Visualization of Text Annotations for Corpus Development

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

Department of Artificial Intelligence and Informatics, Mayo Clinic, Rochester, MN, USA

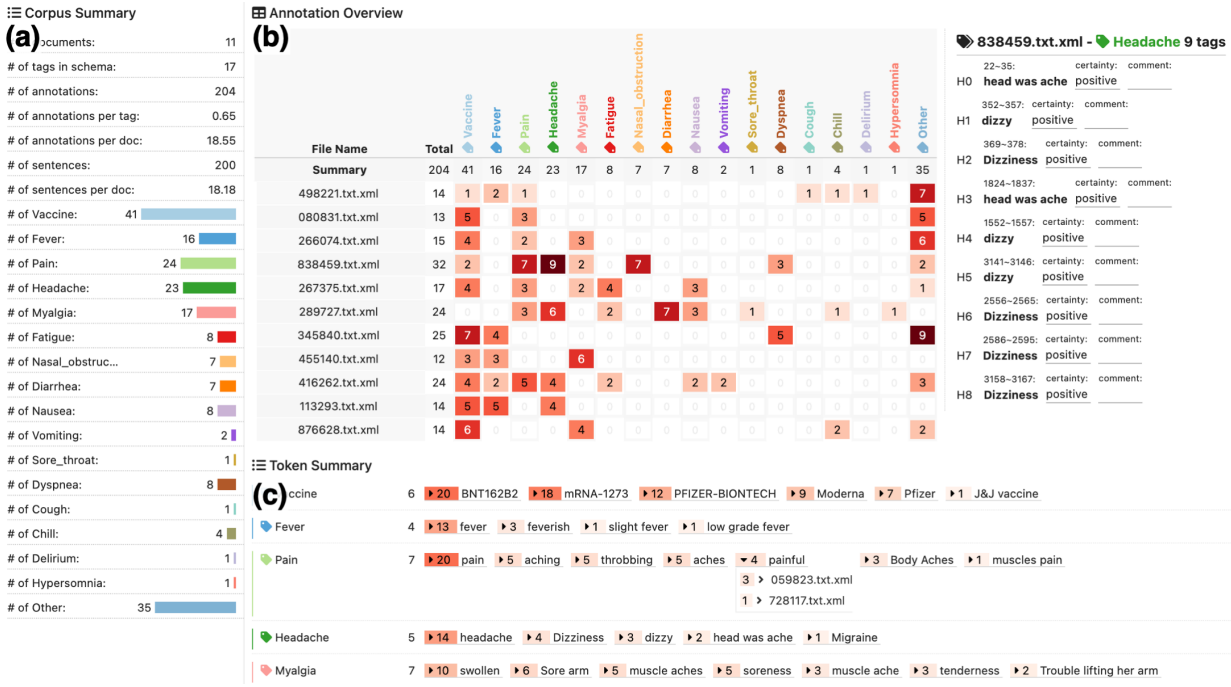

Figure 1: A snapshot of the corpus visualization module in MedTator for corpus development showing: (a) the Corpus Summary view listing the basic statistics of a corpus, (b) the Annotation Overview showing the distribution of annotated tags from the document and tag perspectives, and (c) the Token Summary view listing the detailed annotated tokens and corresponding documents.

Abstract

A gold standard annotated corpus is usually indispensable for building high-quality language models for clinical natural language processing (NLP) tasks. Existing text annotation tools can provide powerful features to cover various needs of text annotation during 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.

Keywords: Corpus visualization, text annotation.

1 Introduction

The development of a gold standard annotated corpus is an integral part of natural language processing (NLP) algorithm development and evaluation. Multiple annotation tools have been developed with many powerful features to support the annotation process [1]. We developed MedTator [2], a serverless web-based annotation tool aiming to provide an intuitive and interactive user interface that focuses on the core steps related to corpus development.

