Biomedical Knowledge Graph Generation With Seq2seq Model

CS6120 Final Report
Members: Hyun Seung Lim, Xinan Wang, Zhen Wang, Bereket Faltamo

Source Code: https://github.com/thomaslim6793/rebel

Abstract:

Knowledge Graphs visually represent information as a graph, consisting of nodes connected by edges. They offer a powerful way to comprehend information in terms of entities and relation types. While traditional plain-text presentation is more condensed, it often obscures the relationships between ideas. In our work, we demonstrate Knowledge Graph generation using seq2seq modeling, specifically fine-tuning REBEL, an encoder-decoder transformer model, on the BioRel dataset, which annotates biomedical text. By training REBEL on this dataset and optimizing the training process, we developed a specialized model. This model prioritizes relation types specific to the biomedical domain, enhancing its ability to extract meaningful information.

Introduction:

The decision to focus on the biomedical domain for developing a model to generate knowledge graphs from textual inputs is guided by several compelling reasons.

Firstly, the biomedical field is known for its vast and intricate data landscape, which includes clinical records, research publications, genetic data, and extensive drug databases. This diversity provides a solid foundation for extracting relationships and building detailed knowledge graphs. These graphs have the potential to significantly impact public health by mapping complex relationships between diseases, treatments, symptoms, and genetic markers. They can enhance our understanding of diseases, predict outcomes, and support personalized medicine approaches.

Related Works:

The goal of the Relation Extraction task (RE) is to extract a set of relations from a given text. There are two main approaches widely used for this task **[Figure 1]**:

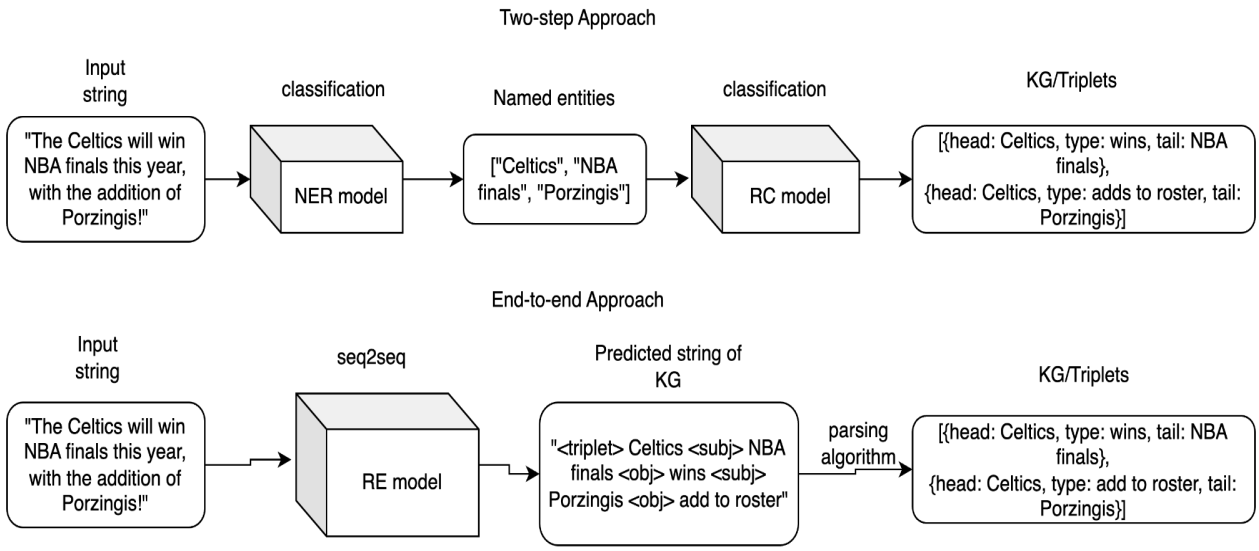
1. **Two-step approach:** In this approach, Named Entity Recognition (NER) is performed in the first step, which converts the input text into a set of named entities. In the second step, Relation Classification (RC) is performed, where the set of named entities is the input, and the output is the predicted relation type for every pair of named entities **[1]**.

2. **One-step (or end-to-end) approach:** This approach involves using a single model to convert the input text into an output text that represents the knowledge graph. **[2]**

The one-step approach is considered favorable to the two-step approach in several ways:
- The two-step approach requires two models, which necessitates separate training and thus incurs greater computational cost and overhead.

- The RC step in the two-step approach is limited to a predefined set of relation types. In contrast, the end-to-end approach theoretically can perform zero-shot learning, enabling it to predict relation types that it has not been trained on.

