Building a Large Language Model for Croatian

Vanja Štefanec[0009−0007−7110−5343], Daša Farkaš[0000−0001−6394−4568], Gaurish Thakkar[0000−0002−8119−5078], and Marko Tadić[0000−0001−6325−820X]

University of Zagreb, Faculty of Humanities and Social Sciences, Zagreb 10000, Croatia {vstefane, dfarkas, gthakkar, marko.tadic}@ffzg.unizg.hr

Abstract. Large Language Models (LLMs) have demonstrated significant advancements in processing natural language tasks. Consequently, sizeable resources have been allocated to developing and comprehending them. As a immense number of monolingual and multilingual LLMs are being developed and deployed in last few years, it has become crucial to assess their capacities. This document presents a summary of the current project HR-XR-XTEND to develop a monolingual LLM for Croatian as a moderately resourced language with less than 6 million speakers in total. The paper covers the project objectives, relevant previous work, explanation why such a model is needed, the intended methodology, and expected projected outcomes.

Keywords: Large Language Models · LLM · GPT · Croatian

1 Introduction

In recent years, there has been significant progress in the development of Language Models (LMs), particularly Large Language Models (LLMs). Today, many LLMs are trained on massive datasets of natural language text and they have the capability to perform a wide range of natural language processing (NLP) tasks at a state-of-the-art level.

Most of the LLMs created so far are multilingual, i.e., they were trained using texts in different languages and the number of languages vary from 2 to 200+. One inherent problem with multilingual LLMs is the unbalanced dataset used for training. Languages with large amount of training data are usually well represented in the model, whereas languages with limited resources are constrained by the paucity of training data. In spite of the predominant general opinion that models trained in many languages (multilingual LLMs) perform better in every NLP task, studies [10], [12], [16], have shown that models trained in a few languages (bilingual/trilingual LLMs) or in a single language (monolingual LLMs) perform better in a number of NLP tasks than multilingual LLMs. There are even recent attempts to distil the monolingual models from a larger multilingual LLMs in order to come up with smaller and more efficient monolingual LLMs [14]. That is the main reason for instigating this project.

The other reason for this project is to check whether such LLM can be implemented in the production process of the industrial partner since their process

Proceedings of New Trends in Translation and Technology, pages 204–209
3–6 July, 2024, Varna, Bulgaria. ©NeTTT 2024
https://doi.org/10.26615/issn.2815-4711.2024_017

includes human translators working with translator's workbench that combines human translation with machine or machine aided translation. The produced LLM will be used for smoothing the final target language text. Later it also can be fine-tuned for different domains. Such successful implementation can serve as a role-model for companies using similar production processes.

The main objective of the project HR-XR-XTEND[1], a FSTP subproject of a Horizon Europe funded project UTTER[2], is to build a large monolingual generative pre-trained transformer model for Croatian (HR-LLM). Although Croatian has already been included in multilingual LLMs, it remains unclear whether the presence of other languages within the LLM influences their performance with Croatian, particularly across different tasks. Since at this point of development of LLMs we still lack the standardised and balanced evaluation tools for measurement of multilingual LLM's performance applied to a single language tasks, one of the ways to establish that performance baseline is to construct a monolingual LLM and use its evaluation results as baseline. Any increase or decrease in performance of a multilingual LLM could be measured in comparison to that baseline set by monolingual LLM.

The LLM type, that is targeted, stems from the European AI initiative OpenGPT-X[3] and recent achievements in building the first GPT-3 monolingual model for Hungarian [17] (PULI GPT-3SX[4]). The model was trained using the GPT-NeoX [2] implementation on a dataset of 36.3 billion tokens of Hungarian texts. The official OpenGPT-X at this moment features only a few major European languages (English, French, German, Italian, and Spanish), with over 50 million speakers, while other European languages are not covered. On the other hand, the Hungarian GPT-3 LLM demonstrates that such an endeavour could be achieved for a language with approximately 13 million speakers [8]. We propose the creation of a LLM tailored to address the needs of lower-resourced languages, focusing on those with less than 6 million speakers [7]. The success of this project could serve as proof of a concept that similar building process could be applicable to other languages with less than 10 million speakers.

