# Learn-to-Retrieve-and-Generate for Article Generation

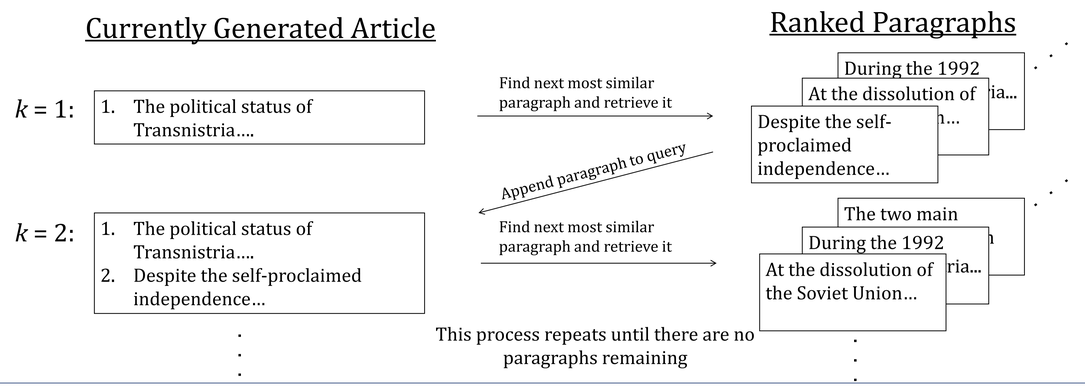
Joseph Guidoboni

*Department of Computer Science, University of New Hampshire, Durham, NH 03824*

## Introduction

In recent years, the development of new Natural Language Processing (NLP) and Information Retrieval (IR) models has led to great strides in the field of text retrieval and article generation. This project explores using one such model, Google's Bidirectional Encoder Representation from Transformers model (BERT), to predict the following paragraph of a Wikipedia article given it's first $k$ paragraphs. In doing this, we look to rebuild Wikipedia articles continuously paragraph by paragraph.

## Retrieval and Generation Process




## Results

As shown in the Model Accuracy section, the fine-tuned models are slightly less accurate than the pre-trained BERT model. Both `bert-base-uncased` and `batch-size-1` rank the correct next article roughly in it's top three choices (0.36034 and $0.31699 \approx 1/3$) , with `batch-size-4` ranking it in the model's top four choices on average ($0.27453 \approx 1/4$) for the given example. While the fine-tuned models are less effective than base BERT, they do still perform significantly better than the random retrieval and generation baseline.

## Methodology

The project utilizes an implementation of BERT provided by HuggingFace's `transformers` library, where fine-tuning is performed on top of the base implementation. Two models are trained, one with a batch size of 1, and another with a batch size of 4, labelled `batch-size-1` and `batch-size-4`, respectively. These model are trained to positively associate two sequential paragraphs and two non-sequential paragraphs. During article generation, the model constructs a query using the first paragraph of the article, retrieves the most similar paragraph and appends it to the query. This process repeats as shown in the Retrieval and Generation Process section, with an example using the Wikipedia article on Transnistria in the Article Generation Example section.

## Model Accuracy

### Reciprocal Ranking for Next Paragraph Predictions

Figure 1

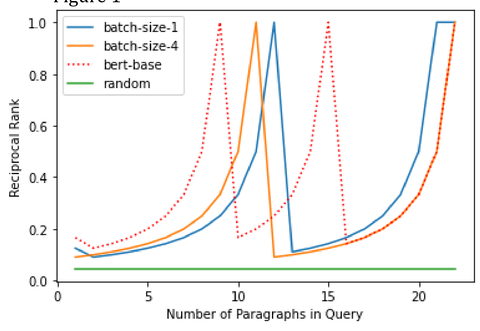



Fig.1 shows reciprocal rankings of how well each model predicts following paragraphs for the Wikipedia article on Transnistria. A value of 1.0 denotes that the model correctly predicts the next paragraph. Table 1 below shows the mean of the Reciprocal rankings for this example.

