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 13, 2024

Contents

1 Introduction 2

2 Methodology 2
2.1 Datasets 2
2.2 Fine-tuning and Prompting 2

3 Training 2

4 Evaluation 3
4.1 Mistral-7b-instruct-v0.2 3
4.2 LLama3-8b-instruct 3
4.3 LLama3-8b-instruct-finetuned 3
4.4 csmpt7b 3
4.5 csmpt7b-finetuned 4
4.6 CSTinyLlama-1.2B 4

5 Conclusion 4

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 corresponding headlines and abstracts. This dataset was created by scraping czech news article websites. For training we used 45 000 samples from dev split and for testing we used samples from test split. Since the dataset is already well structured we didn't perform any further postprocessing.

We also tried to fine-tune czech LLMs which are not trained for following instructions or chat style conversation. Since we realized the best approach to have a all-round model for any kind of summarization is to have a large llm with great language understanding that can follow instructions. This will enable allow for summarizing data in any kind of domain and can be further improved with specific prompt instructions. We tried to finetune czech LLMs for instruction following with the following datasets.

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, CSTinyLlama-1.2B and csmpt7b.

Each model required different prompt as for training and also for base evaluation. Mistral only supports user and assistant without a system prompt. For czech models we used following template:

```
### instruction
{given instruction}

### input
{input if present}

### output`
{desired output}
```

As these are general-purpose language models not specifically trained on conversational data, we opted for a simple and straightforward prompt format without incorporating special tokens or explicit conversational cues.

For finetuning LLama and Mistral model we used their provided templates generated a few different instructions so there is more variability for each task, that is summarizing from text to abstract and to headline and also from abstract to headline.

3 Training

The training process for fine-tuning the language models was carried out on the MetaCentrum computational infrastructure, utilizing an NVIDIA A40 GPU. The Hugging Face libraries were employed for training and working with

LoRA (Low-Rank Adaptation) adapters. Various training parameters were explored, including batch size, number of epochs, learning rate, weight decay, and gradient accumulation steps. For LoRA, different rank (r) values ranging from 16 to 512 were used, while the alpha value was chosen to maintain the recommended range of $(\alpha \cdot r)/\alpha = (1, 2)$. The following LoRA adapters were trained during the project:

```
csmpt7b-ft-SumeCzech-v0.1
csmpt7b-ft-SumeCzech-v0.2-r64-64
csmpt7b-InstructCS-r128-256
CSTinyLlama-1.2B-ft-Instruct-r64-128
CSTinyLlama-1.2B-ft-InstructCS-r128-256
CSTinyLlama-1.2B-ft-InstructCS-r256-128
CSTinyLlama-1.2B-ft-SumeCzech-v0.1
CSTinyLlama-1.2B-ft-SumeCzech-v0.2
Mistral-7B-Instruct-v0.2-ft-r64-10k
Mistral-7B-Instruct-v0.2-ft-r16-10k
Mistral-7B-Instruct-v0.2-ft-r512-10k
Llama-3-8b-ft-Sumeczech-r16-32
```

The fine-tuning process involved either the SumeCzech dataset or the instruction datasets (Databricks/databricks-dolly-15k and Open-Orca/SlimOrca-Dedup) for the Czech language models. The number of epochs was typically kept between 1 and 5, with the training time ranging from 2 to 4 hours on average, depending on the model and dataset size. During fine-tuning with LoRA, all linear layers were fine-tuned, not just the attention matrices. The training process aimed to improve the language models' performance on the text summarization task for Czech newspaper articles, exploring different prompting techniques and fine-tuning strategies to enhance the summarization quality.

4 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.

4.1 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. As for testing, when using czech language, it had a problem with generating stop token and generated until stopped. This wasnt fixed by any of our finetunes.

4.2 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. It showed great understanding in following specific instructions in czech.

4.3 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 or any other metrics.

4.4 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.

When fine-tuning using LoRA, we noticed a potential limitation due to the model's fused attention matrices. Typically, these model architectures *MPT* combine multiple attention matrices into a single matrix, forcing the adapter to train only one weight matrix for the entire attention block. This approach may restrict the model's ability to effectively capture and adapt to the nuances and complexities present in the target domain or task during fine-tuning.

4.5 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.

4.6 CSTinyLlama-1.2B

Smaller 1.2 billion parameter model trained on czech corpus is CSTinyLlama. We used the same training approach as for csmpt7b, since these 2 models are only base models.

We didn't manage to get any great results. We could see that sometimes it could resemble a instruction following model but this happened only on a few prompts and was mostly hallucinating after our training.

5 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.

All the code is available on GitHub: `https://github.com/SyntaxErr0r1/KNN`