Creation of a Training Dataset for Question-Answering Models in Serbian

Ranka Stanković, Jovana Rađenović, Maja Ristić, Dragan Stankov

[2] Frontini, F., Brando, C., Byszuk, J., Galleron, I., Santos, D., Stanković, R. (2021). Named Entity Recognition for Distant Reading in ELTeC. In Constanza Navarretta; Maria Eskevich (ed.), *CLARIN Annual Conference 2020,* 36-41. http://hdl.handle.net/10400.26/39114.

[3] Ikonić Nešić, M., Stanković, R., & Rujević, B. (2022). Serbian ELTeC Sub-Collection in Wikidata. *Infotheca - Journal for Digital Humanities*, 21(2), 60-87. https://doi.org/10.18485/infotheca.2021.21.2.4

[4] Krstev, Cvetana, Ivan Obradović, Miloš Utvić, and Duško Vitas. (2014). A system for named entity recognition based on local grammars. *Journal of Logic and Computation,* 24 (2): 473–489.

[5] Milinković, M. (2022). Application of TXM Tools for Spatial Plan Corpus Analysis. *Infotheca - Journal For Digital Humanities, 22*(1), 32-51. https://doi.org/10.18485/infotheca.2022.22.1.2

[6] Milinković, M. (2022a). *The development of library and language resources for organizing and finding information on spatial planning*. PhD dissertation. University of Belgrade - Faculty of Philology. https://hdl.handle.net/21.15107/rcub_raumplan_683

[7] Nielsen, F.Å., Mietchen, D., Willighagen, E. (2017). Scholia, Scientometrics and Wikidata. In: Blomqvist, E., Hose, K., Paulheim, H., Ławrynowicz, A., Ciravegna, F., Hartig, O. (eds) *The Semantic Web: ESWC 2017 Satellite Events. ESWC 2017. Lecture Notes in Computer Science*. Vol 10577. Springer, Cham. https://doi.org/10.1007/978-3-319-70407-4_36

[8] Stanković, R. (2022). Distant Reading Training School 2020: Named Entity Recognition & Geo-Tagging for Literary Analysis. *Infotheca - Journal For Digital Humanities, 21*(2), 167_171. Retrieved from https://infoteka.bg.ac.rs/ojs/index.php/Infoteka/article/view/223

[9] van Veen, Theo. (2019). Wikidata: From "an" Identifier to 'the' Identifier. *Information Technology and Libraries* 38 (2):72-81. https://doi.org/10.6017/ital.v38i2.10886.

Ranka Stanković, Jovana Rađenović, Maja Ristić, Dragan Stankov
University of Belgrade, Faculty of Mining and Geology
E-mail: {ranka.stankovic|jovana.radjenovic|maja.ristic|dragan.stankov}@rgf.bg.ac.rs

Creation of a Training Dataset for Question-Answering Models in Serbian

The development and application of artificial intelligence in language technologies have advanced significantly in recent years, especially in the domain of the task of answering questions (Question Answering - QA). While existing resources for QA tasks have been developed for major world languages, the Serbian language has been relatively neglected in this area. This work represents an initiative to create an extensive and diverse set of data for training models for answering questions in the Serbian language, which will contribute to the improvement of language technologies for the Serbian language.

In addition to the numerous research on language models in the last few years, much work has also been done on the reference datasets needed to track modeling progress. A lot has been done when it comes to answering questions and understanding what is read, although, mostly, when it comes to big languages (Rogers et al. 2023). The paper provides an overview of the various formats and domains of available multilingual and monolingual resources, with special reference to the Serbian language (Cenić & Stojković 2023; Cvetanović & Tadić 2024). We will also consider the implications that follow from an excessive focus on the English language

As part of the TESLA (Text Embeddings - Serbian Language Applications) project, we are working on the preparation of a set of data: context, questions and answers, collected from different domains. The set will be made up of three smaller ones. To create the first set, a subset of the Stanford set SQuAD (Rajpurkar et al. 2018), where the answer is a segment of text, is translated and adapted, choosing topics such as: Nikola Tesla, climate change, construction, geology, etc. The subset will have around 7000 questions with accompanying answers. The second set that is being prepared will mainly be related to environmental protection, informatics and energy and will contain about 5000 questions with answers and given context excerpted from the textbook. The third set will contain automatically generated contexts based on the content of the Wikidata knowledge base.

The questions are carefully formulated to cover different types of queries: questions that require specific facts, questions with descriptive answers (which seek explanations or descriptions), and procedural questions, that is, questions that require a series of instructions or steps as a response. Data is collected in a variety of ways and verified through a manual annotation process to ensure accuracy and relevance of responses. The lack of manually annotated datasets in the Serbian language makes the contribution of this research particularly important.

The conclusion of the paper indicates the importance and potential of the application of this data set in various fields, including educational technologies, digital assistants, and information retrieval systems. The presented results contribute to the improvement of language technologies for the Serbian language, and we hope that they will encourage further research and development in this area.

Keywords: *artificial intelligence, natural language processing, language resources, annotated sets, information extraction, question answering*

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

References

[1] Rogers, Anna, Matt Gardner, and Isabelle Augenstein. "QA Dataset Explosion: A Taxonomy of NLP Resources for Question Answering and Reading Comprehension." ACM Computing Surveys 55, no. 10 (2023): 1-45. https://arxiv.org/pdf/2107.12708

[2] Rajpurkar, Pranav, Robin Jia, and Percy Liang. "Know what you don't know: Unanswerable questions for SQuAD." arXiv preprint arXiv:1806.03822 (2018). https://arxiv.org/abs/1806.03822 , https://rajpurkar.github.io/SQuAD-explorer/

[3] Cenić, Aleksandar B., and Suzana Stojković. "A Serbian Question Answering Dataset Created by Using the Web Scraping Technique." In 2023 58th International Scientific Conference on Information, Communication and Energy Systems and Technologies (ICEST), pp. 147-150. IEEE, 2023.

[4] Cvetanović, Aleksa, Predrag Tadić, "Synthetic Dataset Creation and Fine-Tuning of Transformer Models for Question Answering in Serbian", 2024, https://arxiv.org/html/2404.08617v1, https://paperswithcode.com/paper/synthetic-dataset-creation-and-fine-tuning-of

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