New Language Models for South Slavic Languages

Mihailo Škorić

the complexity of human emotions in the current digital world within the Serbian-speaking community. The corpus will be published under open license CC-BY-4.0.

Keywords: *emotions, Plutchik's model, annotation, corpus, social media, Serbian language*

Acknowledgment: *This research was supported by the Science Fund of the Republic of Serbia, #7276, Text Embeddings - Serbian Language Applications - TESLA.*

Mihailo Škorić
Društvo za jezičke resurse i tehnologije JeRTeh
E-mail: mihailo@jerteh.rs

New Language Models for South Slavic Languages

The report will present the challenges and perspectives of modeling South Slavic languages, especially the general language models built on the transformer architecture (BERT, GPT), available corpora of texts for training those models, and the quantity and quality of those corpora. The presentation will offer an overview of the available data and models, primarily the latest textual corpora. The first corpus, *Umbrella*, represents the umbrella web corpus of South Slavic languages and at the same time the largest corpus of texts in the region, includes all other currently available regional web corpora and contains over eighteen billion words. The second corpus, *S.T.A.R.S*, gathers academic works written in the Serbian language and includes, most notably, eleven thousand dissertations downloaded from the NARDUS platform, and a large number of scientific and professional works downloaded from various open repositories that are included in the *eScience* system. In addition, academic corpora of other South Slavic languages will be discussed, which were created from works stored on various web platforms: DABAR (for the Croatian language), the repositories of the universities in Maribor, Ljubljana, Primorska and Nova Gorica, and the DiRROS and REVIS repositories (for the Slovene language), the repository of the universities in Zenica, Sarajevo and East Sarajevo (for the Bosnian language), the repository of the University of Goce Delčev and St. Kliment Ohridski (for the Macedonian language) and the repository of the University of Montenegro (for Montenegrin). Finally, we will talk about new models for text vectorization in South Slavic languages, which were trained using the aforementioned corpora. An analysis of their performance on a number of previously established tasks will be presented, with reference to the model performance and improvements over models trained on the previous generation of the corpora.

Keywords: *Large text corpora, language models, South Slavic languages*

Acknowledgment: *This research was supported by the Science Fund of the Republic of Serbia, #7276, Text Embeddings - Serbian Language Applications - TESLA.*

CIP - Каталогизација у публикацији
Народна библиотека Србије, Београд

811.163'322(048)(0.034.2)
004.8(048)(0.034.2)

INTERNATIONAL Conference South Slavic Languages in the Digital Environment JuDig (2024 ; Beograd)

Book of abstracts [Elektronski izvor] / International Conference South Slavic Languages in the Digital Environment JuDig, International Conference South Slavic Languages in the Digital Environment JuDig, November 21-23. 2024, [Belgrade] ; [organisers University of Belgrade - Faculty of Philology [and] Society for Language Resources and Technologies JeRTeh]. - Belgrade : University, Faculty of Philology, 2024 (Belgrade : University, Faculty of Philology). - 1 elektronski optički disk (CD-ROM) : tekst ; 12 cm

Sistemski zahtevi: Nisu navedeni. - Nasl. sa naslovnog ekrana. - Tiraž 100.

ISBN 978-86-6153-754-7

a) Јужнословенски језици -- Рачунарска лингвистика – Апстракти

COBISS.SR-ID 157072905