INCEpTION
Corpus-based Data Science from Scratch

Richard Eckart de Castilho, Jan-Christoph Klie, Naveen Kumar, Beto Boullosa and Iryna Gurevych
Ubiquitous Knowledge Processing (UKP) Lab
Technische Universität Darmstadt, Germany
http://www.ukp.tu-darmstadt.de

In recent years, corpus-based data science has seen rapid adoption both in science and in industry. Developing corpus-based models for text mining from scratch has penetrated a huge number of application fields. This renders common approached to corpus annotation unscalable. Instead machine-assisted annotation with a human-in-the loop is becoming crucial for the adoption of NLP by data scientists.

INCEpTION [1] is a web-based annotation platform for machine-assisted annotation which provides such a tool. The platform targets users in any domain or application scenario that are in need of text that is annotated with specific categories and relations or linked to knowledge bases. It uses machine learning to provide annotation suggestions including active-learning driven guidance, thus improving annotator efficiency and quality. The modular architecture allows using different external annotation services to provide such suggestions. It supports entity disambiguation and linking, cross-document coreference, as well as fact linking using custom domain-specific RDF-based internal knowledge bases or using local or remote external knowledge bases through SPARQL. Annotation interoperability is ensured through the use of UIMA [2] as well as through the support of various annotation formats including CLARIN TCF [3]. At the level of the annotation scheme, the platform is compatible with the DKPro Core [4] type system facilitating interoperability with many of the NLP tools integrated within DKPro Core.

INCEpTION is a multi-user platform. Users assume different roles (e.g. admin, project creator, normal user) on the platform as well as in individual projects (manager, annotator, adjudicator). User authorization can be delegated to an external mechanism users to be authenticated against infrastructure identity providers. This is essential for the deployment of the platform at the level of local or national infrastructures where it is used by users from many different organizations. Being a web-based tool these geographically distributed users can also conveniently collaborate on annotation projects within the platform.

Further connectivity with other services is possible through a remote access API compatible with the OpenMinTeD AERO protocol[1] that permits the automated setup and management of annotation projects. This allows projects to embed the annotation tool into a larger annotation campaign management process. It can also be used in a classroom scenario to automatically set up and tear down projects for students.

INCEpTION is fully open-source, openly developed on GitHub and published under the liberal Apache License 2.0. It is our goal to not only develop a comprehensive semantic text annotation platform, but also to grow a community around it and thus to promote a community-driven sustainability model for the platform. We believe the high level of interoperability, the generic nature of the tool, the open development process and the liberal license are key factors in this strategy.

Acknowledgments

This work was supported by the German Research Foundation under grant No. EC 503/1-1 and GU 798/21-1 (INCEpTION).

Note

This document is a nicely rendered version of the abstract for the poster "INCEpTION - Corpus-based Data Science from Scratch" presented at the Digital Infrastructures for Research (DI4R) 2018 conference in Lisboa, Portugal[2]. We prepared it in the present format for convenient reading since the conference does not publish traditional proceedings.

References

[1] J.-C. Klie, M. Bugert, B. Boullosa, R. Eckart de Castilho, and I. Gurevych, "The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation," in *Proceedings of the 27th International Conference on Computational Linguistics - COLING 2018*, Santa Fe, New-Mexico, USA, 2018, pp. 5–9.

[2] D. Ferrucci, A. Lally, K. Verspoor, and E. Nyberg, "Unstructured information management architecture (UIMA) version 1.0," OASIS Standard, Mar. 2009. [Online]. Available: http://docs.oasis-open.org/uima/v1.0/uima-v1.0.html

[3] M. Hinrichs, T. Zastrow, and E. Hinrichs, "WebLicht: Web-based LRT Services in a Distributed eScience Infrastructure," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, Eds. Valletta, Malta: European Language Resources Association (ELRA), May 2010, pp. 489–493.

[4] R. Eckart de Castilho and I. Gurevych, "A broad-coverage collection of portable nlp components for building shareable analysis pipelines," in *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT*. Dublin, Ireland: Association for Computational Linguistics and Dublin City University, Aug. 2014, pp. 1–11. [Online]. Available: http://www.aclweb.org/anthology/W14-5201

[1] https://openminted.github.io/releases/aero-spec/1.0.0/omtd-aero/

[2] https://www.digitalinfrastructures.eu