Visualization of Text Annotations for Corpus Development

Huan He *, Sunyang Fu, Liwei Wang, Andrew Wen, Sijia Liu, Sungrim Moon, Kurt Miller, and Hongfang Liu †

Department of AI and Informatics Research, Mayo Clinic, Rochester, MN, USA

* He.Huan@mayo.edu † Liu.Hongfang@mayo.edu

Background

A gold standard annotated corpus is usually indispensable for building high-quality language models for clinical natural language processing.

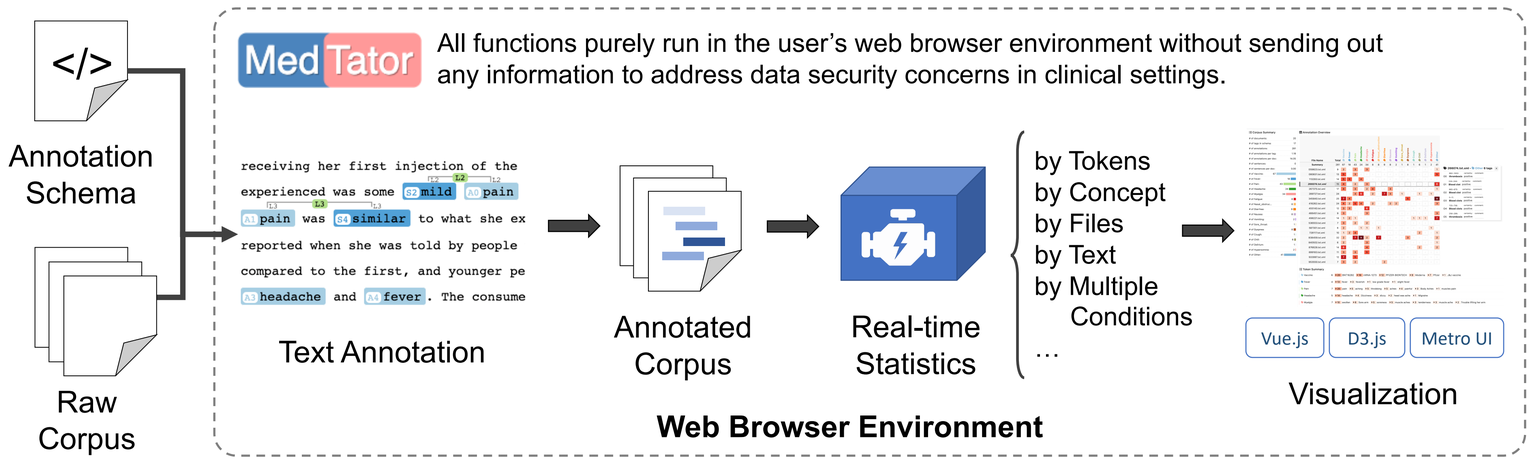
Existing text annotation tools can provide powerful features to cover various needs for corpus development, but few tools provide real-time visualization to support the needs of annotation analysis during the annotation process.

To address this need, we developed a corpus visualization module in ***MedTator***, a **serverless** annotation tool.

MedTator doesn't require any runtime installation! You can check our online demo at:
https://ohnlp.github.io/MedTator/

