Slovenian Language Assistance Bot (SLAB)

Luka Škodnik, Robert Jutreša, Valter Hudovernik

Abstract

Slavko je car!

Keywords

LLM, ...

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, discuss our idea to use a GPT-based model, and explore various data sources that we plan to use to train our model. Finally, we will discuss the importance of estimating the amount of data required to train a model effectively.

Initial Plan

Due to large amounts of conversational datasets being available in other languages. We focus on translating the most widely used and well-composed datasets into Slovenian. For this reason, one of our main tasks is developing a framework for robust translation from English to Slovene.
However recent works, specifically the GPT-4 report by Openai has pointed out that pre-training had much more of an effect on the performance of the model as opposed to fine-tuning. For this reason, we also put an emphasis on using and looking into additional resources for unsupervised pre-training for the Slovenian language, such as the NLP project by Jurkovič et. al.

Related Work

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.

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 [2].

A corpus of automatically annotated TV series subtitles for dialogue construction was developed in [3]. 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 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 [4].

BLOOM [5] a multilingual LLM was trained on the ROOTS corpus [6], amounting to 1.61 terabytes of text that span 46 nat-

ural 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 [7], 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.

SloBERTa 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 [8].

Methods

Results

Discussion

Acknowledgments

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] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.