Linking Text and Knowledge using the INCEpTION annotation platform

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

***Abstract*—In the Digital Humanities (DH), linking text collections to general or domain-specific knowledge bases or authority files is important to enable a contextualised analysis. Automatic named entity recognition and linking tools require training data or domain-specific methods. Interactive annotation tools do often not support the tasks of entity linking, fact-linking, cross-document reference resolution, etc. We aim to address this gap with the INCEpTION annotation platform, which not only provides these capabilities in the context of a generic annotation tool, but also combines them with machine learning methods to improve annotation efficiency.**

***Index Terms*—interactive entity linking, cross-document co-reference, generic tool annotation tool**

The INCEpTION[1] annotation platform is an integrated web-based environment for text annotation and knowledge management. Besides of offering the functionalities expected from a modern generic annotation tool, such as a versatile and yet intuitive user interface, a flexible configuration of the annotation schema, the ability to run multiple annotation projects concurrently and workflow-support with annotation and adjudication stages, it also integrates a generic knowledge management. This permits a wide range of use-cases including the obvious tasks of entity identification, entity disambiguation, entity linking, and consequently cross-document co-reference annotation. Moreover, due to the flexibility of INCEpTION, it is also possible to model advanced tasks such as fact linking or aspect-oriented entity linking where the role an entity takes in context is categorized.

A multi-perspective analysis and exploration of texts requires a tight integration of knowledge management and annotation support. From a text-oriented perspective, the user reads a text, identifies entity mentions and links them to the knowledge resource, potentially creating the resource in the process. From a knowledge-oriented perspective, the user searches for mentions of specific entities, e.g. to analyse their behaviour or to gather additional detail knowledge about them. When text exploration and analysis are performed for the purpose of gaining insight –as it is the case in many Digital Humanities use-cases– and not simply to create some gold standard dataset –as it is often the case in natural language processing– it is necessary to often switch between these perspectives. INCEpTION supports both perspectives and allows the user to assume one or the other as needed.

Linked entities, annotations, and full texts are indexed and can be searched using a powerful pattern-based query language. MTAS [1] is one of the few indexing systems for annotated corpora which supports frequent updates to the index as they are necessary in the context of an interactive annotation tool. It is based on the popular *Corpus Query Language* (CQL), providing a familiar syntax for many users.

***Recommenders* show annotation suggestions to render entity recognition, entity linking and other annotation tasks more efficient**. They can be either static or dynamic, actively learning from the user's input. For example, an entity recommender can learn to recognize domain-specific entities as they are identified by the user. An entity linking recommender can then suggest suitable knowledge base entries for linking, taking into account the context of the entity mention. Active learning strategies can be activated to guide the user to judging suggestions that are particularly helpful for training.

Flexible configuration and the use of standards make INCEpTION a generic and interoperable platform. Annotations are based on the UIMA standard and support custom annotation schemata. Knowledge bases can be created from scratch in the tool, be imported from RDF files, or be remotely accessed via SPARQL. Since the resources may follow different conventions (e.g. RDF Schema, OWL, Wikidata RDF mapping), a configurable mapping between resource and tool conventions is performed. Custom recommenders can be implemented as web services, allowing users to employ the recommender system for unforeseen and novel tasks.

The open and generic approach is presently unique to the INCEpTION platform, to the best of our knowledge. Current generic annotation tools often have no or only limited support for entity linking, e.g. supporting only specific resources such as DBPedia or Wikidata. Other tools focussing on cross-document coreference, may offer more flexibility in terms of the knowledge resources that can be used and may even make use of classifiers to render the entity linking more efficient, but fall short when it comes to the customizability of the annotation schema.

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

[1] https://inception-project.github.io

References

[1] M. Brouwer, H. Brugman, and M. Kemps-Snijders, "MTAS: A Solr/Lucene based Multi Tier Annotation Search solution," in *Selected papers from CAC 2016*, no. 136. Aix-en-Provence, France: Linköping University Electronic Press, Linköpings universitet, Oct. 2017, pp. 19–37.