Source codes are available at:
https://github.com/OHNLP/MedTator

System Overview



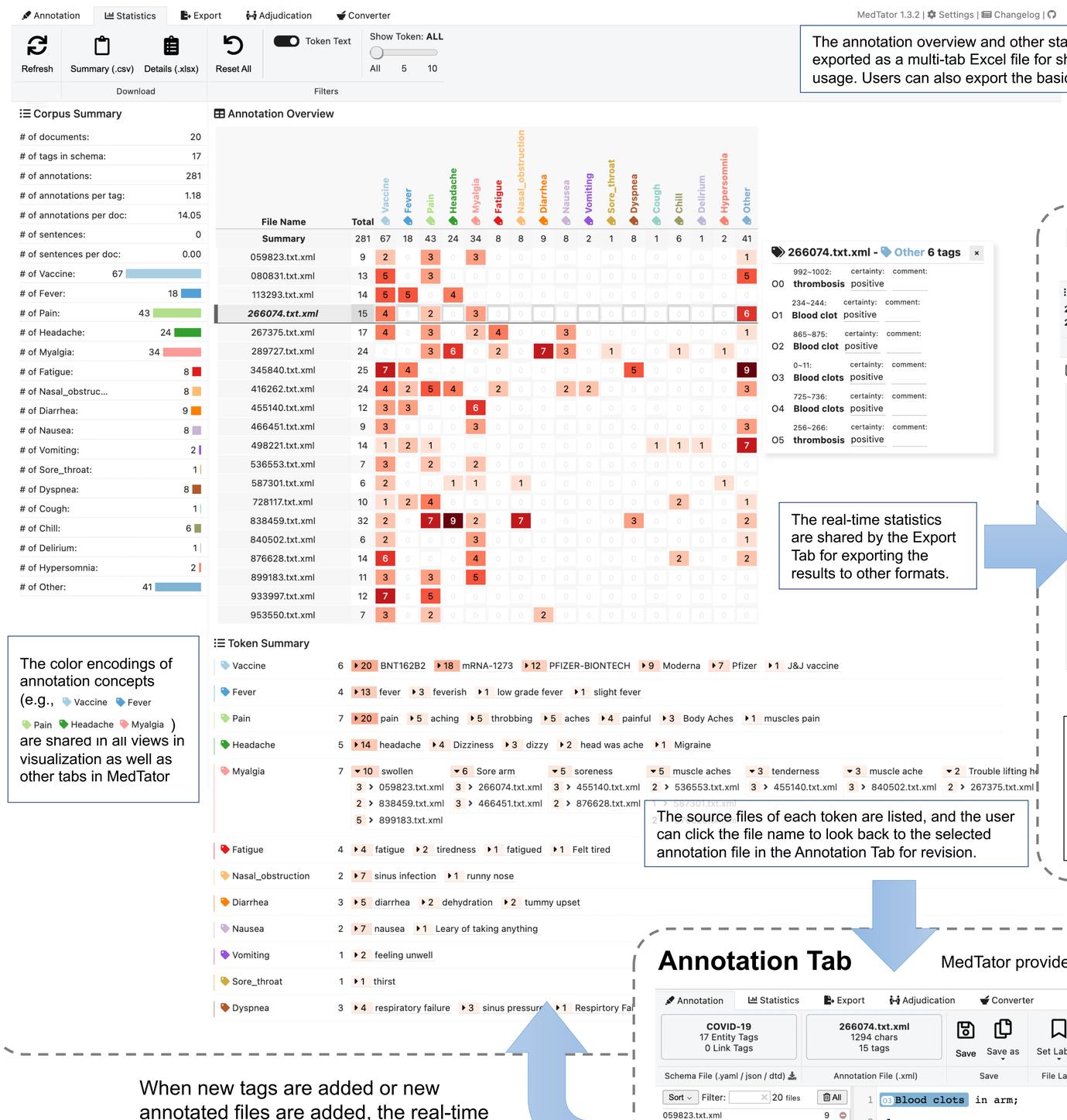
Users can annotate the corpus by using the Annotation Tab in MedTator, and the text annotations are saved in memory.

The annotations are analyzed in real-time and shown in the Statistics Tab.

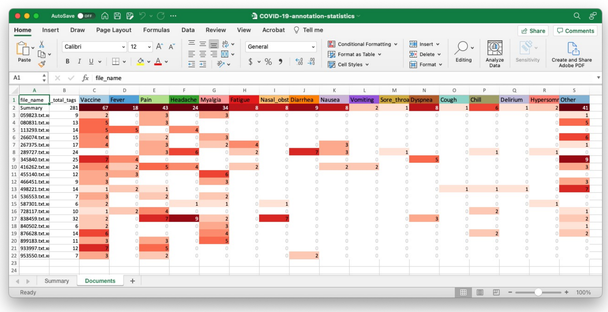
Visualization and Interactivity Designs

Statistics Tab

MedTator provides a separate Statistics Tab to visualize the annotated corpus



The annotation overview and other statistical results can be exported as a multi-tab Excel file for sharing and further usage. Users can also export the basic statistics to a CSV file.



The real-time statistics are shared by the Export Tab for exporting the results to other formats.

