Natural Language Processing 2023
Second submission report

Slovenian Language Assistance Bot (SLAB)

Luka Škodnik, Robert Jutreša, Valter Hudovernik

Abstract

Keywords
Large Language Models, Machine Translation, Conversational AI, Slovene, Question Answering

Advisors: doc. dr. Slavko Žitnik

Introduction

Natural Language Processing (NLP) is an exciting field of Artificial Intelligence that focuses on teaching machines how to understand and respond to human language. In this report, we will discuss our project for the Natural Language Processing course 2022/23, where we aim to develop a Slovenian Language Assistance Bot (SLAB) using NLP techniques. We will provide an overview of our preliminary research, current implementation and explore various data sources that we plan to use and discuss possible models to train or finetune our model.

Related Work

Transformers & Large Language Models

Attention [1] is a key component of transformer models, which are widely used in natural language processing tasks such as language translation and text summarization. It is also a part of many modern large language models (LLM) that we will use during this project. In transformer models, attention mechanisms allow the model to focus on the most relevant parts of the input sequence at each step of processing. This is achieved by assigning weights to each element in the input sequence based on its relevance to the current step. By doing so, the transformer can capture long-range dependencies between different parts of the input sequence, which is particularly important for language processing.

SloBERTa [2] is a Slovene large language model. It is a large pre-trained masked language model based on the BERT architecture. The authors trained the model on a large corpus of Slovene text and evaluated it on various downstream tasks such as text classification, named entity recognition, and part-of-speech tagging. The results show that SloBERTa outperforms existing Slovene language models and achieves competitive performance with state-of-the-art multilingual language models.

We examine the alignment of language model paradigm using reinforcement learning such as InstructGPT [3]. However due to the lack of resources we instead focus our attention to using already compiled high quality datasets for a one-time training of our model. However their work points out the total number of collective prompts used in training their model, which was $77k$. This gives us an approximate estimate of how much data to use in the fine-tuning process. Additionally they discuss deduplication, however in future works such as [4] have pointed out that repeating prompts is not a big issue.

Datasets

The P3 dataset is a collection of prompted English datasets covering a diverse set of NLP tasks. The use of prompts allows for the creation of consistent and standardized data examples across different datasets, which can facilitate the development of new models and the comparison of results across different tasks [5].

A corpus of automatically annotated TV series subtitles for dialogue construction was developed in [6]. The authors used a combination of rule-based and machine-learning techniques to identify speaker turns and assign speaker identities. They evaluated their method on a corpus of subtitled TV series episodes and achieved high accuracy in speaker identification and turn-taking recognition. The resulting corpus was used for various downstream tasks such as emotion recognition and dialogue act classification.

GOS [7] is a reference corpus of spoken Slovene language. The methodology used to collect the corpus involved recording conversations of native Slovene speakers in various domains such as business, education, and social interactions. The transcription process involved annotating the recordings with orthographic, phonetic, and prosodic information. The resulting corpus was used for various research tasks such as acoustic modeling, speaker recognition, and speech synthesis.

Alpaca [4] by Standford University is a set consisting of a data generation procedure, dataset, and training recipe. It is a fine tuned model from 7B LLaMA [8] on 52K instruction-following data generated by the techniques in the Self-Instruct [9] paper, with some modifications. In a preliminary human evaluation, it was found that the Alpaca 7B model behaves similarly to the GPT-3 model on the Self-Instruct instruction-following evaluation suite.

BLOOM [10] a multilingual LLM was trained on the ROOTS corpus [11], amounting to 1.61 terabytes of text that span 46 natural languages and 13 programming languages. Unfortunately Slovenian is not one of the available languages. However, due to vast collections of datasets available such as Hugging Face datasets [12], we focus on developing tools for robust translation in order to translate the large amounts of available data to Slovenian in order to facilitate the development of a conversational LLM.

SuperGLUE [13] is a benchmark for evaluating the performance of natural language understanding models. It consists of eight challenging natural language understanding tasks, including both textual entailment and commonsense reasoning tasks. The benchmark is designed to test the ability of models to handle more complex linguistic phenomena and to generalize to new examples. The authors compare the performance of several state-of-the-art models on SuperGLUE and find that there is still a significant gap between the best models and human performance.

The paper [14] discusses the challenges of question answering for less-resourced languages and presents an adaptation of the English UnifiedQA approach to the Slovene language. The adaptation uses encoder-decoder transformer models (SloT5 and mT5) to handle different question-answering formats, and existing Slovene adaptations of four datasets, as well as machine translation of the MCTest dataset. The study shows that a general model can perform at least as well as specialized models for answering questions in different formats. However, the performance of the Slovene model still lags behind that of English, and cross-lingual transfer from English is used to improve the results.

The paper ”OpenAssistant: Aligning Large Language Models with Human Preferences using Open-Source Conversations” [15] presents a new corpus of human-generated, human-annotated assistant-style conversations called OpenAssistant Conversations. The corpus consists of 161,443 messages distributed across 66,497 conversation trees, in 35 different languages, annotated with 461,292 quality ratings, and was created through a worldwide crowd-sourcing effort involving over 13,500 volunteers. The authors also release their code and data under fully permissive licenses, making their work easily accessible to the wider research community.

