NLPre Benchmarking Platform

Evaluation campaign proposal by Alina Wróblewska (2023-09-08, Istanbul)

Drawing inspiration from the GLUE benchmark, the NLPre benchmarking platform would be an environment for evaluating and ranking NLP models predicting sentence segmentation, tokenisation, part-of-speech tags, morphological features, lemmata, multiword expressions, dependency trees, etc. These NLP tasks may serve as the foundation components for advanced NLU tasks, and this is the rationale behind labelling them as natural language preprocessing tasks and thus naming the platform NLPre.

The platform will be language-centric, meaning that the evaluation will be carried out in a single language. On the other hand, it will be designed to cover intra-linguistic diversity, e.g. diversely annotated datasets could be used to evaluate NLP systems.

The web-based platform will encompass an assessment module and a leaderboard. The assessment module will evaluate submitted test set predictions output by external NLPre systems or LLMs. Concurrently, the platform will automatically track the quality of these predictions and rank the evaluated system/model on the leaderboard.

The leaderboard will provide insights into the advancement of preprocessing the particular language, thereby offering valuable information to developers engaged in implementing NLPre tools, as well as other advanced systems utilising these NLPre tools. The scores on the leaderboard could be accessed via an API.

The concept of NLPre benchmarking closely resembles the idea of shared tasks (e.g. the CoNLL 2018 shared task), although the distinguishing feature is the ability to continuously update the leaderboard compared to static shared task performance tables. The leaderboard will therefore contain the up-to-date and complete quality of processing a given language. As all NLPre systems or LLMs' outcomes will be eligible for submission and assessment, it could be necessary that test datasets be kept secret and unpublished to ensure the reliability of the evaluation process and prevent any potential manipulation of the ranking.