Figure 1: Two-step approach vs one-step approach



Methods:

1. **Data**

Our model is fine-tuned on BioRel, a large-scale dataset for biomedical relation extraction problem, by using Unified Medical Language System as knowledge base and Medline as corpus **[3]**. It consists of biomedical text extracted from PubMed abstracts and annotations of semantic relations between biomedical entities such as genes, diseases, and chemicals. The relations annotated in the dataset span across various biomedical domains, including genetics, molecular biology, and cancer research.

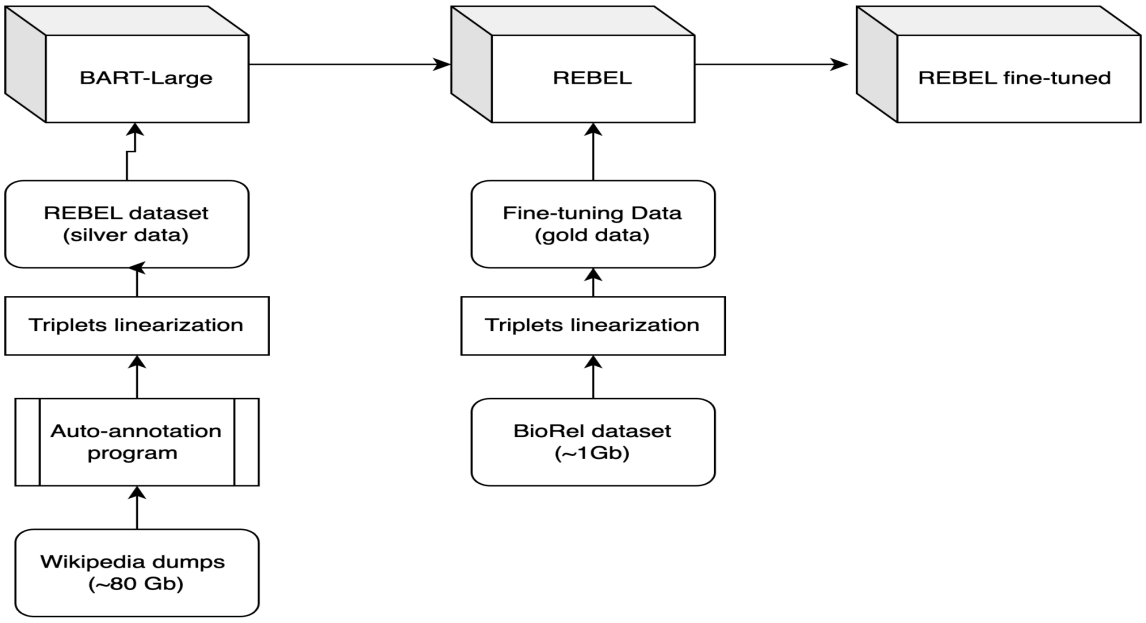
We utilized only a test split of the original dataset to fine-tune our model as BioRel is a large-scale dataset. The dataset is then pre-processed to get linearized triplets in a string representation **[Figure 2]**. The raw training data was 800Mb in size.

The data pre-processing step involves transforming the triplets (head entity, tail entity, relation) into a linearized representation using a template format. This linearized representation is then used as input to the model to perform relation extraction. Here's how the the linearization scheme works:

1. For each sentence, we collect the unique subject entities (head entities) and sort them by their start index in the sentence.
2. For each subject entity, we create a linearized triplet string starting with "*<triplet> subject_entity*".
3. For each triplet associated with the current subject entity, we append the following to the linearized triplet string:
    a. " *<subj> object_entity <obj> relation_type* ", where *object_entity* is the tail entity and *relation_type* is the relation between the subject and object entities.

4. The linearized triplet string for the sentence is created by concatenating the linearized triplets for each subject entity.

Figure 2: Project scheme



For example, consider a sentence with two triplets:

Triplet 1: (head_entity="gene1", tail_entity="disease1", relation="relation1")
Triplet 2: (head_entity="gene1", tail_entity="disease2", relation="relation2")
The linearized triplet string for this sentence would be:
<triplet> gene1 <subj> disease1 <obj> relation1 <subj> disease2 <obj> relation2

By linearizing the triplets in this way, the model can learn to generate the correct triplet string given a sentence, effectively performing relation extraction. It also improves performance compared to traditional supervised learning approaches as the same information is presented in a more condensed format for the model to process.

