration of knowledge graphs into the training corpora of Large Language Models (LLMs). This may improve the performance of LLMs in knowledge-intensive tasks like question-answering, where factual responses are essential. Finally, models may be trained for link prediction, in order to find novel relationships between terms. In the context of biology, this could aid in the discovery of new molecular and cellular interactions.

While these applications are promising, the creation of high-quality knowledge graphs is still a work in progress. Several ongoing projects, including ConceptNet and Cyc, aim to develop comprehensive KGs for use as tools (Kejriwal et al., 2022). Over recent years, these projects have shifted towards novel NLP approaches, only possible thanks to significant advances in artificial intelligence (AI). This can be seen in the increasing use of Bidirectional Encoder Representations from Transformers (BERT), a family of deep learning models.

In recent years, KGs have gained attention in the biomedical field. A recent research publication on this topic carried out RE by parsing each sentence in the corpus into fundamental components - a noun phrase and a verb phrase. From this, *subject, predicate, object* triplets were derived. This was aided by Semantic Role Labeling (SRL), which identifies the semantic roles words and phrases play in sentences. Secondary relationships were also extracted, such as noun modifiers (Rosannez et al., 2020). A limitation of this study was the reliance on shallow algorithms to carry out RE, which perform worse than modern deep relationship extraction models.

Another study used BERT models to build a knowledge graph of PubMed information such as author, publication date and topic (Xu et al., 2020). But rather than aiming just to represent the semantic relationships between terms in abstracts, this knowledge graph aimed towards the representation of information relating to PubMed information such as author, publication date and general topics. Since this paper, major advancements relating to RE have been made, including the development of BART (Lewis et al., 2019) and REBEL, a model based on BART that is fine tuned for relationship extraction, able to achieve more accurate results than the BERT-based-approach used in the study.

Despite their potential, previous knowledge graphs representing biological corpora have fallen short of being truly useful practical tools for researchers. Further effort is needed to produce KGs that are genuinely useful, and can aid people in understanding concepts. Throughout this paper, we shall investigate the potential of recent transformer models to aid in the creation of KGs from a corpus of PubMed extracts, utilising REBEL for RE and BERT models for key-word extraction and entity resolution.

Methods

Our data processing protocol involved several key stages, ultimately aimed towards the generation of *subject, relationship, object* triples. The steps for data collection and processing are outlined below.

1. Downloading the Data with Entrez-Utilities and Biopython:

Entrez-Utilities is a set of tools developed to facilitate the process of programmatically downloading bioinformatic data from the NCBI database (Sayers, 2022). We used Biopython, a library designed for leveraging the E-Utilities, to search for and retrieve XML data for the top 15,000 PubMed articles relating to the search term ‘biology’. The abstracts themselves were extracted from the XML data, then stored to a text file. Abstracts, which aim to be concise representations of full research articles, were an ideal choice given our lack of access to computing power.

2. Relationship Extraction with REBEL:

The next task was to extract relationship triples from the corpus. This was performed using REBEL, an open-source relationship extraction seq2seq model that achieved state-of-the art RE performance (Huguet Cabot and Navigli, 2021). The corpus was segmented into sentences using NLTK, which were then passed into the model. The outputted triples were parsed and saved to a CSV file.

3. Keyword Extraction with KeyBERT and BioBERT:

Many terms and relationships captured by REBEL did not relate to biology, for example '*Pisa, capital, Italy'*. A keyword extraction method was required to filter irrelevant terms from the triples dataset. For this we used KeyBERT, a library that leverages BERT embeddings for the purpose of keyword extraction (Grootendorst, 2022). Unigram, bigram and trigram keywords were generated for each abstract, then used to filter the triples dataset.

4. Entity Resolution with Stanza's Biomedical Model:
The filtered dataset contained many variants of the same term with slight differences in spelling. To normalise the dataset, lemmatization was applied. Standard lemmatization techniques using SpaCy or NLTK proved insufficient for linking biological terms. Stanza's biomedical model (*REFERENCE*) aims to address this, with models specialised for the task of biomedical lemmatization. Unigrams were lemmatized and rows with identical term pairs were filtered such that the data frame contained only one triple; the two terms and their modal relationship. Doing so reduced the dataset from 33000 unique word pairs to around 8000.

5. Visualisation using NetworkX:
NetworkX is a Python package for the creation and study of complex networks. The triple dataframe was converted to a NetworkX graph object for visualisation as KGs. In this object, nodes represent unique terms in the dataframe and edges represent relationships between terms. Subgraphs were created for further visualisation.