The color encodings of annotation concepts (e.g., Vaccine Fever Pain Headache Myalgia) are shared in all views in visualization as well as other tabs in MedTator

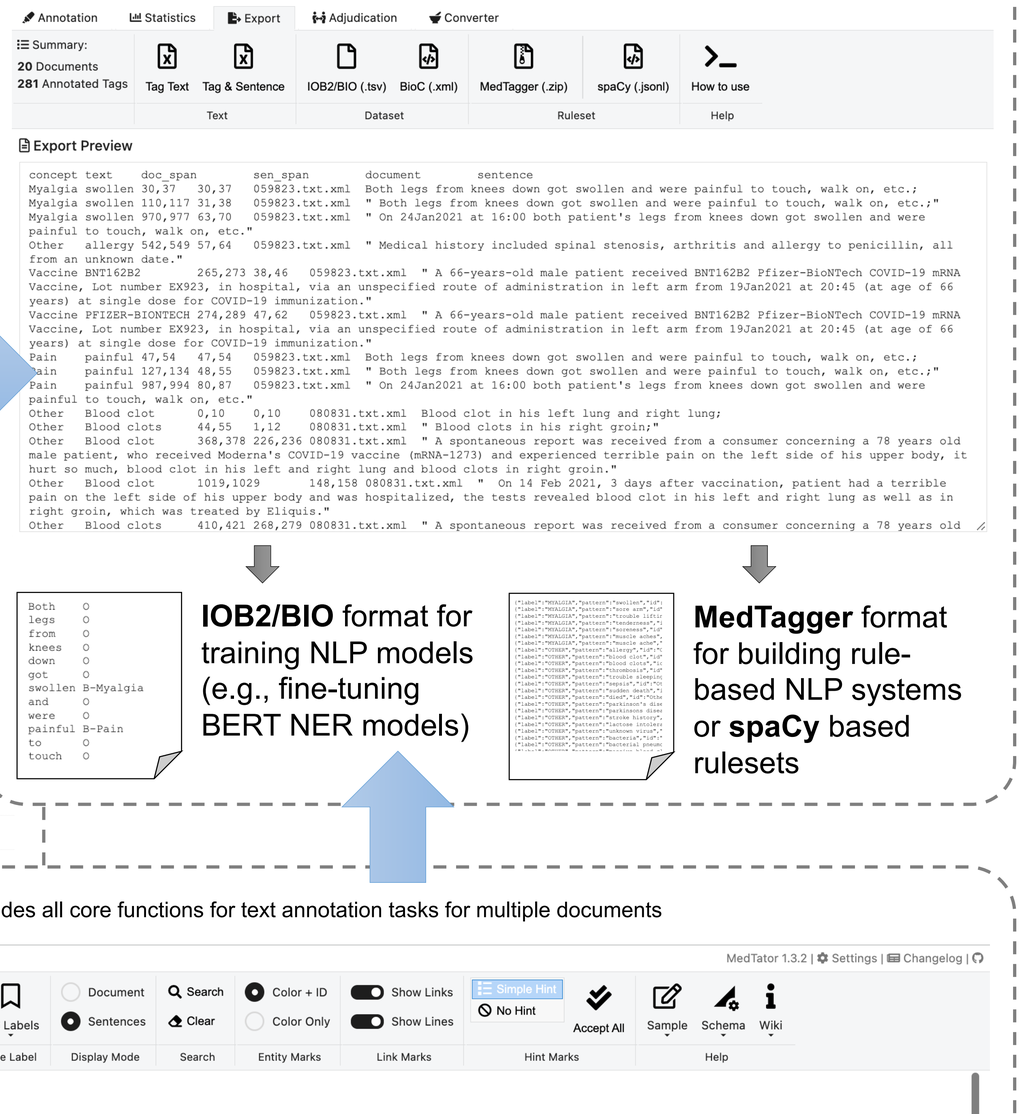
The source files of each token are listed, and the user can click the file name to look back to the selected annotation file in the Annotation Tab for revision.

When new tags are added or new annotated files are added, the real-time statistics will be updated in the backend.

Users can switch to the Statistics Tab to check the annotation progress or analyze the imported corpus at any time.

