MDS Writer
Annotation tool for creating high-quality multi-document summarization corpora

Christian M. Meyer, Darina Benikova, Margot Mieskes, and Iryna Gurevych
Research Training Group AIPHES, Technische Universität Darmstadt, Germany

Motivation
- Amount of textual information (e.g., in the web) becomes intractable
  ➢ We need automatic summarization!
- Single-document summarization cannot reduce the number of documents
  ➢ We need multi-document summarization!
- Corpora are crucial for training and evaluating automatic methods
- Existing corpora limited in terms of size, domains, genres, languages
  ➢ We need summarization corpora!

- Summarization is one of the most challenging NLP tasks, as it requires solving multiple subtasks, incl. content selection, redundancy removal, coherent writing
- Existing corpora focus on a single subtask (e.g., the final summary text), which prevents us from evaluating intermediate steps
- Currently no freely available tools for modeling a complex multi-document summarization setup and storing intermediate results and system–user interactions
  ➢ We need tools for complex annotation setups!

Summary
MDS Writer is an open-licensed software for designing complex annotation setups with multiple steps. It is particularly useful for the creation of summarization corpora.

Highlights:
- Divide complex annotation setups into multiple steps
- Support cross-document tasks such as multi-document summarization
- Link annotation tool and guidelines to support human annotators in creating high-quality corpora
- Separately store results of intermediate steps and system–user interactions
- Improve evaluation of automatic summarization methods by assessing intermediate results
- First step towards our vision for next-generation summarization systems that learn the human summarization process rather than only replicating its result
- Easy to extend to other summarization setups and other cross-document tasks
- Available as open-source software from GitHub under the Apache License (ASL)
  ➢ https://github.com/UKPLab/mdswriter

Prototypical Annotation Workflow

(Step 1: Nugget identification)

(Step 4: Co-reference resolution)

(Step 2: Redundancy detection)

Extensibility
- MDS Writer is flexible to deviate from our prototypical annotation workflow:
  ➢ Define the steps of your own workflow, create and test annotation guidelines
  ➢ Add, modify, or reorder existing steps
  ➢ Extend 18N if necessary (currently: en/de)
- Enables a wide range of complex annotation setups such as:
  ➢ Summarization (single- and multi-document, structured and opinionated summaries,...)
  ➢ Information extraction (separate steps for entity, event, relation identification)
  ➢ Text compression (sentence fusion, subclause removal, co-references, lexical substitution,...)
  ➢ Cross-document discourse structure annotation

Technology
- Java/JSP and JavaScript
- Multi-user, multi-step, multi-doc. support
- Individual user actions sent to server application in real-time based on WebSockets connection
- Results stored in SQL database
- Simple text-based protocol allows for easy and fast extensions

Source code, documentation, annotation guidelines, video tutorial

https://github.com/UKPLab/mdswriter

https://www.aiphes.tu-darmstadt.de

This work has been supported by the DFG-funded research group “Adaptive Preparation of Information from Heterogeneous Sources” (AIPHES, GRK 1864/1) and the Volkswagen Foundation as part of the Lothar-Liebknecht-Program under grant No. 182806.