CorefAnnotator

A New Annotation Tool for Entity References

Nils Reiter

Introduction

This poster presents a new tool for the annotation of co-referring entities in texts. While coreference resolution is an established task in computational linguistics (cf. Poesio et al. 2016), the operationalization of this task is tailored to texts prevalent in this field. This holds true for automatization approaches, but also annotation guidelines and annotation tools. The tool we present here has been developed in the context of an annotation project in which literary texts (drama and prose) are annotated with coreferences.

Resolving coreference is the task of determining whether two linguistic units (mentions) in the text refer to the same (text-external) entity. Linguistic units are typically proper names (Mary), pronouns (she), or noun phrases (the doctor). In the example in (1), the name "Mary" and the noun phrase "The doctor" both refer to the same entity, similarly to "the bike" and "it".

(1) Mary bought a bike. The doctor used it to go to work.

Entities play a major role in many text analysis projects in the DH community (e.g., social networks (Trilcke 2013) typically consist of entities and show their relations as expressed in the text). Entity references in a text also form the basis of attempts to extract traits for characters. Detecting all references to an entity is therefore an important preprocessing step for many DH applications.

Existing Annotation Tools

Most annotation tools that have been developed within the computational linguistics (CL) community have been developed for the use in specific projects, and make assumptions based on the project setup. WebAnno (WebAnno), for instance, works only well for texts of a certain length, as it becomes unresponsive if book-long texts are loaded. Since coreference chains are displayed within the text as lines, many coreference chains make it impossible to tell them apart. MMAX2 (Müller/Strube 2006) focuses on the annotation of binary relations between mentions, and requires input data to be preprocessed substantially. File formats that contain structuring information can not be used.

CorefAnnotator

The annotation tool we present here is a desktop application that is made available with an open source license (Apache 2.0)[1]. Coreference annotations are conceptualized as equivalence sets. All mentions that belong to one coreference chain form a set, and are treated equally. Annotating a mention into a chain adds it directly to the set. Each entity is represented by a color, and can optionally be named. All mentions that belong to the same entity are underlined with the same color in the text view, multiple annotations on the same span result in multiple underlines on different levels. As the human eye is only able to distinguish a small number of colors reliably, selected entities in the entity list will also be highlighted in the text view, thus allowing to identify mentions of specific entities quickly.

If the texts contain appropriate annotation (e.g., stage directions or headings), they can be used to control the formatting (bigger headings and italic stage directions, for instance). This makes reading and annotating more accessible, in particular for long texts.

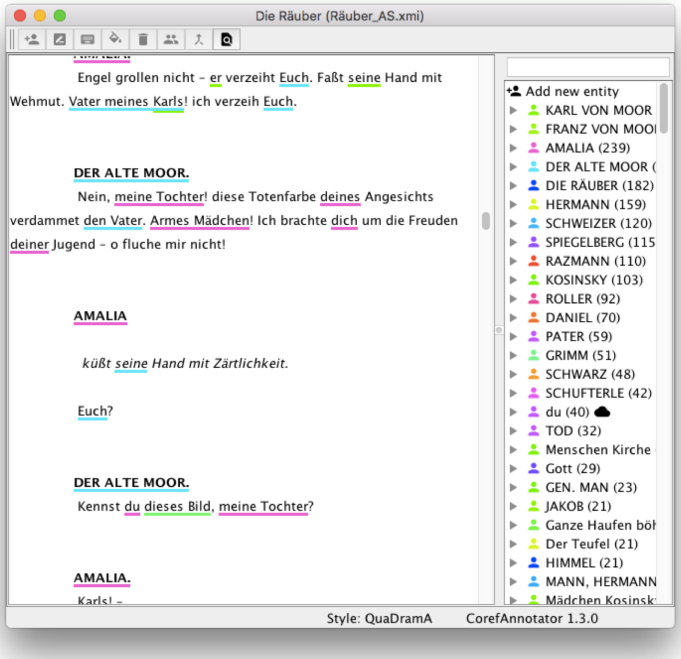

Figure 1: A Screenshot of the Annotation Tool

The tool makes no assumptions on related tasks. Arbitrary text spans can be annotated as mentions, including sub-token annotations (by default, partial token annotations are expanded to the full token). Internally, all annotations are represented as stand-off annotation using the UIMA frame-