However, during the development of MedTator, we found that our users (e.g., annotators, adjudicators, and researchers) have strong and diverse needs for analyzing the annotated data in addition to text annotation. For example, annotators need to track their annotation progress and ensure that no documents are missed, adjudicators need to check the annotated tokens to ensure the corpus quality, and researchers need to analyze the annotation patterns and improve the schema design. These needs are usually addressed by some customized scripts or manual reviews, requiring extra effort in the data processing and analysis not supported by the annotation tool. The extra effort not only reduces the efficiency of corpus development but can also introduce unexpected errors, ultimately affecting the corpus quality.

To address these needs, we propose using web-based data visualization techniques to show real-time statistics of the annotations in MedTator during the annotation process. The major contribution of this work is a corpus visualization module in MedTator, which has been open-sourced and public accessible without any requirements for runtime installation.

2 Related Work

Text visualization and visual text analytics have become a growing and increasingly important field, and actually, there have been many text visualization techniques proposed to address a variety of text analytic tasks, such as topic analysis and sentiment analysis. However, a few previous studies [3], [4] focused on visualizing the annotated tokens and narratives of a single document or general statistics. Our study focuses on corpus-level visualization by showing the distribution of all the annotated tokens from different aspects to facilitate corpus development.

* He.Huan@mayo.edu

† Liu.Hongfang@mayo.edu

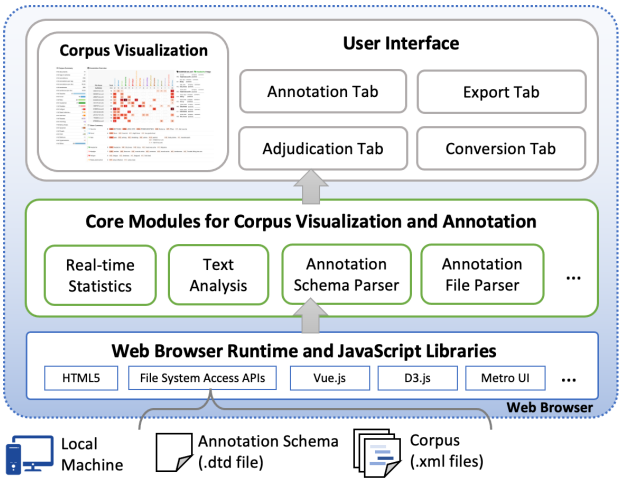

Figure 2: The architecture of our proposed corpus visualization tool, which shares the modules in MedTator for interoperability.

3 System Overview

To reduce user barriers and enhance the interoperability between the corpus visualization tool and the text annotation tool, we implemented our visualization tool as an integrated component in MedTator. As shown in Figure 2, the corpus visualization tool can leverage the core modules for text annotation tasks based on the web browser runtime and other libraries. As a result, our tool can access any corpus information directly from the annotation tool and traceback to the original documents and annotations. Moreover, the corpus visualization tool can provide real-time statistics to users while annotating a corpus.

We have released the source code on GitHub under Apache 2.0 license: https://github.com/OHNLP/MedTator. The user manual and other materials can be found in the same GitHub repository. Moreover, MedTator is freely accessible without any runtime installation at https://ohnlp.github.io/MedTator.

Due to the security concerns and patient privacy of our clinical data, we can only share the annotations of the public datasets from the Vaccine Adverse Event Reporting System (VAERS) as sample datasets in our online demo. Users can load the sample datasets from the MedTator menu directly.

4 Task Analysis and Visual Designs

We worked with the domain experts from our clinic and the users of MedTator on several annotation tasks of clinical documents, such as family history extraction and adverse event of vaccine identification. Based on their suggestion and feedback, we defined the visualization tasks as follows:

- **T1**: Visualize the basic statistics of a corpus, such as the number of annotated tokens at the document level and tag level.
- **T2**: Show the detailed statistics of annotated tokens for each tag so that users can review annotations to fix mistakes.

As shown in Figure 1, our proposed corpus visualization tool features three views: the Corpus Summary view, the Annotation Overview, and the Token Summary view.

