Sumarização de Contos do Dataset *Tiny Stories*

Caio Ribeiro de Paula
Prova Final: *Natural Language Processing*
Insper
São Paulo, Brasil

I. Visão Geral do Projeto

Neste projeto, realizou-se um processo de sumarização de contos utilizando uma porção do dataset *Tiny Stories [1]*, composto exclusivamente por textos curtos, mas divididos em várias sentenças. O desafio principal foi desenvolver uma sumarização que capturasse a essência de cada história de forma concisa. Assim como no artigo *Fine-tune BERT for Extractive Summarization [2]*, a proposta envolveu técnicas de sumarização extrativa, mas também técnicas de sumarização abstrativa usando BART.

II. Implementação

A. Adaptação do Modelo BERT para Sumarização Extrativa

O projeto inspirou-se diretamente no artigo ao explorar o uso de BERT para tarefas de sumarização. Para selecionar as sentenças mais representativas de um conto em específico do dataset, o projeto utilizou o algoritmo KMeans, seguido da análise de similaridade coseno com a média dos embeddings das sentenças. Enquanto o artigo foca em modificações estruturais no BERT para realizar essa tarefa, o projeto optou por uma abordagem prática, utilizando algoritmos externos de agrupamento e similaridade. Além disso, o projeto incorporou técnicas de visualização, como mapas de calor e projeção PCA, para explorar a relação semântica entre as sentenças de cada conto. Como cada conto era sobre um determinado tema ou acontecimento, não havia, na maioria das vezes, grande variedade de clusters. Confira exemplos de sentenças tidas como as mais representavivas após a extrativização do texto usado como exemplo no código:

- *Lily wanted to share the needle with her mom, so she could sew a button on her shirt.*
- *Together, they shared the needle and sewed the button on Lily's shirt.*

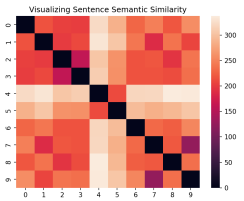

Fig. 1. Similaridade Semântica das Sentenças

B. Ajustando com a Sumarização Abstrativa

Para resolver a limitação da sumarização extrativa, que se restringe a selecionar sentenças relevantes de um texto para formar um resumo, foi incorporada também uma sumarização abstrativa no projeto, que visa gerar resumos mais naturais e condensados, abordando o conteúdo de forma mais sucinta e direta no resumo produzido.

A sumarização abstrativa tem a capacidade de gerar novas frases que não estão necessariamente presentes no texto original, mas que mantêm o significado essencial do conteúdo. Para isso, foi utilizado o modelo BART (Bidirectional and Auto-Regressive Transformers), um modelo de linguagem poderoso para tarefas de geração de texto, combinando os transformadores bidimensionais (como o BERT) com a capacidade de gerar texto fluente, ao estilo de modelos como o GPT. Esta etapa começou com a criação de um dataset personalizado para treinar o modelo BART.

C. Fine-Tuning e a Geração do Resumo

Com o dataset pronto, o modelo BART, que foi pré-treinado pela Facebook AI, é carregado e fine-tuned com os dados do projeto. O fine-tuning ajusta os parâmetros do modelo para a tarefa específica de sumarização, utilizando os textos do dataset de treino e seus resumos correspondentes, adaptados manualmente para melhor orientação do estilo do resumo. Esse processo é realizado em múltiplos ciclos de treinamento, durante os quais o modelo aprende a gerar resumos a partir do conteúdo fornecido.

Dessa forma, o modelo BART é utilizado para gerar um resumo do conto. Ao receber um texto de entrada, o modelo gera um resumo condensado que reflete o conteúdo mais importante, mas de uma maneira mais fluida e compacta. Isso é diferente da sumarização extrativa, que simplesmente seleciona partes do texto original sem alterá-las, havendo alteração aqui, ainda que, por vezes, mínimas ou inexistentes, a depender do quão sucinto o texto original é. Veja como ficou o resumo final do conto, que pode variar em diferentes testes:

- *Lily wanted to share the needle with her mom, so she could sew a button on her shirt.*

D. Avaliação da Qualidade com ROUGE

Para avaliar a qualidade dos resumos gerados, seguindo a abordagem utilizada no artigo, o código utiliza as métricas ROUGE (Recall-Oriented Understudy for Gisting Evaluation). Essas métricas comparam o resumo gerado com o resumo real

em termos de sobreposição de n-gramas (ROUGE-1, ROUGE-2) e subsequências mais longas (ROUGE-L), proporcionando uma medida quantitativa da qualidade do resumo.

Para este teste, em específico, os resultados foram perfeitos (1.0000 para Precision, Recall e F1) indicando que o resumo gerado corresponde exatamente a uma sentença do resumo de referência. Isso pode ocorrer quando a sentença do resumo é suficiente para sumarizar o seu conteúdo geral, como foi neste caso de estudo. No entanto, em casos mais gerais, o modelo gera resumos mais fluentes, resultando em valores ROUGE imperfeitos, já que o resumo varia do texto original, tanto em palavras quanto em estrutura.

III. Conclusões

Em suma, o projeto demonstrou eficácia nas abordagens de sumarização extrativa e abstrativa na sintetização de contos como o do dataset utilizado. A combinação dessas técnicas provou ser valiosa para gerar resumos com desempenho excelente nas métricas ROUGE. Esse projeto não só validou a aplicabilidade de modelos como o BERT e BART em tarefas de sumarização, mas também ofereceu perspectivas sobre como diferentes técnicas podem ser combinadas para aprimorar a qualidade dos resumos gerados.

References

[1] S. Sarkar, "Sentiment Analysis for Mental Health: Unlocking Mental Health Patterns through Statements," 2024. Disponível em: https://www.kaggle.com/datasets/thedevastator/tinystories-narrative-classification. Acesso em: Nov. 20, 2024.

[2] LIU, Yang. Fine-tune BERT for Extractive Summarization. arXiv, 2019, arXiv:1903.10318v2 Disponível em: https://arxiv.org/pdf/1903.10318. Acesso em: Nov. 20, 2024.