6. Predictor Training with PyKEEN:
To evaluate the performance of the generated triples, PyKEEN (a KG embeddings library) was used to train a predictor model. The PyKEEN pipeline works by removing labels and having the model predict term associations. Triples were split into 20% test data and 80% training data then passed into the PyKEEN pipeline, for 100 epochs of training followed by statistical tests. Adjusted Mean Rank Index, Adjusted Geometric Mean Rank Index and Adjusted Hits at K scores were used to quantify performance.

Results and Analysis

Following the method described above, triples were successfully generated and visualised on a KG (Figure 1). Evaluation was carried out on both the lemmatized and unlemmatized triple data. 'Adjusted Arithmetic Mean Rank Indicator', 'Adjusted Geometric Rank Indicator' and 'Adjusted Hits at K' statistical tests were used to analyse the models. The scores vary between 0 and 1, where 1 represents perfect performance (Figure 2). The predictor trained from the unlemmatized triplets marginally outperformed the model trained from the lemmatized corpus across the tests. The predictors' average AGMRI performances were high (~0.91). However, Arithmetic Mean Rank scores were significantly lower, around ~0.49. As Geometric Mean Rank scores are less sensitive to extreme deviation, the significant difference between the two test scores suggests the existence of outliers in the data; relationships that the predictor significantly misjudged. AH@K returned much lower values (~0.13), meaning that correct entities were rarely assigned to the top probabilities of the predictor's output. Head and tail scores were similar, suggesting no discrepancy in quality between outgoing and incoming edges.

Discussion

Our triples successfully captured many important relationships between key terms, as can be observed through their visualisation on a KG (Figure 1). But our methodology still has significant room for improvement. For example, many captured relationships were inaccurate or entirely missing, leading to subpar predictor model performance during the evaluation. A more thorough investigation of the performance of predictor models could aid in determining the kinds of associations the predictor failed to intuit (which reflect the groups represented worse by the KG). For example, NER and/or clustering could be used to create subgraphs, for training a range of predictors. Their performance could then be compared to identify particular categories of entities / groups of terms with which the predictor performed more poorly.

One major limitation to our methodology lies in the fact that we used the basic REBEL model rather than fine tuning one for the task of extracting relationships in a biological text. As a result, many untrue relationships - for example, '*dna, has part, histone'* - were captured by REBEL. We could not find a dataset for training a biological corpus for relationship extraction. The problem of lacking a dataset for the task of fine tuning a model came up once more when trying to carry out entity resolution. When our initial approach (using a combination of BioBERT embeddings and levenshtein distance) did not perform, we attempted to finetune a BioBERT model for this task. Unfortunately, we lacked the ability to make a dataset of the required length for entity resolution, and no resources were available online to do so. Future research should aim to create both datasets mentioned.

Another potential improvement to the methodology involves *association strength.* Our KG treats every relationship the same - it does not measure how strong associations are. In our methodology, the final triplets contained modal relationships identified in the corpus, and were given an equal weight on the graph. But one would expect some associations to be stronger than others. During discussion, we came up with several ideas, such as using the number of repetitions or surface co-occurrences measures to determine edge weights in the final KGs. But, after some discussion, it was clear to us that given that the distribution of our data is very sparse, using repetitions might overrepresent the relationships of negligible importance. We are leaving the challenge of how to determine association strength between terms as an open question to the community that, when answered, could provide a clearer path to exploring semantic relationships in the biological field.

A final improvement could be made by making use of predictive models, which were used in this study for the purpose of evaluation. Similar machine learning techniques could be utilised to fill in missing associations. For example, a KG might have two links: A causes B and B causes C. Logic dictates that in this case, A causes C. Such 'missing links' could be found by predictive models. We recommend further investigation of this topic.

As can be seen from the end product, biological relationships indeed constitute a chaotic network! Further studies should explore the methods mentioned above to improve our ability to generate high quality biological knowledge graphs.

Appendices

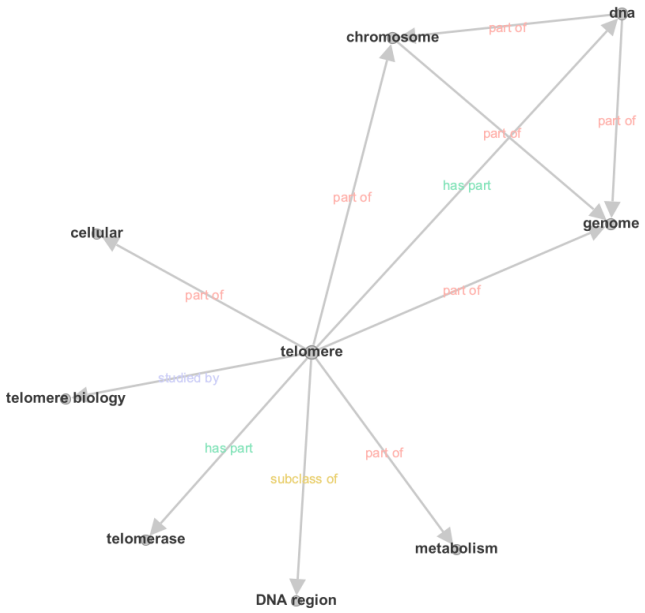

Figure 1: An example of a knowledge graph generated by our triples. (Further examples are available as PDFs in this report's supplementary files.)