2 Related Work

In this section we mention LLMs that have been trained in Croatian. All of them are multilingual, and so far there is no monolingual Croatian model built from monolingual data alone. Language-agnostic BERT Sentence Encoder (LaBSE) [4] is a BERT-based model trained for sentence embedding in 109 languages. X-MOD [13] is a MLM trained on filtered Common Crawl data containing 81 languages. This model reuses the tokeniser of XLM-R. The model has been pre-trained with language-specific modular components (language adapters). CroSlo-Engual BERT [15] is a trilingual model (110 million parameters) using BERT-

[1] https://hr-xr-xtend.ffzg.unizg.hr
[2] https://he-utter.eu
[3] https://www.aleph-alpha.com
[4] https://nytud.github.io

base architecture, trained on Croatian, Slovenian, and English corpora. Focusing on three languages, the model performs better than multilingual BERT while still offering an option for cross-lingual knowledge transfer. BERTić [9] is a joint LLM of four today distinct languages: Bosnian, Croatian, Montenegrin and Serbian. Cro-CoV-cseBERT [1] is a LLM based on the CroSloEngual BERT and has been further trained on a large corpus of Croatian texts related to COVID-19 (Cro-CoV-Texts) which contains 186,738 news articles, 500,504 user comments related to COVID-19 published on Croatian news portals, and 28,208 COVID-19 tweets in Croatian. Several additional language models (LLMs) have been developed under the InfoCoV project, such as Cro-CoV-BERTić, Senti-CoV-cseBERT, and Multi-Cro-CoV-cseBERT. These models are specifically trained on text data connected to COVID-19. The multilingual parliamentary model (XLM-R-parla) [11] is the XLM-R-large model additionally pre-trained on texts of parliamentary proceedings. Texts for the additional pre-training, 1.7 billion tokens in size, came from the ParlaMint corpus and the EuroParl corpus. EU-BERT [3] is a pre-trained BERT uncased model that has been trained on a vast corpus of documents in all 24 EU official languages and published by the European Publications Office. HR-RoBERTa [6] model is a LLM pre-trained on Macedonian using a MLM objective and has been further trained on Croatian data. The HR-GPT2 [5] is a LLM pre-trained on a large corpus of Croatian data in a self-supervised fashion to provide text generation capabilities. TwHIN-BERT [18] is a multilingual Tweet language model (250 and 550 million parameters) that is trained on 7 billion Tweets from over 100 distinct languages.

3 Objectives

The project goals are to:

1. collect at least 6 billion tokens of Croatian texts and prepare that data for LLM training;
2. create a LLM for the Croatian language using monolingual data only;
3. evaluate the LLM for downstream tasks.

To build a monolingual LLM, a large number of tokens need to be collected from various sources: corpora and text collections from the existing repositories (although often under-performing for Croatian, e.g. results from the Oscar project[5]) and newly collected data (online and offline) that were not available in previous data collection campaigns. The minimal amount of 6 billion tokens was established upon estimates of the available Croatian data collected in crawling campaigns so far, but also upon corpora collected at the University of Zagreb, Faculty of Humanities and Social Sciences in previous projects. However, we have surpassed this minimal size as presented in the Table 1. The performance of the monolingual LLM will be benchmarked for various downstream tasks such as named-entity recognition and classification (NERC), as a representative of a

[5] https://oscar-project.org/

task with lower level complexity, and sentiment analysis (SA) as a task with higher level complexity. The well-known instruction and prompting based evaluation dataset Alpaca[6] with more than 51,000 prompts has been translated from English to Croatian and is being manually checked by project collaborators.

The direct outputs of the project are the LLM for Croatian and LLM evaluation tools for Croatian. Both will be available through HR-CLARIN[7] repository with permissive licenses.

