6th International Workshop on Deep Learning in Computational Physics (DLCP-2022)

Contribution ID: **35**

Type: **Presentation**

Relation Extraction from Texts Containing Pharmacologically Significant Information on base of Multilingual Language Models

Thursday 7 July 2022 10:45 (15 minutes)

In this paper we estimate accuracy of solving the task of relation extraction from texts containing pharmacologically significant information on the set of corpora in two languages:
1) the expanded version of RDRS corpus, that contains texts of internet reviews on medications in Russian;
2) the DDI2013 dataset containing MEDLINE abstracts and documents from DrugBank database in English;
3) the PhaeDRA corpus containing MEDLINE abstracts in English.

Relation extraction accuracy for Russian and English was estimated with comparison of two multilingual Language models: XLM-RoBERTa-large and XLM-RoBERTa-sag-large. Additionaly we used the State-of-the-Art specialized models aimed at English language: bioBERT, bioALBERT, bioLinkBERT. Earlier research proved XLM-RoBERTa-sag-large to be the most efficient language model for the previous version of the RDRS dataset. We used the same approach to relation extraction included two steps: named entity recognition and relation extraction on predicted entities. Each step was estimated separately.

As a result, it is shown, that multilingual XLM-RoBERTa-sag model achieves relation extraction macro-averaged f1-score equal to 85.42% on the ground-truth named entities, 53.83% on the predicted named entities on new version of RDRS corpus. Additionally, XLM-RoBERTa-sag was estimated on the datasets for relation extraction in English (DDI2013, PhaeDRA) and achieves accuracy comparable with the top specialized models.

Consequently, XLM-RoBERTa-sag model sets the state-of-the-art for considered type of texts in Russian, and achieves accuracy comparable with the SotA results in Engilsh.

Agreement to place

Participants agree to post their abstracts and presentations online at the workshop website. All materials will be placed in the form in which they were provided by the authors

Authors: SELIVANOV, Anton (NRC ”Kurchatov Institute”); Dr RYBKA, Roman (NRC ”Kurchatov Institute”); Dr SBOEV, Alexander (NRC ”Kurchatov Institute”; NRNU ”MEPhI”)

Presenter: SELIVANOV, Anton (NRC ”Kurchatov Institute”)

Session Classification: Session 2. Modern Machine Learning Methods

Track Classification: Track 2. Modern Machine Learning Methods