Methods

Recently the OpenAssistant Conversations Dataset (OASST1) has been released. We focus on translating the dataset in the conversation tree form where multiple replies can be nested to form conversations. We select a subset of trees that are “ready for export” as they have deleted spam messages and do not contain low quality messages and trees with only one prompt. Since OASST1 contains messages in various languages we translated the messages only from the 4 most common languages in the dataset: namely English, Spanish, Russian and German.

We construct pipelines for translation using NLLB-200 [16] and GoogleTranslate [17] using the deep translator python library [18]. Some of the messages also contain code which was replaced with a predefined substring at the time of translation and afterwards substituted back into the translated text. In this way we do not translate the code literally (eg. Translating predefined words such as *import* or *let*.). We translate 8654 trees which totals to 78474 messages. We observe that the translations obtained using google translate are best after manual inspection of translations.

Further Plan

Alongside the translations obtained with Google Translated we will also translate the same data using NLLB-200 which we have already set up. We’re currently in the process of setting up Slovene NMT to get yet another source for translations.

When we have data translated from multiple sources we plan to compare the translations so we only keep the messages with translations of sufficient quality. To enrich the translated dataset we plan the following procedures; Use back-translation for a quick way to filter sentences that are translated well. Additionally, combine multiple translations into a single one using some metric of agreement. Manually inspecting messages for word games, help with error messages, and other semantics potentially lost in translation. Transforming the form of the messages from formal to more informal prompts using lemmatization.

After this we will use this data (possibly only a subset) to train a smaller model such as T5 to see how well it can converse in Slovene. We will also consider the use of the smallest version of LLaMA (7B) however it is possible that even that model is to big for our use case. If we are to use the LLaMA model, we are still considering additional Slovenian pretraining on the following corpora; Gigafida [19] and ccKRES [20], we also consider the Šolar corpus [21], with considerations that the written language of primary and high school students could potentially have an impact on the models performance, and GOS [7] with similar considerations about spoken language.

Our plans for evaluation are comparison of responses to ChatGPT in Slovene and OpenAssistant replies translated to Slovene. We plan to do a qualitative comparison of 100 replies for each model, which will be scored by users who will not know which model produced the reply. Additionally we consider inspecting similarity and correlations of BERT embeddings for the model replies.

Results

Discussion

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[2] Matej Ulčar and Marko Robnik-Šikonja. Sloberta: Slovene monolingual large pretrained masked language model. *Proceedings of SI-KDD within the Information Society 2021*, pages 17–20, 2021.

[3] 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.

[4] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

[5] Stephen H Bach, Victor Sanh, Zheng-Xin Yong, Albert Webson, Colin Raffel, Nihal V Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, et al. Promptsource: An integrated development environment and repository for natural language prompts. *arXiv preprint arXiv:2202.01279*, 2022.

[6] Leilan Zhang and Qiang Zhou. Automatically annotate tv series subtitles for dialogue corpus construction. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1029–1035, 2019.

[7] Darinka Verdonik, Iztok Kosem, Ana Zwitter Vitez, Simon Krek, and Marko Stabej. Compilation, transcription and usage of a reference speech corpus: the case of the slovene corpus GOS. *Language Resources and Evaluation*, 47(4):1031–1048, January 2013.

[8] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

[9] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language model with self generated instructions, 2022.

[10] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

[11] Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Moral, Teven Scao, Leandro Werra, Chenghao Mou, Eduardo Ponferrada, Huu Nguyen, Jörg Frohberg, Mario Šaško, Quentin Lhoest, Angelina McMillan-Major, Gerard Dupont, Stella Biderman, Anna Rogers, Loubna allal, Francesco Toni, and Yacine Jernite. The bigscience roots corpus: A 1.6tb composite multilingual dataset. 03 2023.

[12] Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, et al. Datasets: A community library for natural language processing. *arXiv preprint arXiv:2109.02846*, 2021.

[13] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32, 2019.

[14] Katja Logar and Marko Robnik-Šikonja. Unified question answering in slovene, 2022.

[15] Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. Openassistant conversations – democratizing large language model alignment, 2023.

[16] NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang.

No language left behind: Scaling human-centered machine translation. 2022.

[17] Google LLC. Google Translate. https://translate.google.com/?hl=sl, Accessed 2023, March 21.

[18] deep-translator 1.10.1.

[19] Simon Krek, Tomaž Erjavec, Andraž Repar, Jaka Čibej, Špela Arhar Holdt, Polona Gantar, Iztok Kosem, Marko Robnik-Šikonja, Nikola Ljubešić, Kaja Dobrovoljc, Cyprian Laskowski, Miha Grčar, Peter Holozan, Simon Šuster, Vojko Gorjanc, Marko Stabej, and Nataša Logar. Corpus of written standard slovene gigafida 2.0, 2019. Slovenian language resource repository CLARIN.SI.

[20] N Logar Berginc, M Grčar, M Brakus, T Erjavec, Š Arhar Holdt, and S Krek. *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba*. Institut za slovenski jezik Frana Ramovša ZRC SAZU, Ljubljana, 2012.

[21] Iztok Kosem, Tomaž Rozman, and Mateja Stritar. How do slovenian primary and secondary school students write and what their teachers correct: A corpus of student writing. In *Proceedings of Corpus Linguistics Conference*, pages 20–22, July 2011.