4 Experiments

The experimental phase is focused on developing and evaluating the model architecture and training process. This involves:

1. Collecting and pre-processing of the training dataset: Data will also be partially used within newly established initiatives such as Language Data Space and ALT-EDIC. Table 1 lists the large repositories or collection campaigns that we used so far as sources of data with approximate sizes in tokens.
2. Training the model: For training the model, we will use the publicly available library at gpt-neox[8].
3. Local GPU infrastructure will be used for experimentation, while EuroHPC and University of Zagreb Computing Centre supercomputers with GPU nodes will be used for advanced experimenting and final version training.
4. Configuring the training hyperparameters: hyperparameters like learning rate, batch size, number of epochs will be tuned for optimal performance.
5. Evaluating the model: the model will be evaluated using language model evaluation Harness[9] and Alpaca.
6. Additional dataset creation for fine-tuning: Datasets will be created for tasks like NERC, sentiment analysis and prompting.

5 Conclusion

Recent advancements in language technology have demonstrated that large language models are essential for reliable and efficient language processing capabilities. Imbalanced data distributions during multilingual LLM pre-training, demonstrably weakens the monolingual proficiency of a model applied to the specific monolingual tasks, particularly in a lower-resourced languages.

This paper presented the work-in-progress on the project HR-XR-XTEND. The main objectives of this project are to collect the relevant monolingual data, clean and prepare that data for training, and finally to develop a LLM for the Croatian language. Additionally, the key findings may have broader implications for lower-resourced languages in Europe and around the globe in subsequent developments.

[6] https://huggingface.co/datasets/yahma/alpaca-cleaned
[7] https://www.clarin.hr
[8] https://github.com/EleutherAI/gpt-neox
[9] https://github.com/EleutherAI/lm-evaluation-harness

Table 1. Non-exhaustive list of largest data sources to be used for training the LLM with approximate size in tokens for sources with 50+ million tokens.

Name	Approx Size
CC100-Croatian Dataset 1.0	3.3 billion
CLASSLA Hr Web corpus 1.0	2.5 billion
Corpus of Croatian News Feeds	2.25 billion
HPLT Croatian/Bosnian/Serbian Corpus, Croatian texts only	4 billion
Parallel data for En-Hr on OPUS Resources, Croatian texts only	1.48 billion
Corpus of Croatian Academic Theses	312 million
Joel Niklaus, Multi Legal Pile, Croatian	258 million
Leipzig Corpora	182.40 million
ParlaMint 4.0, Croatian texts only	88.16 million
ParaCrawl, Croatian texts only	79.06 million
hrWikipedia	66.48 million
MARCELL Croatian legislative subcorpus	56 million
Total	14.57 billion

6 Acknowledgements

This project has received funding from the European Union's Horizon Europe Research and Innovation programme under Grant Agreement No 101070631 and from the UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (Grant No 10039436).

References

1. Babić, K., Petrović, M., Beliga, S., Martinčić-Ipšić, S., Matešić, M., Meštrović, A.: Characterisation of covid-19-related tweets in the croatian language: framework based on the cro-cov-csebert model. Applied Sciences **11**(21), 10442 (2021)
2. Black, S., Biderman, S., Hallahan, E., Anthony, Q., Gao, L., Golding, L., He, H., Leahy, C., McDonell, K., Phang, J., Pieler, M., Prashanth, U.S., Purohit, S., Reynolds, L., Tow, J., Wang, B., Weinbach, S.: GPT-NeoX-20B: An open-source autoregressive language model. In: Fan, A., Ilic, S., Wolf, T., Gallé, M. (eds.) Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models. pp. 95–136. ACL, virtual+Dublin (May 2022). https://doi.org/10.18653/v1/2022.bigscience-1.9, https://aclanthology.org/2022.bigscience-1.9
3. Huggingface/europeanparliament/eubert (2023), https://huggingface.co/EuropeanParliament/EUBERT, accessed: 19-02-2024
4. Feng, F., Yang, Y., Cer, D., Arivazhagan, N., Wang, W.: Language-agnostic bert sentence embedding (2022)
5. Huggingface/macedonizer/hr-gpt2 (2021), https://huggingface.co/macedonizer/hr-gpt2, accessed: 19-02-2024
6. Huggingface/macedonizer/hr-roberta-base (2021), https://huggingface.co/macedonizer/hr-roberta-base, accessed: 19-02-2024

