FRI Natural language processing course 2024
Project report

University *of Ljubljana*
Faculty *of Computer and*
Information Science

Slovenian Instruction-based Corpus Generation

Andraž Čeh, Tilen Miklavič, Tom Sojer

Abstract

This study focuses on enhancing the capabilities of Large Language Models (LLMs) to interpret and execute instructions in the Slovene language accurately. As AI assistants become increasingly prevalent across various sectors, there's a need for these technologies to support a wide range of languages, including less commonly supported ones like Slovene. The project focuses on assembling and fine-tuning a dataset of Slovene conversations to enhance the performance of a multilingual LLM in understanding and responding in Slovene.

Keywords
data gathering, corpus, finetuning

Advisors: Slavko Žitnik

Introduction

The NLP field has for a very long time focused on training different models for different tasks. Some of these tasks may include text classification and question answering. Since training these models is resource-intensive and by extension also expensive, a new field has emerged. Large language models (LLMs) have gained popularity in recent years. Just like other models, these too are trained on some corpus of text, but their goal is to create a more general-purpose text-to-text model, which could then further be fine-tuned to serve specific needs. What we will do in our project is to gather quality Slovene conversation text input, organise it and then use it to fine-tune one of the existing LLMs.

In order to broaden our knowledge about the project, we have looked at three articles from this field.

Related work

0.1 Training Language Models with Human Feedback

The authors of this study implemented a methodology for improving language model performance by incorporating human feedback into the training process[1]. This method began with collecting a set of prompts, both from labellers and the OpenAI API, to fine-tune the GPT-3 model using supervised learning. Subsequently, they gathered a dataset of model output rankings based on their alignment with provided instructions for further refinement through reinforcement learning.

This reinforcement learning phase utilized human feedback as a reward signal, enhancing the model's ability to follow user instructions and implicit expectations. The refined model, named InstructGPT, showed improved performance in executing instructions, as confirmed by human evaluators. Outputs from InstructGPT were consistently preferred over those from the base GPT-3 model, highlighting the success of the fine-tuning approach.

This approach demonstrates the potential of using human feedback to train language models for better understanding and executing instructions, representing a notable step towards models that align more closely with human intentions and produce contextually relevant responses.

0.2 Llama 2: Open Foundation and Fine-Tuned Chat Models

The article from Hugo Touvron, Louis Martin and Kevin Stone [2] talks about the development process and the release of several pre-trained large language models by the name of Llama 2. Mentioned models have anywhere from 7 to 70 billion parameters. These models have been optimised for dialogue and are surpassing other LLMs across multiple benchmarks. The article goes in-depth with explaining the pretraining process, fine-tuning and safety considerations of said model. They have explained the process of cleaning and optimising the data that was used for training the model. They also talk about increasing the pretraining corpus. For fine-tuning, they have used supervised fine-tuning, reinforced learning with human feedback and Ghost Attention, this is the new method they have used. The article also covers the observations made during the development of the Llama 2 model. It talks about the organisation of knowledge, limitations and ethical considerations. It also covers the release strategy and emphasises the importance of having such a strategy for this kind of model. The article concludes with practical and consumer applications of the developed model.

0.3 BLOOM: A 176B-Parameter Open-Access Multilingual Language Model

In order to make large language models more accessible to the public, researchers work towards developing open-source solutions. One solution is BLOOM, a 176 billion parameter language model that was built by hundreds of researchers [3]. This decoder-only transormer language model was trained on the ROOTS corpus which consists of 1.6 TB of text and spans 46 natural languages as well as 13 programming languages. The corpus is a collection of 498 Hugging face datasets [4]. First, the eight largest languages by the number of speakers were chosen, and then additional languages were added based on the knowledge of contributors. After data was collected from various sources like the web, Github and large PDF archives, the obtained dataset was preprocessed. This step included deduplication, filtering non-natural language text and removing personal information. The authors present the architectural decision of building such a model. For instance, an additional normalization layer was added after the embedding layer which improved the training stability. The tokenizer on the other hand, was designed to ensure accurate tokenization across all the languages used. BLOOM was trained on a French-funded supercomputer for 3,5 months. To present how the model compares to existing LLMs, it was evaluated and trained against a set of models like mGPT, GPT-Neo, M2M, etc.

The model has shown significant performance when evaluated against other existing models. BLOOM showed that it is excellent at creating language-diverse content. It also represents a great asset as a code assistant. It has become a useful option due to its multi-lingual nature and the fact that it is open source.

Initial ideas

The project will consist of multiple stages. Firstly, we will determine which large language model should be used. It is important to first decide on data collection as data should be as suitable for the chosen model as possible. The decision on whether to use Llama 2 or Bloom as the LLM to finetune will be based on multiple factors, such as resources available for finetuning, model performance metrics, architecture and also community support.

Secondly, we will need to gather a large set of data in Slovene. This includes first implementing a web crawler that while moving through pages, collects data. The program should crawl through different forums (mojforum.si, med.over.net), web portals (MMC, 24ur) and other sources. Because the obtained data will contain a lot of duplicated text as well as personal information, we will preprocess the data using different methods. For this part, we will need to follow existing articles that already describe the process. The obtained corpus should be usable to finetune an existing model.

Lastly, and if time permits, we will use the generated corpus and finetune the model. The results will be evaluated by comparing the model to existing LLMs.

References

[1] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022.

[2] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

[3] Bloom: A 176b-parameter open-access multilingual language model, 2023.

[4] BigScience Workshop. Bloom (revision 4ab0472), 2022.