INCEpTION: Interactive Machine-assisted Annotation

Extended Abstract

Jan-Christoph Klie
Technische Universität Darmstadt
Ubiquitous Knowledge Processing Lab
https://www.ukp.tu-darmstadt.de

The demand for high-quality annotated text corpora in science and industry has sky-rocketed over the past years. To address this need, we introduce INCEpTION,[1] a web-based platform for efficient text annotation. The platform generically supports use cases that require span or relation annotations as well as entity and fact linking. To the best of our knowledge, INCEpTION is the first text annotation platform that integrates interactive annotation support, knowledge management and offers extensibility.

Annotation support. To minimize the required human effort and to increase annotation speed and quality, possible annotations are suggested by machine learning algorithms, so-called *recommenders*. When the user accepts, rejects or corrects these suggestions, this feedback is used to update the recommender model in the background. This interactive process creates a tight feedback loop between human and machine which continually provides better suggestions. A non-obtrusive active learning mode can be used to navigate the suggestions in the order of the largest estimated improvement in recommendation quality. State-of-the art generic annotation tools only support static pre-annotations or manual intervention to trigger re-training, making INCEpTION the first generic tool to offer interactive annotation support.

Knowledge management. INCEpTION supports entity and fact linking with knowledge bases (KB), a common requirement for tasks like cross-document text discovery/exploration or knowledge base population/completion. Recommendation support is also implemented for entity linking: for existing named entity annotations, suitable KB entries to link with are suggested. When annotating new named entities, auto-completion with contextual re-ranking facilitates finding the right entity, even for very large knowledge bases like Wikidata. While some existing applications also support entity and fact linking, editing knowledge bases often can only be performed in a different tool. Integrating knowledge management in INCEpTION enables performing KB population and KB completion tasks during the annotation process.

Extensibility. By using an open and extensible architecture, INCEpTION aims to induce a shift from implementing new annotation tools for new corpus projects towards adding project-specific functionality through customization. INCEpTION provides many extension points where custom Java code can easily be called through events or dependency injection. Additionally, web-service APIs are provided, e.g. in order to support external annotation web services as custom recommenders. The latter allows users to reuse already trained machine learning models, even in different programming languages. Furthermore, the platform uses standards to be interoperable, e.g. for knowledge management, SPARQL and RDF are used, allowing straightforward access to remote KBs like Wikidata or DBPedia. Different knowledge base schemata are supported through a configurable mapping mechanism.

The FAMULUS project [1] is one of the early adopters of INCEpTION. It is an interdisciplinary effort which aims at improving the diagnostic reasoning competence of aspiring medical doctors (diagnosing diseases) and trainee teachers (diagnosing behavioral problems). This skill is often trained by solving case simulations: given a description of a prototypical situation (e.g. the medical records of a patient), the student has to arrive at the right diagnosis (e.g. the causing disease). To improve their reasoning skills, students need individual feedback. However, providing feedback is time consuming and requires experts, which is why FAMULUS aims to develop an automated feedback system. This requires automatic detection of reasoning steps like generating *hypotheses* or drawing *conclusions*. To capture these, students write down their chain of reasoning and final diagnosis in the form of a short text. INCEpTION is used here to create a gold standard corpus where text spans containing reasoning activities are annotated. This corpus then serves as training data for the automated feedback system. It takes little time to model the annotation scheme for this task in INCEpTION. Additionally, FAMULUS leverages an external Python-based Bi-LSTM neural sequence tagger linked to INCEpTION to provide suggestions for spans to be annotated. Annotators report an improvement in annotation speed and ease while using this external recommender.

To summarize: in order to cope with the increasing demand of annotated corpora, we present INCEpTION. At this time it is the only tool that combines annotation support and knowledge management into an extensible and modular platform. INCEpTION is still under development, but its alpha version has already been released and employed in real-world use cases. We welcome early adopters and encourage feedback for a continued alignment of the platform with the needs of the community.

ACKNOWLEDGMENTS

This work was supported by the German Research Foundation under grant No. EC 503/1-1 and GU 798/21-1 (INCEpTION) and by the German Federal Ministry of Education and Research (BMBF) under the promotional reference 16DHL1040 (FAMULUS).

REFERENCES

[1] C. Schulz, M. Sailer, J. Kiesewetter, C. M. Meyer, I. Gurevych, F. Fischer, and M. R. Fischer. 2017. Fallsimulationen und automatisches adaptives Feedback mittels Künstlicher Intelligenz in digitalen Lernumgebungen. *e-teaching.org Themenspecial "Was macht Lernen mit digitalen Medien erfolgreich?"* (2017), 1–14.

[1] https://inception-project.github.io ; software is licensed under the Apache License 2.0

DESIRES 2018, August 2018, Bertinoro, Italy