What's in this Collection Dataset?
Semantic Annotation with GATE

Felicitas Löffler†, Birgitta König-Ries†
† Friedrich Schiller University Jena, Jena, Germany

Abstract

Semantic annotations of datasets are very useful to support quality assurance, discovery, interpretability, linking and integration of datasets. However, providing such annotations manually is often a time-consuming task. If the process is to be at least partially automated and still provide good semantic annotations, precise information extraction is needed. The recognition of entity names (e.g., person, organization, location) from textual resources is the first step before linking the identified term or phrase to other semantic resources such as concepts in ontologies. A multitude of tools and techniques have been developed for information extraction. One of the big players is the text mining framework GATE (Cunningham et al. 2013) that supports annotation rules, semantic techniques and machine learning approaches. We will run GATE’s default ANNIE pipeline on collection datasets to automatically detect persons, locations and time. We will also present extensions to extract organisms (Naderi et al. 2011), environmental terms, data parameters and biological processes and how to link them to ontologies and LOD resources, e.g., DBPedia (Sateli and Witte 2015). We would like to discuss the results with the conference participants and welcome comments and feedbacks on the current solution. The audience is also welcome to provide their own datasets in preparation for this session.
Keywords

semantic annotation, collection data, GATE

Presenting author

Felicitas Löfler

Presented at

Biodiversity_Next 2019

References