| Model | Mean Reciprocal Rank |
|---|---|
| `bert-base-uncased` | 0.36034 |
| `batch-size-1` | 0.31699 |
| `batch-size-4` | 0.27453 |
| `random` | 0.04545 |

Table 1

## Conclusions

While this research is still on-going, there are a few take-aways from the current results. The first being that each model ranks the correct paragraph more accurately towards the end of article generation. This is to be expected, as there are less paragraphs to choose from as article generation nears completion. Second, training over a larger data set may contribute to more successful results, as 198 Wikipedia articles is a small subset of the entire Wikipedia corpus. Finally, modifications to how the model is trained to consider two paragraphs as dissimilar should be adjusted, as the reciprocal ranking is low at the beginning of generation, so the model considers several other incorrect paragraphs as better choices.

## Data Set and Training

Training and evaluation are done through HuggingFace's pre-built Trainer. The data set that is utilized is a subset of the Year 1 TREC Complex Answer Retrieval (TRECCAR) Data Set. It is a collection of English Wikipedia articles, with fine-tuning performed using `train-200`, a set of 198 Wikipedia articles from TRECCAR Year 1. In the ~200 Wikipedia articles in this data set, there is a total 7137 paragraphs from which 13878 total pairs of sequentially correct and incorrect paragraph pairs are generated. An 80/20 train/evaluation split is constructed over this data set and training and evaluation is done through HuggingFace's pre-built Trainer. Accuracy of the generated articles is determined using Mean Reciprocal Rank, a metric that describes how well the model correctly ranks the next paragraph on average. One example of each model's accuracy on generating the Wikipedia article for Transnistria can be seen in the Model Accuracy section.

## Article Generation Example

### Ground and Random Baseline

| Model | Article |
|---|---|
| `ground` | The political status of Transnistria...<br>Moldova lost de facto control of...<br>The two main political parties in...<br>Only three polities recognize... |
| `random` | The political status of Transnistria...<br>The Soviet Union in the 1930s..<br>The national poet Mihai Eminescu...<br>At the dissolution of the Soviet... |

The first four paragraphs of the Wikipedia article on Transnistria. `ground` is the Wikipedia article itself, and `random` is a randomly generated reconstruction starting with the correct first paragraph.

### Base BERT and our Models

| Model | Article |
|---|---|
| `bert-base-uncased` | The political status of Transnistria...<br>At the dissolution of the Soviet Union...<br>According to PMR advocates...<br>The two main political parties... |
| `batch-size-1` | The political status of Transnistria...<br>Despite the self-proclaimed independence...<br>At the dissolution of the Soviet Union...<br>During the 1992 War of Transnistria... |
| `batch-size-4` | The political status of Transnistria...<br>Despite the self-proclaimed independence...<br>At the dissolution of the Soviet Union...<br>During the 1992 War of Transnistria... |

The first four paragraphs retrieved by each model. `bert-base-uncased` is the pre-trained BERT model, and `batch-size-1` and `batch-size-4` are the fine-tuned models.

## Acknowledgements

Thesis Advisor: Professor Laura Dietz
dietz@cs.unh.edu

Academic Advisor: Professor Marek Petrik
marek.petrik@unh.edu

## References

Devlin, Jacob, et al. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." ArXiv:1810.04805 [Cs], May 2019. arXiv.org, http://arxiv.org/abs/1810.04805.

Dietz, Laura, et al. TREC Complex Answer Retrieval Overview. https://trec.nist.gov/pubs/trec26/papers/Overview-CAR.pdf.

Wolf, Thomas, et al. "HuggingFace's Transformers: State-of-the-Art Natural Language Processing." ArXiv:1910.03771 [Cs], July 2020. arXiv.org, http://arxiv.org/abs/1910.03771.

GitHub Repository: https://github.com/JoeGuidoboni/TRECCarBERT