Contribution ID: **53** Type: **Full presentation: short (15 mins.)**

Using the EGI infrastructure for REPROLANG2020: reproducibility in the context of Natural Language Processing

Monday, 2 November 2020 16:15 (15 minutes)

Introduction

REPROLANG-The Shared Task on the Reproduction of Research Results in Science and Technology of Language, was organized by ELRA with the technical support of CLARIN. This initiative was aimed at eliciting and motivating the spread of scientific work on reproductibility in the area of Natural Language Processing. It built on the previous pioneer LREC workshops on reproducibility 4REAL2016 and 4REAL2018, and followed also the initiative of the Language Resources and Evaluation journal. In this presentation we describe how the computational resources of the EGI Infrastructure were used to support this initiative.

Scientific methodology

This shared task is a new type of challenge: it is partly similar to the usual competitive shared tasks—in the sense that all participants share a common goal; but it is partly different —in the sense that its primary focus is on seeking support and confirmation of previous results, rather than on overcoming those previous results with superior ones. Thus instead of a competitive shared task, this is a cooperative shared task, with participants struggling for systems to reproduce as close as possible the results to an original complex research experiment and thus eventually reinforcing the level of reliability on its results by means of their eventually convergent outcomes.

3. Technical approach

Each submission provided a link to the input dataset and to a gitlab repository containing a docker image associated with a tag. The association of the tag to the image was ensured by building the docker image via the gitlab CI pipeline on each tag. The structure of the gitlab repository, the entrypoint script and parameters and the mount points for the input and output datasets where all predefined. With all container images available and a well-defined process in place to run the submissions, we provisioned four virtual private servers on EGI infrastructure. Some of the submissions ran without issues, some had obvious errors, but others had subtle, unexpected issues, such as a dependency on specific CPU instructions which one of the VPS instances did not have. Another experiment failed due to a lack of GPU memory. Our instances had 8GB of GPU memory, after provisioning a new instance with 12GB on commercial infrastructure we were able to successfully run the submission.
To check for possible hard-coded results, we proceeded to

rerun the experiments that successfully finished before the review deadline with ablated input data.

Conclusion

Through the efforts made in the context of REPROLANG2020, we learnt that a meticulous replication process requires substantial efforts from all sides involved. In this paper we focused on the technical replication, highlighting the importance of having access to flexible and adequate computing resources. While in the end we were able to successfully complete the exercise, we also hope that our experience can contribute to further improvements of the EGI infrastructure, like better pre-configured GPGPU instances with more dedicated GPU RAM, and the inclusion of more information about the CPU models and supported instruction sets in the provisioning systems.

Primary authors: VAN UYTVANCK, Dieter (CLARIN ERIC); SILVA, João (University of Lisbon, Department of Informatics); GOMES, Luís (University of Lisbon, Department of Informatics); MOREIRA, André (CLARIN ERIC); ELBERS, Willem (CLARIN ERIC); BRANCO, António (University of Lisbon, Department of Informatics); CALZOLARI, Nicoletta (Istituto di Linguistica Computazionale, CNR); VOSSEN, Piek (Vrije Universiteit Amsterdam); VAN NOORD, Gertjan (University of Groningen)

Presenters: VAN UYTVANCK, Dieter (CLARIN ERIC); ELBERS, Willem (CLARIN ERIC)

Session Classification: Data Analytics and thematic services - part 1