Export Tab

the annotations and statistics can be exported to multiple formats

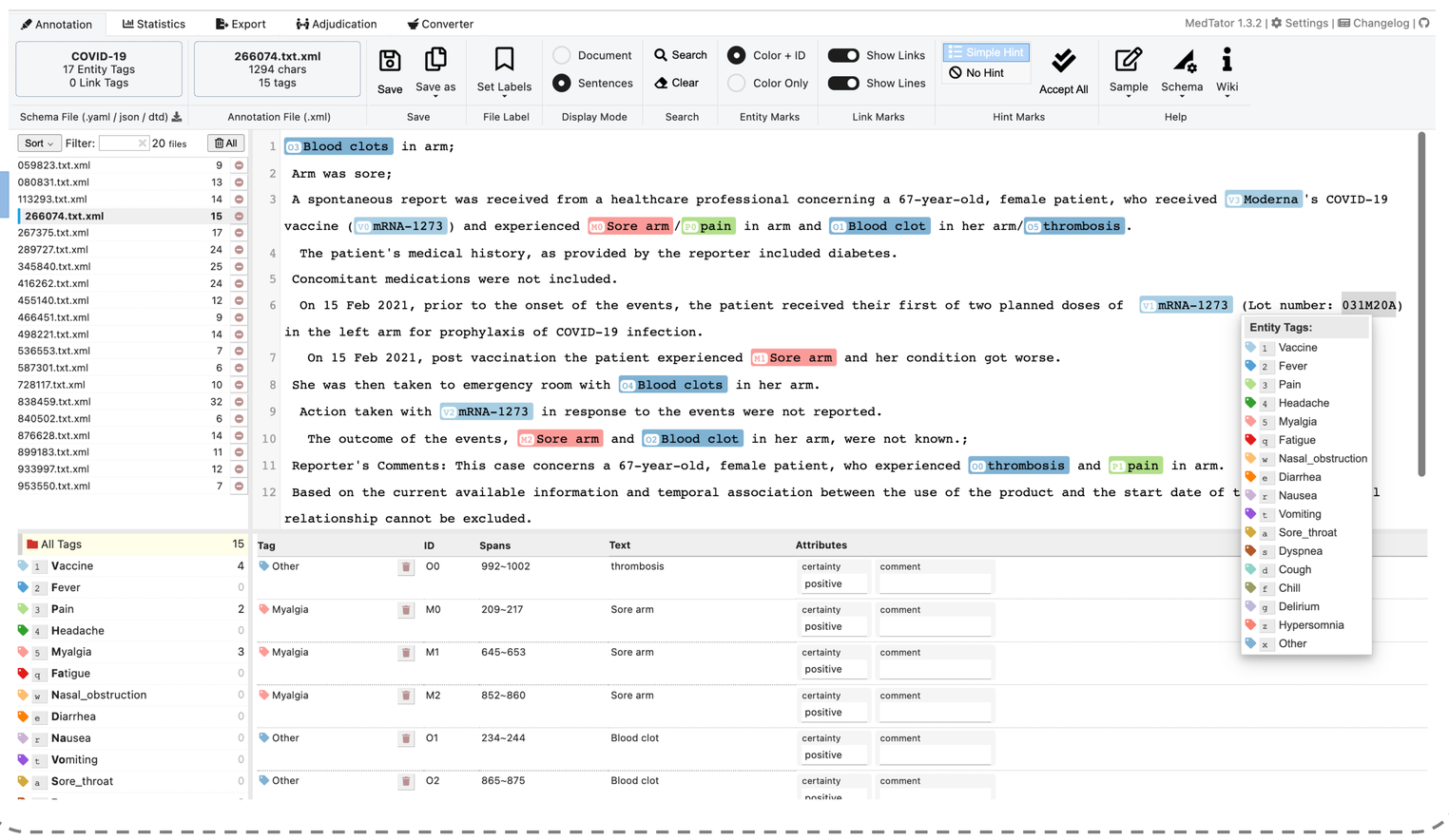


IOB2/BIO format for training NLP models (e.g., fine-tuning BERT NER models)

MedTagger format for building rule-based NLP systems or **spaCy** based rulesets

Annotation Tab

MedTator provides all core functions for text annotation tasks for multiple documents



Future Work

We are still working on improving the performance and usability of MedTator and its corpus visualization module, but we have been using MedTator in our internal NLP projects for corpus development, adjudication, and building rule-based NLP systems.

We welcome you to use MedTator in your projects and leave any comments on the MedTator's repo issues:
https://github.com/OHNLP/MedTator/issues

This work has been supported by the National Center for Advancing Translational Science of the National Institutes of Health under award number U01TR002062.