7. European language equality (2022), https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_7__Language_Report_Croatian_.pdf, accessed: 19-02-2024
8. European language equality (2022), https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_18__Language_Report_Hungarian.pdf, accessed: 19-02-2024
9. Ljubešić, N., Lauc, D.: BERTić - the transformer language model for Bosnian, Croatian, Montenegrin and Serbian. In: Babych, B., Kanishcheva, O., Nakov, P., Piskorski, J., Pivovarova, L., Starko, V., Steinberger, J., Yangarber, R., Marcińczuk, M., Pollak, S., Přibáň, P., Robnik-Šikonja, M. (eds.) Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing. pp. 37–42. ACL, Kiyv, Ukraine (Apr 2021), https://aclanthology.org/2021.bsnlp-1.5
10. Martin, L., Muller, B., Ortiz Suárez, P.J., Dupont, Y., Romary, L., de la Clergerie, É., Seddah, D., Sagot, B.: CamemBERT: a tasty French language model. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 7203–7219. ACL, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.645, https://aclanthology.org/2020.acl-main.645
11. Mochtak, M., Rupnik, P., Ljubešić, N.: The parlasent multilingual training dataset for sentiment identification in parliamentary proceedings. arXiv preprint arXiv:2309.09783 (2023)
12. Papadimitriou, I., Lopez, K., Jurafsky, D.: Multilingual BERT has an accent: Evaluating English influences on fluency in multilingual models. In: Beinborn, L., Goswami, K., Muradoğlu, S., Sorokin, A., Kumar, R., Shcherbakov, A., Ponti, E.M., Cotterell, R., Vylomova, E. (eds.) Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP. pp. 143–146. ACL, Dubrovnik, Croatia (May 2023). https://doi.org/10.18653/v1/2023.sigtyp-1.16, https://aclanthology.org/2023.sigtyp-1.16
13. Pfeiffer, J., Goyal, N., Lin, X.V., Li, X., Cross, J., Riedel, S., Artetxe, M.: Lifting the curse of multilinguality by pre-training modular transformers. arXiv preprint arXiv:2205.06266 (2022)
14. Singh, P., Maladry, A., Lefever, E.: Too many cooks spoil the model: Are bilingual models for Slovene better than a large multilingual model? In: Piskorski, J., Marcińczuk, M., Nakov, P., Ogrodniczuk, M., Pollak, S., Přibáň, P., Rybak, P., Steinberger, J., Yangarber, R. (eds.) Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023). pp. 32–39. ACL, Dubrovnik, Croatia (May 2023). https://doi.org/10.18653/v1/2023.bsnlp-1.5, https://aclanthology.org/2023.bsnlp-1.5
15. Ulčar, M., Robnik-Šikonja, M.: Finest bert and crosloengual bert: less is more in multilingual models. In: Text, Speech, and Dialogue: 23rd International Conference, TSD 2020, Brno, Czech Republic, September 8–11, 2020, Proceedings 23. pp. 104–111. Springer (2020)
16. Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., Ginter, F., Pyysalo, S.: Multilingual is not enough: Bert for finnish (2019)
17. Yang, Z.G., Laki, L.J., Váradi, T., Prószéky, G.: Mono-and multilingual gpt-3 models for hungarian. In: International Conference on Text, Speech, and Dialogue. pp. 94–104. Springer (2023)
18. Zhang, X., Malkov, Y., Florez, O., Park, S., McWilliams, B., Han, J., El-Kishky, A.: Twhin-bert: a socially-enriched pre-trained language model for multilingual tweet representations. arXiv preprint arXiv:2209.07562 (2022)