2. Model:

BART: The base model is BART-large, a seq2seq transformer model with an encoder-decoder architecture used for Relation Extraction (RE). The encoder includes two bidirectional LSTM layers and an attention layer, which allow for comprehensive contextual analysis of tokens by considering influences from both adjacent tokens and local text segments. BART is semi-supervised trained to denoise text by correcting artificially introduced errors like masking or rearranging tokens. This capability is crucial for transforming unstructured text into structured knowledge graphs. **[4]**

REBEL: REBEL is a model fine-tuned on BART using Wikipedia dump data to create a silver dataset of knowledge graph annotations. It employs a novel annotation format called 'linearized triplets,' which

condenses the target sequence by using a single head token for all triplets with that token as the head entity. This preprocessing optimization is crucial, as the performance of seq2seq models generally improves with shorter target sequences, making REBEL more effective compared to other seq2seq models. **[5]**

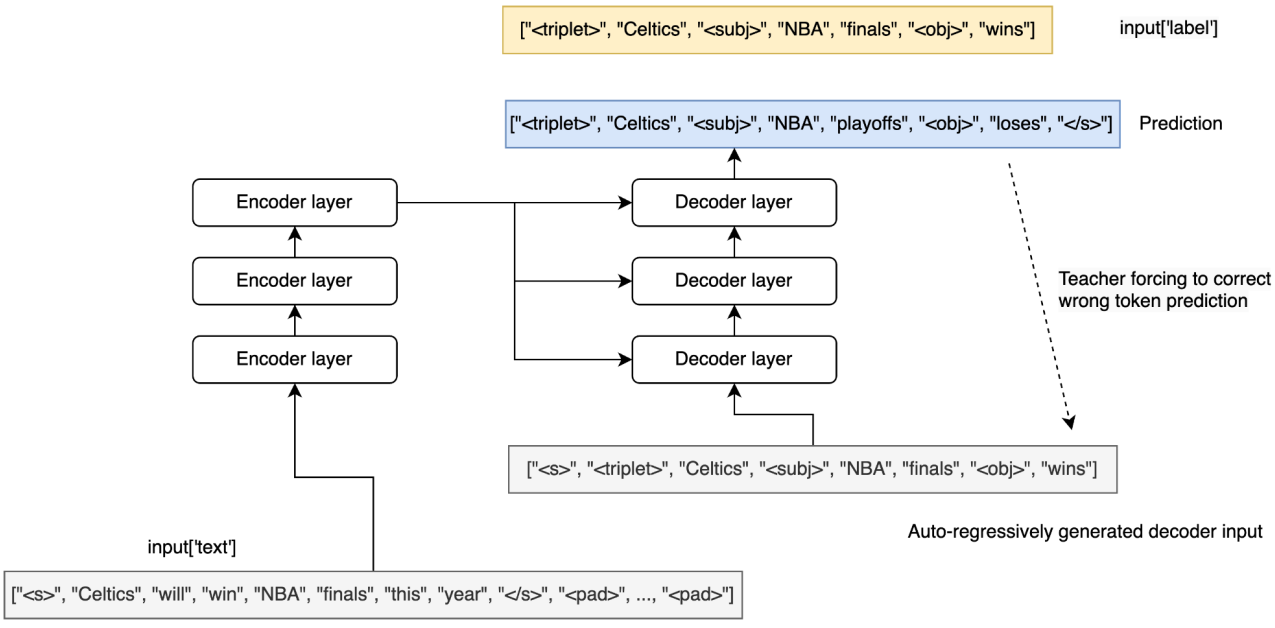
3. Training

The fine-tuning process of REBEL involves supervised learning using the BioRel dataset, which includes 75K knowledge graph (KG) annotated sentence-level documents. The model was trained for 9 epochs on a single V100 GPU with an Adam optimizer at a learning rate of 0.00005, using a cross-entropy loss function. This function calculates the negative log loss for each predicted token by the auto-regressive decoder compared to the target token, with a weighted average determining the overall loss **[Figure 3]**.

Two key optimizations were implemented: 1) ignoring padding tokens in loss calculations to prevent skewing the model towards predicting them, since padding tokens should not contribute to the loss, and 2) applying 'teacher forcing' to ensure that incorrect predictions during training do not influence subsequent tokens. While 'teacher forcing' directs the decoder to use actual subsequent tokens from the target sequence during training, the decoder reverts to auto-regressive generation at inference.

These strategies enhance the training process by focusing the model on relevant data and maintaining the integrity of sequence generation.

Figure 3: REBEL model architecture, showing training for a single sample.



Experiments:

We fine-tune REBEL on BioREL train and validation split for 9 epochs on the HCP Discovery cluster for 8 hours. After training, we evaluate the model on BioREL test split and another relation extraction dataset.

Two measurement metrics were adopted in this project. First, we evaluated the model using F1 score based on the predicted outputs and labeled triplets. The prediction would only be considered correct when the entities and the relation types have to be the same as those in the labeled triplets. As F1 measure is rather strict considering the training objective is sequence to sequence task, we also used Rouge-L to better evaluate the fine-tuned model's performance on datasets with different annotation rules and compare it with the original REBEL.

