BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Text Summarization
Project KNN

Juraj Dedič, xdedic07@stud.fit.vutbr.cz
Matej Horník, xhorni20@stud.fit.vutbr.cz

May 10, 2024

Contents

1 Introduction

The KNN project at VUT FIT focuses on text summarization for Czech newspaper articles using large language models (LLMs). The primary objective is to improve the quality of summarization models for this task. The project involves fine-tuning various LLMs, exploring different prompting techniques, and evaluating the results using standard metrics.

2 Methodology

2.1 Datasets

The main dataset used for training and evaluation is the SumeCzech dataset, which contains Czech newspaper articles with their summaries. To augment the training data, two additional datasets were utilized:

1. **Databricks/databricks-dolly-15k**: This dataset was fully translated using the Facebook M2M_100_1.2B model.
2. **Open-Orca/SlimOrca-Dedup**: For this dataset, the OPUS MT models were used for translation, along with a mix of the Facebook model. However, only 18,000 out of 300,000 samples were successfully translated.

2.2 Fine-tuning and Prompting

The project involved fine-tuning several LLMs using the LoRA (Low-Rank Adaptation) technique. Different prompting templates and ideas were explored to improve the summarization performance. The fine-tuned models included Mistral-7B, LLama3-8b, and csmpt7b.

3 Evaluation

The summarization quality was evaluated using several metrics, including RougeRAW, Rouge, BertScore, BLEU, and METEOR. The evaluation was performed on the SumeCzech test dataset, with a subset of 150 examples. The results for each model and prompting strategy are presented in the following sections.

3.1 Mistral-7b-instruct-v0.2

The instruction-tuned version of Mistral 7b is reasonably accurate for zero-shot tasks. The main problem in our case was the use of the model in Czech language tasks which resulted in significantly worse results.

3.2 LLama3-8b-instruct

This model was released shortly before the end of our project development. Although trained mostly on English corpora, the model has shown the best results on our task.

3.3 LLama3-8b-instruct-finetuned

The fine-tuning was initially planned for 2 epochs with a matrix rank of 16. Due to time constraints, it was interrupted after the first epoch. There wasn't a significant improvement in the RougeRaw metric.

3.4 csmpt7b

The next model we used was csmpt7b pre-trained by BUT FIT. We chose it was the largest model with the largest corpora in the Czech Language, and in our tests, it performed better than BUT-FIT CSTinyLlama-1.2B.

3.5 csmpt7b-finetuned

We fine-tuned this model in multiple ways. Either using a summarization dataset, instruction dataset, or both. The final fine-tuning consisted of both rank and alpha 64. We observed a significant improvement in all of the metrics after fine-tuning of this model.

4 Conclusion

We found that for this task, using Llama-3 is the best option. Possibly better results could be achieved by effectively fine-tuning Llama-3, perhaps using a larger rank. The second finding is that BUT model csmpt 7b can be effectively fine-tuned for the summarization task, although the performance does not exceed Llama-3.