[1]https://github.com/nilsreiter/CorefAnnotator. Each release version is archived and can be accessed with a DOI: https://doi.org/10.5281/zenodo.1228105.

work[2] for text and annotation representation. This allows flexible import and export in a variety of formats: **CoNLL 2012** is the most commonly used format in the CL community. Annotations can be exported into CoNLL directly. For this an automatic sentence splitting and tokenization is done. Texts can be imported from **TEI/XML**. Texts that have been imported from TEI/XML can also be exported again as the same TEI/XML, without loosing any of the TEI-encoded information.

The tool is built to support keyboard based annotation to increase annotation speed. Text spans can be selected with the keyboard, and the appropriate entities can be searched for. In addition, selected text spans can be dragged onto the entity. Fast, large scale annotation can be performed via the search function. It supports regular expressions and all or some found spans can be annotated as a new or existing entity with a single click or press. Annotations created by different annotators can be loaded into a compare view which makes annotation mistakes easily identifiable. Similarly, if parts of the texts have been annotated by different annotators, the resulting files can be merged again.

Finally, the tool includes a context-sensitive candidate generation: After selecting a span, the tool suggests likely entities that the span might refer to. This candidate list is based on contextual information as well as the existing entity list.

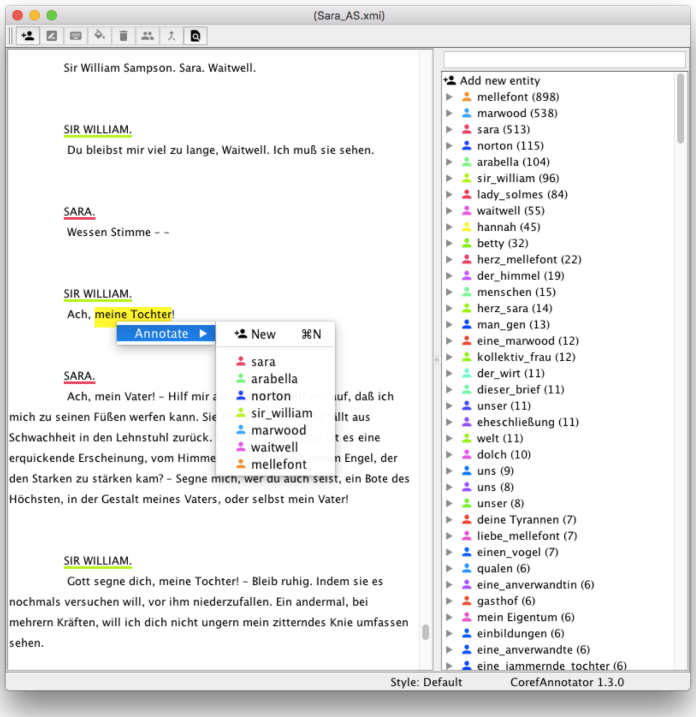

Figure 2: The candidate generation function

[2]http://uima.apache.org

References

Müller, Christoph/Strube, Michael: Multi-level annotation of linguistic data with MMAX2. In: *Corpus technology and language pedagogy: New resources, new tools, new methods*. Eds. von Sabine Braun/Kurt Kohn/Joybrato Mukherjee. Peter Lang, 2006, 197–214.

Poesio, Massimo/Stuckardt, Roland/Versley, Yannick (Eds.): *Anaphora resolution*. Springer, 2016.

Trilcke, Peer: Social Network Analysis (SNA) als Methode einer textempirischen Literaturwissenschaft. In: *Empirie in der literaturwissenschaft*. Eds. von Philip Ajouri/Katja Mellmann/Christoph Rauem. 2013, 201–247.

Yimam, Seid Muhie/Gurevych, Iryna/Eckart de Castilho, Richard/Biemann, Chris: WebAnno: A flexible, web-based and visually supported system for distributed annotations. In: *Proceedings of the 51st annual meeting of the association for computational linguistics: System demonstrations*. 2013, 1–6.