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

Slovenian Instruction-based Corpus Generation

Žan Horvat, Bine Markelj, Anže Glušič

Abstract

This paper presents a comprehensive approach to constructing a specialized (conversational) corpus for the Slovene language, aimed at enhancing the performance of Mistral large language model through fine-tuning with locally relevant data. Recognizing the need for a rich linguistic dataset that captures the nuances of Slovene, we target content from two of the largest Slovenian online forums, known for their wide range of topics and dialectical diversity.

Keywords

natural language processing, large language models, corpus dataset, GPT, model fine-tuning

Advisors: Slavko Žitnik

Introduction

Slovene falls into the category of languages that are underrepresented in the digital landscape and has not been the focus of (extensive) NLP research. Recognizing this, the paper aims to address the problem by leveraging the rich linguistic content available in Slovenian online forums for corpus generation that can be used for fine-tuning large language models. In this paper, we will try to generate a high-quality conversational corpus that can be used to train Mistral models.

Related Work

The task of fine-tuning language models to perform specific tasks and the methods and datasets used for such training have been explored and presented in multiple articles.

In the article Training language models to follow instructions with human feedback [1] they showed, that truthfulness and usability of the model improves with the use of human feedback in the evaluation. That way the model gets more aligned with our intent (for example of making it a conversational agent). Such supervised training can have up to 100 times fewer parameters and return better outputs than far larger models without the human evaluation.

Similarly, authors of the article Llama 2: Open Foundation and Fine-Tuned Chat Models [2] also emphasize the importance of RLHF – reinforcement learning with human feedback, where the model is penalized for bad answers and rewarded for good ones. The rating of the outputs is done by humans. Such supervised technique greatly improves the model and also makes it more safeguarded for the user.

Creators of the Bloom LLM [3] put a great emphasis on the quality of the fine-tuning dataset. In their case, they collected lots of data from 539 reputable sources to build their large ROOTS corpus. Afterwards, they also performed filtering of the non-natural language and non-desired parts of websites. The last step of their corpus creation was deduplication of the data and removing of any personal or identifiable information. The corpus created in such a way was the basis for the training of their LLM.

Methods

0.1 Mistral

For testing our corpus, we chose **Mistral-7B** (accessible from *https://huggingface.co/*), which nice implementation documentation and fine-tuning capabilities. This model is claimed to be the most powerful language model for its size to date. We chose it over GPT, as it is more accessible and can be fine-tuned locally.

It was also shown, that Mistral-7B can be fine-tuned (with no proprietary data) for chat application, and it outperformed all models with 7 billion parameters on MT-Bench [4]. It's comparable to models with 13 billion parameters.

As we couldn't find any indication of how much of pretraining was done on Slovenian language, we also leave possibility of using *"CohereForAI/aya-101"* model from *Hugging face*. If time allows us, we will test our data on both of them and compare the results.

Corpus generation (Dataset)

Focus of this paper is on corpus generation for Slovene language. Our goal is enhancing the performance and local relevance of large language models. We are going to construct a comprehensive corpus by leveraging the rich linguistic landscape of Slovenian online forums. To achieve this, we have chosen to scrape content from two largest Slovenian online forums: med.over.net and slo-tech.

Decision to choose these two forums are based on several advantages. These forums are a great collection of natural language data because they encompass a range of *topics, dialects* used in daily communication of Slovenes. Additionally, the sheer volume of user-generated content available on these platforms guarantees a substantial dataset, critical for training. Our plan for data gathering includes:

1. **Categorization of forum sections that are likely to yield diverse linguistic data**
2. **Web scraping** to extract text data
3. **Cleaning process**: remove any non-textual elements (HTML tags for example)

The generated corpus will serve as a foundational dataset for fine-tuning GPT-model specifically for the Slovene language. It is expected that this effort will significantly enhance the model's performance. Crawler was configured to crawl each individual forum topic within **slo-tech.com**, capturing the entirety of discussions within those topics.

For each topic, Crawler extracted forum content to JSON format, with each entry structured as:

```
{role: "assistant" or "user", text: ...}
```

This structured format facilitates easy parsing and analysis of the forum data. By crawling each forum topic comprehensively, we aimed to capture a broad spectrum of discussions and interactions present in the forum. The generated corpus will serve as a foundational dataset for fine-tuning our model specifically for the Slovene language. It is expected that this effort will enhance the model's performance in Slovene language.

Pre-processing

Corpus pre-processing is a crucial step in natural language processing (NLP) tasks. It involves preparing the raw text data extracted from sources like online forums for further analysis or model training.

At this stage we prepared a basic preprocessing pipeline. Firstly we remove all threads, where only 1 person is talking.

We want each thread to be its own conversation, we can feed to the model. We rename the person asking the initial question/starting conversation to "user". And every further person answering them is named "assistant_n". We also find any further mention of each person in conversation and rename them there also, so the model will know who each person is talking to.

We also remove extra empty spaces and clean the text of some additional double symbols. Such data is then saved in the correct form for model fine-tunning.

This is a very basic preprocessing, but we are also exploring possibilities of using augmentation on slovenian data to create a larger corpus from fewer data, slightly changing it.

The next logical step is also ranking of responses. We would rank them according to other user's input (upvotes, likes,..), but this is not possible on the data from slo-tech.com. A step we will explore in the future development is also discussion classification, so our corpus would be more clearly seperated to different thematics (health, dangerous discussions, technical discussions, ..).

Fine-tuning Mistral model

Once we collect the scraped data into a well organized corpus, we will use it to fine-tune the Mistral model. Our goal will be to make a Slovene conversational agent.

Our fine-tuning approach would be based on the underlying understandings of the specifics of the Mistral/Aya model. We will have to observe, and evaluate, how our new data affects the performance of the model. The corpus would be organized into a structure that resembles real-world conversations (questions-answers, dialogue, ...), which will fine-tune our model to become proficient in flowing human-like conversation.

Based on the performance of such fine-tuned model (evaluated with methods described in the next subsection) we will then change and optimize our corpus structure and training methods to be efficient and capable of training Slovene conversational agents.

Results

Evaluation methods

Evaluating the performance of our fine-tuned model is essential. We could use *human evaluation* for measuring performance of our model. Human evaluation of a model involves engaging native speakers (and/or linguists) to assess the quality of outputs. We would ask a human evaluator which output they prefer more (given outputs by both pre-trained and our fine-tuned model for the same prompts) [1]. We could also assess fluency and coherence, relevance and contextual understanding.

Discussion

This will be place for discussion in the final report.

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] BigScience Workshop, :, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim

Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. Bloom: A 176b-parameter open-access multilingual language model, 2023.

[4] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.