Orbis2 - A Natural Language Processing Benchmarking Framework That Supports Drill Down Analyzes

Norman Süsstrunk and **Roger Waldvogel** and **Andreas Murk**
André Glatzl and **Albert Weichselbraun**
norman.suesstrunk@fhgr.ch

Abstract

Competitive benchmarking of natural language processing (NLP) systems has contributed considerably towards improving the performance of NLP methods. Orbis2 is an open-source benchmarking framework designed towards addressing the need of (i) evaluating natural language processing systems, and (ii) obtaining insights that help in further enhancing them. The framework was developed to support the development of information extraction components in multiple Innosuisse-funded projects such as CareerCoach, IMAGINE, Job-Cockpit and Future of Work. Traditional evaluation tools like GERBIL focus on aggregated statistics such as accuracy, precision, recall, and the F1 metric which indicate the overall performance of the evaluated systems. Although these aggregated metrics are well-suited for comparing systems, they provide little help in understanding evaluation results. Orbis2 addresses this shortcoming by enabling drill down analyzes, which contextualize evaluation results (e.g., by visualizing correct and incorrect annotations within their textual context), therefore, helping researchers in better understanding the strengths and weaknesses of their systems, and in systematically addressing them. The framework integrates with existing NLP annotation tools such as Label Studio and Doccano, enabling users to seamlessly import and utilize corpora from these platforms. Orbis2 also supports a variety of evaluation types, including named entity classification, named entity linking, named entity recognition, and page segmentation. Work on complex slot-filling evaluation tasks is currently underway. The Orbis2 developer team, has invested significant efforts towards enhancing user experience. Its design allows users to easily navigate, compare and analyze metrics without the need to switch between multiple interfaces. Orbis2 is licensed under the Apache 2.0 license, publicly available on GitHub (github.com/orbis-eval), and encourages contributions from the community, to further improve and innovate in the area of visual benchmarking.