OntoHuman: User Interface for Ontology-based Information Extraction from Technical Documents with Human-in-the-loop interaction

Kobkaew Opasjumruskit, NFDI4Ing Conference 2022

In this talk, we present DSAT (Document Semantic Annotation Tool), a tool to automatically extract information from technical documents based on ontologies and natural language processing techniques, within the context of the OntoHuman project[1].
The OntoHuman project aimed to enrich ontologies, which contain semantic information describing objects or concepts, with information extracted from documents. The central component of the OntoHuman Project is DSAT, which was originally designed for assisting users to annotate key-value-unit tuples on technical documents.

Besides the user interface, there are other modules used in OntoHuman: an ontology enrichment module (ConTrOn - Continuously Trained Ontology), a DSAT database (DSAT DB) for storing annotations and custom ontologies, and an information extraction module (PLIX[2]). These components are integrated into OntoHuman to achieve the automatic information extraction.

Prior to OntoHuman, the ontologies used by DSAT for the automatic extraction were fixed and limited to one specific domain, i.e. spacecraft engineering. To update and customize an ontology manually is tedious and requires additional efforts to use ontology modelling tools. Therefore, a semi-automatic process to enrich ontologies can assist domain experts, who are not necessarily ontology experts, to map their knowledge into ontologies.

To enable the customization of ontologies, we improved DSAT and ConTrOn in the OntoHuman project. We also pursue the Human-in-the-Loop (HiL) approach, which requires humans to verify the results of an automatic process by providing feedback to the system. We combined the HiL component to generalize the automatic information extraction process. In contrast to the prototypical solution, we now can apply and customize ontologies to extract data from documents of other domains. Feedback from users can now be collected via a web-based user interface and used for updating ontologies further.

---

[1] The project was finished in June 2022, and the source code is publicly available on Zenodo (10.5281/zenodo.6783007)
[2] PLIX (Information Extraction module) version 1.0, license Apache-2.0, authors: Sarah Böning, Christian Kiesewetter

The following proposed features were implemented: correction of automatically extracted data, resolution of word ambiguities, adding new annotations, and export function for annotations. Additionally, we simplified the UI according to feedback from workshops participants from the NDFI4Ing community. We also conducted a user survey and received rather good rating for the tool (DSAT). Regarding the user experience, the tool is considered to be easy to use (6 points out of 7), supportive (5.5/7), efficient (6/7) and novel (5/7). The workshop's participants rated the domain of usage of DSAT to generic purpose (rated 3.5 points out of 5), somewhat relevant to their colleagues' work (3/5), and not very relevant to their own work (2/5). However, since the participants of the workshops were limited to 9 and 6 persons, we hope to collect further feedback and attract more users from various fields of work during this conference.

Since the automatic annotation of documents depends largely on the used ontologies, to fully use the tools for other domains, users should know where to find relevant ontologies. An ontology search API could be used to assist the users to find the right ontologies in the future. Furthermore, the suggested topics we collected from the workshops, such as semantic disambiguation, multi-language support, and graph value extraction are rather complicated topics. Therefore, we decided to research these topics beyond the project period. They are currently studied and could be integrated into DSAT in the future.