# Visual Text Analysis for NLP System Evaluation and 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

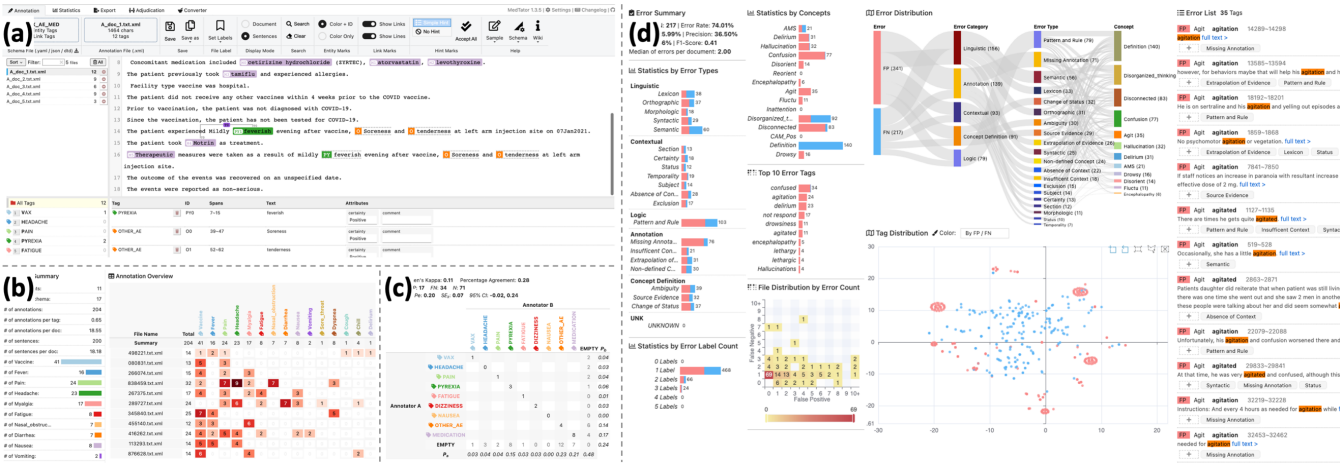


Figure 1: Screenshots of the data visualization for text analysis in MedTator, including (a) text visualization of the entity annotations for building gold standard corpus, (b) statistical result visualization for analyzing corpus, (c) visualization of adjudicating annotation results of two annotators for evaluating inter-annotator agreement, and (d) visual error analysis of the NLP system output for improving NLP system performance.

## Abstract

With the rapid development of big data and deep learning techniques in recent years, natural language processing (NLP) systems have been widely used in the healthcare domain. Due to the low error tolerance for clinical use cases, such as clinical decision support and precision medicine, performance evaluation is a crucial task in developing and refining an NLP system. However, evaluating an NLP system for a specific clinical use case is challenging due to the complexity of the gold standard corpus creation and the evaluation process. In response, we proposed using web-based data visualization techniques to promote the whole process and developed several visual analysis modules based on a text annotation tool.

## 1 Introduction

Natural language processing (NLP) has been widely applied in healthcare-related domains for clinical practice and research field [1]. Due to the low error tolerance for clinical use cases, such as clinical decision support and precision medicine, performance evaluation is a crucial task in developing and refining an NLP system in clinical settings.

However, evaluating an NLP system for a specific clinical use case is challenging. First, a high-quality gold standard corpus (GSC) is needed for evaluation. But due to the complexity and diversity of the clinical practices, the GSC creation requires manual annotation from a raw corpus with iterative evaluations, which is a labor-intensive process. Second, there are several tasks in the GSC creation, including corpus annotation, annotation analysis, adjudication, and error analysis. Each task may involve different data formats and analysis needs, requiring the user to have various skills and tools. Thus, these tasks are not straightforward processes to interpret the results and improve the performance of the NLP system accordingly.

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

To address these challenges in the GSC creation and NLP system evaluation, we proposed using web-based data visualization techniques to promote the whole process. Moreover, we designed and developed several visual analysis modules based on a serverless text annotation tool, MedTator [2], to support evaluation tasks, including corpus annotation, corpus analysis, adjudication, and error analysis. The source code, online demo, and sample datasets are available on GitHub under Apache 2.0 license: https://ohnlp.github.io/MedTator.

## 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 focused on visualizing the annotated tokens [3] and narratives of a single document or general statistics [4]. Our study focuses on corpus-level visualization from different aspects to facilitate GSC development and NLP system evaluation.

## 3 System Design

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.

## 4 Visual Designs

We worked with the domain experts from our clinic and the users of MedTator on several annotation tasks of clinical documents to collect the visualization tasks and design the user interface for each task in creating the GSC.