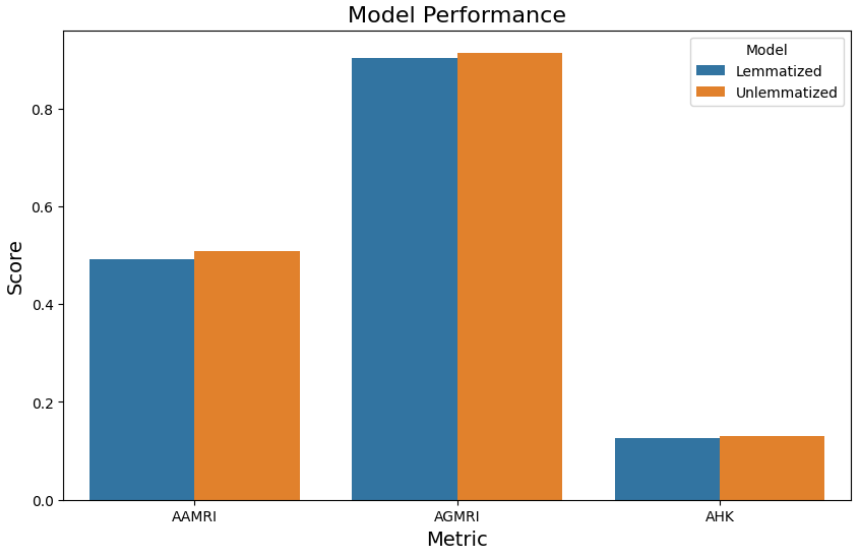

Figure 2: The relative performance of the two models across three statistical tests.

Supplementary Materials

Supplementary materials can be found on our GitHub repository, *https://github.com/aidanlowrie/text-mining-project.*

Contribution

Aidan: Implementing the pipeline discussed in the ‘methods’ section, writing the project notebook, presentation, contributed to the final report.
Boray: Implemented code related to the visualisation of the data, contributed to the final report..
Eliott: Contribution to the final report and presentation.

Bibliography

Cabot Huguet, Pere-Lluís, and Navigli Roberto. ‘REBEL: Relation Extraction By End-to-End Language Generation’. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2370–81. Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021. https://doi.org/10.18653/v1/2021.findings-emnlp.204.

Grootendorst, Maarten, ‘KeyBERT’, 2022 Accessed 4 June 2023. https://maartengr.github.io/KeyBERT/.

Kejriwal, Mayank. ‘Knowledge Graphs: A Practical Review of the Research Landscape’. *Information* 13, no. 4 (April 2022): 161. https://doi.org/10.3390/info13040161.

Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. ‘BART: Denoising Sequence-to-Sequence Pre-Training for Natural Language Generation, Translation, and Comprehension’. arXiv, 29 October 2019. http://arxiv.org/abs/1910.13461.

Manrique, Rubén, Bernardo Pereira, and Olga Mariño. ‘Exploring Knowledge Graphs for the Identification of Concept Prerequisites’. *Smart Learning Environments* 6, no. 1 (12 December 2019): 21. https://doi.org/10.1186/s40561-019-0104-3.

Rossanez, Anderson, Julio Cesar dos Reis, Ricardo da Silva Torres, and Hélène de Ribaupierre. ‘KGen: A Knowledge Graph Generator from Biomedical Scientific Literature’. *BMC Medical Informatics and Decision Making* 20, no. 4 (14 December 2020): 314. https://doi.org/10.1186/s12911-020-01341-5.

Sayers, Eric. ‘A General Introduction to the E-Utilities’. In *Entrez Programming Utilities Help [Internet]*. National Center for Biotechnology Information (US), 2022. https://www.ncbi.nlm.nih.gov/books/NBK25497/.

Siamak Shakero, Oshin Agarwal, ‘KELM: Integrating Knowledge Graphs with Language Model Pre-Training Corpora’, 20 May 2021. https://ai.googleblog.com/2021/05/kelm-integrating-knowledge-graphs-with.html.

Xu, Jian, Sunkyu Kim, Min Song, Minbyul Jeong, Donghyeon Kim, Jaewoo Kang, Justin F. Rousseau, et al. ‘Building a PubMed Knowledge Graph’. *Scientific Data* 7, no. 1 (26 June 2020): 205. https://doi.org/10.1038/s41597-020-0543-2.