The Corpus Summary view (Fig. 1a) is a table that displays the basic statistics of a corpus, such as the total number of annotations in the corpus and the number of annotations for each tag (**T1**). In addition, the number of annotations for each tag is displayed as a horizontal bar chart that is color-encoded by the corresponding tag. The color encoding for each tag is decided when the annotation schema is imported to MedTator. It is also used in other views and the other user interfaces in MedTator. Users can export the basic statistics and detailed information in a spreadsheet for downstream analysis in other software.

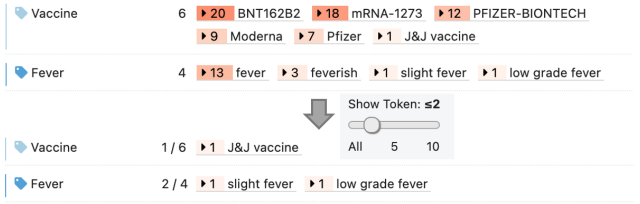

Figure 3: Filtered tokens for by the Token Summary view.

The Annotation Overview (Fig. 1b) is a heatmap that displays the distribution of the annotated tags of a corpus from different aspects (T1). The x-axis of the heatmap is the tags defined in the annotation schema, while the y-axis is the filenames of documents sorted in the import order. The number of annotated tokens of each tag in each document is colored-encoded and displayed as a cell in this chart. Users can click each cell to browse the corresponding tokens. In addition, users can also click the filename or cell to open the related document and tokens in the annotation tab for checking the annotation context.

The Token Summary view (Fig. 1c) is a dynamic table that displays the annotated tokens of each tag (**T2**). The tokens of each tag are sorted and displayed one by one in each row. The annotation count is displayed before each token and color-coded by the same rule in the Annotation Overview. Users can specify a threshold to select those less annotated tokens to identify rare cases or mistakes. For example, by setting the threshold to two, the tokens annotated more than two times are hidden, and only those less annotated tokens are left for examination (Fig. 3). Moreover, the documents where the token is annotated are displayed when clicking the token, as well as the annotation count in the document. Users can click the document file name to open the document in the annotation tab for revision.

5 Discussion and Conclusion

Iterative refinement is necessary for delivering user-friendly tools. When domain experts and annotators use MedTator, we continuously evaluate the functionality and the effectiveness of visual designs. We have not carried out formal user tests since it is still under development. Our users have appreciated the interactive user interface and can effectively explore the statistical results.

In addition to the positive feedback, they commented that the current tool has limited abilities in adjudicating corpus and error analysis. For example, some common annotation mistakes should be identified automatically and summarized for fixing, such as "missing annotation" and "unnecessary white space". They also commented that the corpus visualization could be deeply integrated with MedTator to empower other annotation functions, such as comparing the token distribution of two corpora for improving the adjudication and evaluating corpora quality. We plan to incorporate the suggestions and enhance the visual design of our corpus visualization tool in the future.

References

[1] M. Neves and J. Ševa, "An extensive review of tools for manual annotation of documents," *Briefings in Bioinformatics*, vol. 22, no. 1, pp. 146–163, Jan. 2021, doi: 10.1093/bib/bbz130.

[2] H. He, S. Fu, L. Wang, S. Liu, A. Wen, and H. Liu, "MedTator: a serverless annotation tool for corpus development," *Bioinformatics*, p. btab880, Jan. 2022, doi: 10.1093/bioinformatics/btab880.

[3] E. Amorim *et al.*, "Brat2viz: a tool and pipeline for visualizing narratives from annotated texts," 2021.

[4] J. S. Grosman, P. H. T. Furtado, A. M. B. Rodrigues, G. G. Schardong, S. D. J. Barbosa, and H. C. V. Lopes, "Eras: Improving the quality control in the annotation process for Natural Language Processing tasks," *Information Systems*, vol. 93, p. 101553, Nov. 2020, doi: 10.1016/j.is.2020.101553.