Results:

Testing with the BioREL test dataset, we observed an improvement in Rouge-L from the original REBEL model to the trained REBEL, with the latter achieving an F1 score of 0.8 compared to 0 by the former. The BioREL dataset, a large-scale distantly supervised set, tends to feature high-level summarizations rather than direct extractions from sentences, which may not capture all possible triplets.

For instance, consider the sentence: "Since some monoclonal or polyclonal human serum antibodies of rheumatoid arthritis (RA) or mixed connective tissue disease (MCTD) have been reported to recognize shared epitopes of denatured IgG and DNA-histone (nucleosomes), this MRP-2 monoclonal antibody with similar activity derived from a lupus-prone mouse will be useful for studies on the etiology of autoantibodies associated with RA, MCTD, and systemic lupus erythematosus (SLE)."

The following are predictions from REBEL and REBEL trained:

REBEL: "<s><triplet> monoclonal <subj> antibodies <obj> subclass of <triplet> polyclonal <subj> antibodies <obj> subclass of</s>"

REBEL trained: "<s><triplet> rheumatoid arthritis <subj> ra <obj> nichd_parent_of <triplet> mixed connective tissue disease <subj> mctd <obj> nichd_parent_of <triplet> histone <subj> nucleosomes <obj> gene_product_has_associated_anatomy</s>"
Label: "<triplet> histone <subj> nucleosomes <obj> gene_product_has_associated_anatomy"

While REBEL was able to extract partial information from the original text, REBEL trained successfully predicted the high level relation between "histone" and "nucleosomes" as the labels provide. As both the example and the statistics showed, REBEL trained performed fairly well on relation extraction from biomedical texts. Meanwhile, 0 F1 score doesn't reflect that the original REBEL is incapable of extracting triplets. Rouge-L is a more suitable measurement when evaluating performance.

Table 1: Comparison (Rouge-L and F1) between REBEL and REBEL trained

	Rouge-L	**F1**
REBEL$_{trained}$	0.88	0.8
REBEL	0.46	0

Additional Results:

Chemprot is a document-level dataset that identifies entities of chemicals and proteins and their likely relations. It comprises 1,820 PubMed abstracts with chemical-protein interactions annotated by domain experts and includes 10 groups of general relation types **[6]**.

We tested both REBEL and REBEL trained with Chemprot. REBEL achieved a 0.35 in Rouge-L, while REBEL trained scored 0.17. The results indicated that the REBEL trained model performed poorly in extracting the generic relations specific to chemical-protein interactions. The original REBEL, without training on BioREL, showed relatively better performance.

Conclusion:
Based on the superior performance of our model on the BioRel test set compared to the base REBEL model, it is evident that the general REBEL model primarily extracts generic relations and struggles with domain-specific types from the test dataset. Conversely, our fine-tuned model's failure on datasets like ChemProt and its inability to predict new relationship types indicate a strong bias towards training-set relations, despite its theoretical capability to generate any arbitrary string of relationship type.

Additionally, both models perform poorly with long text inputs because they were trained on sentence-level documents, leading to issues in capturing long-range relationships in larger texts. Currently, the approach involves processing text at the sentence level and assembling triplets independently, which fails to capture extended relationships across multiple sentences.

Future improvements for our model include:
1. Capturing long-range relationships between entities.
2. Expanding the training dataset to include a wider variety of relationship types.
3. Enhancing the model's understanding of English grammatical structures to better identify and predict verbs as relationship types.

References:

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 2335–2344, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Makoto Miwa and Mohit Bansal. 2016. End-to-end relation extraction using LSTMs on sequences and tree structures. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1105–1116, Berlin, Germany. Association for Computational Linguistics

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7871–7880, Online. Association for Computational Linguistics.

Xing, R., Luo, J. & Song, T. BioRel: towards large-scale biomedical relation extraction. BMC Bioinformatics 21 (Suppl 16), 543 (2020). https://doi.org/10.1186/s12859-020-03889-5

Pere-Lluís Huguet Cabot and Roberto Navigli. 2021. REBEL: Relation Extraction By End-to-end Language generation. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 2370–2381, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Olivier Taboureau, Sonny Kim Nielsen, Karine Audouze, Nils Weinhold, Daniel Edsgärd, Francisco S. Roque, Irene Kouskoumvekaki, Alina Bora, Ramona Curpan, Thomas Skøt Jensen, Søren Brunak, Tudor I. Oprea, ChemProt: a disease chemical biology database, Nucleic Acids Research, Volume 39, Issue suppl_1, 1 January 2011, Pages D367–D372, https://doi.org/10.1093/nar/gkq906