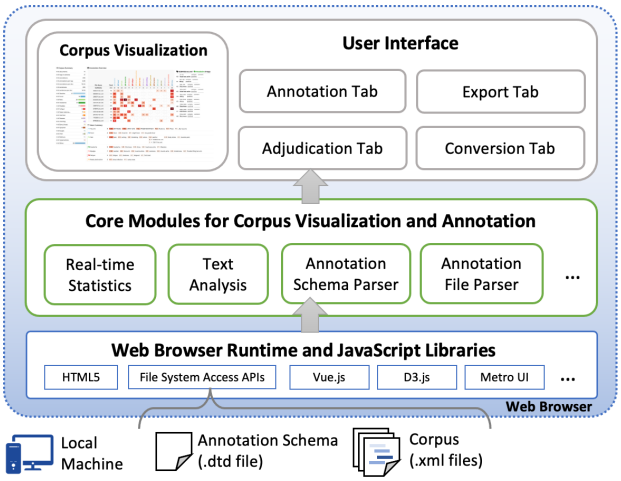


Figure 2: The architecture of our proposed corpus visualization tool, which shares the modules in MedTator for interoperability.

### 4.1 Annotation Visualization

To support corpus annotation, we leveraged web frontend techniques to help annotators to identify the annotated entities and relations in the documents. As shown in Fig. 1a, the annotated word tokens are highlighted in the document with customized color encodings based on the concepts so that users can identify the tokens and related concepts.

At present, MedTator supports visualizing two types of text annotation in this interface: 1) span-based entity annotation that shows as color rectangles, and 2) entity-based relation annotation that shows as labeled lines. To reduce the repetitive workload, MedTator can optionally visualize hints on potential tokens of interest based on real-time statistics of the annotated tags. It takes a simple one-click on the hint box to automatically add a new tag or accept all hints as to new tags. In addition, the real-time statistical results will also be visualized as number hints for tracking annotation progress.

### 4.2 Visual Exploration of Annotated Tags

To support the visual exploration of the annotated tags for annotation analysis, we implemented a module to generate statistical results of an annotated corpus, such as the total number of tokens and the distribution of tokens among concepts and files. As shown in Fig. 1b, we designed a bar chart and a heatmap to show these results from different aspects. Those visualized results can help users track the annotation progress and understand how the tokens are distributed in the annotated corpus.

The bar charts display the basic statistics of a corpus, such as the total number of annotations in the corpus and the number of annotations for each tag. 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.

The heatmap displays the distribution of the annotated tags of a corpus from different aspects. 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 to check the frequency of each token. 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.

### 4.3 Visual Adjudication

To support the annotation adjudication, we implemented commonly used inter-annotator agreement measurements, including F1-score and Cohen's Kappa coefficient. As shown in Fig. 1c, we designed an interactive table to show the confusion matrix of the results. In addition, users can also check multi-level F1 scores and the detailed tokens interactively in a separate tab.

The calculation results can be downloaded as a multi-sheet Excel file for further analysis. In addition, users could accept or reject each annotator's annotation to generate the GSC for further downstream NLP tasks, such as designing rule-based systems, training named entity recognition models, and developing relation extraction models.

### 4.4 Visual Error Analysis

Once the GSC is built, users can use it to compare the output of an NLP system and conduct an error analysis to evaluate the NLP system. Therefore, as shown in Fig. 1d, we designed multiple coordinated views to show the errors, including bar charts, heatmap, Sankey diagrams, scatter plots, and token lists, to help users gain an in-depth understanding of the errors for improving the performance of the NLP system.

Each view of this interface visualizes the error tags from different aspects to show conditional distributions. For example, the bar width represents the number of tags with a specific error label. The user can find out what error types the NLP system mainly has. Moreover, the text embeddings of all error tags are plotted as scatter based on t-distributed stochastic neighbor embedding (t-SNE) to help explore the semantic patterns. Users can click the bars or dots to check the detailed tag information and context in the tag list.

We provide a default error schema as a reference in MedTator, which includes common error categories such as linguistic, contextual, logic, annotation, and concept definition. Users can modify them based on their own context and project needs. To further analyze the error types, we designed web services based on NLP techniques such as transformer-based sentence embedding and classification to facilitate error identification. In addition, users can also manage the error labels in the tag list manually. The user-created error labels and the detailed tag information can be exported to an Excel file or a JSON file for other downstream tasks.

## 5 Discussion and Future Work

The NLP system can be evaluated from many perspectives, which can hardly be covered by a single tool or technique. Our tool currently only focuses on supporting the GSC creation and error analysis of the information extraction task.

We plan to improve the visualization and interactivity designs to enhance the text analysis abilities for the error analysis with some case studies on our real NLP systems to validate the effectiveness and usability of our methods.

## References

[1] A. Wen *et al.*, "Desiderata for delivering NLP to accelerate healthcare AI advancement and a Mayo Clinic NLP-as-a-service implementation," *npj Digit. Med.*, vol. 2, no. 1, Art. no. 1, Dec. 2019, doi: 10.1038/s41746-019-0208-8.

[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," in *Proceedings of Text2Story-Fourth Workshop on Narrative Extraction From Texts held in conjunction with the 43rd European Conference on Information Retrieval (ECIR 2021)*, 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.