Visual Exploration of the Results and Errors of Information Extraction

Huan He, PhD, Sunyang Fu, PhD, Andrew Wen, MS, Liwei Wang, MD, PhD, Sungrim Moon, PhD, Kurt Miller, MS, and Hongfang Liu PhD†

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

† Liu.Hongfang@mayo.edu

Background

Information extraction (IE) is a critical step in facilitating the use of healthcare data in clinical decision support and translational research by leveraging natural language processing (NLP) techniques. To evaluate the results of IE systems or algorithms, it is essential to examine the extracted entities and relations with the contextual information. Moreover, as IE errors are not unavoidable, conducting an error analysis is usually required to gain more insights into the reasons for errors.

However, it's challenging to analyze the IE errors in its raw format due to the complexity. To address this challenge, we proposed a visual exploration tool to visualize and analyze the IE results and errors based on data visualization techniques and our serverless tool, MedTator.

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

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

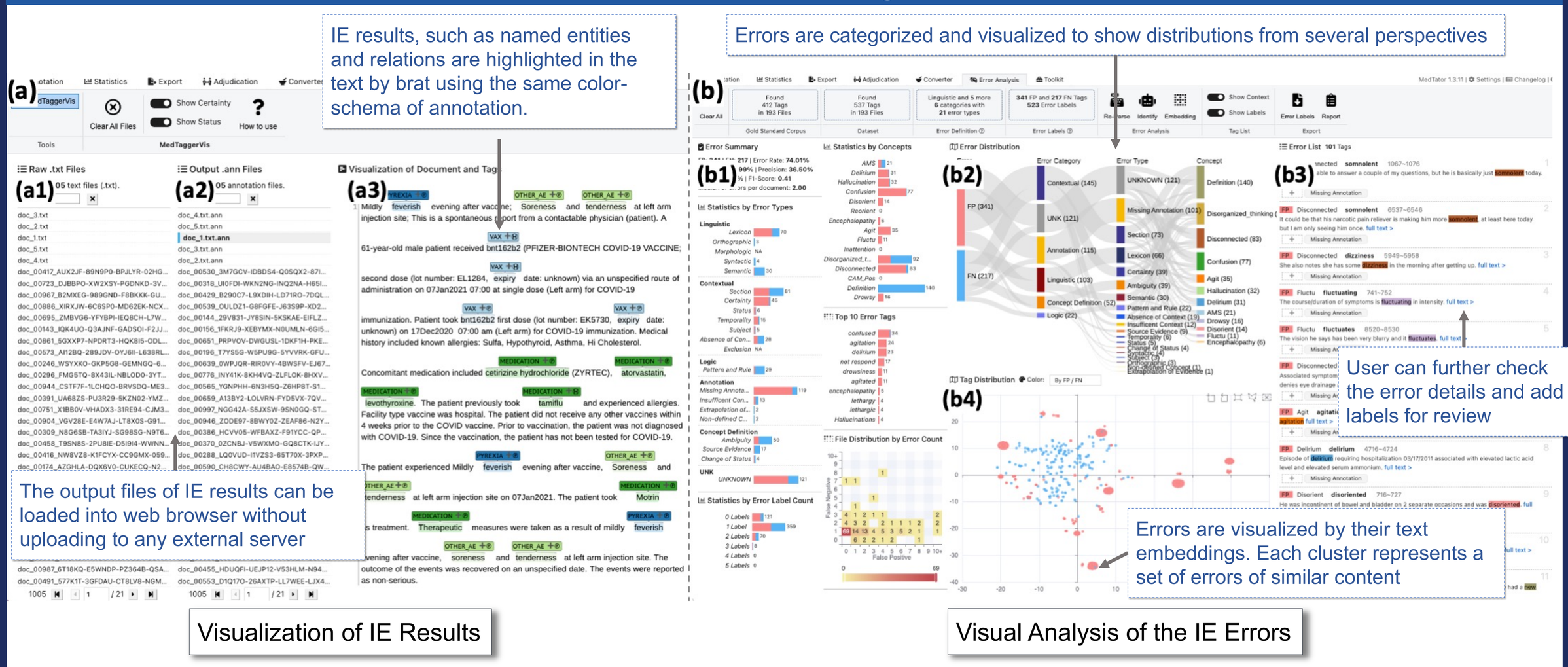
Human-computer interaction design principles

Users do not need to install any programming language runtime or server to use the tool. As shown in the following figure, we designed two tabs in MedTator to visualize the IE results (a) and errors between the gold standard and IE system outputs (b).

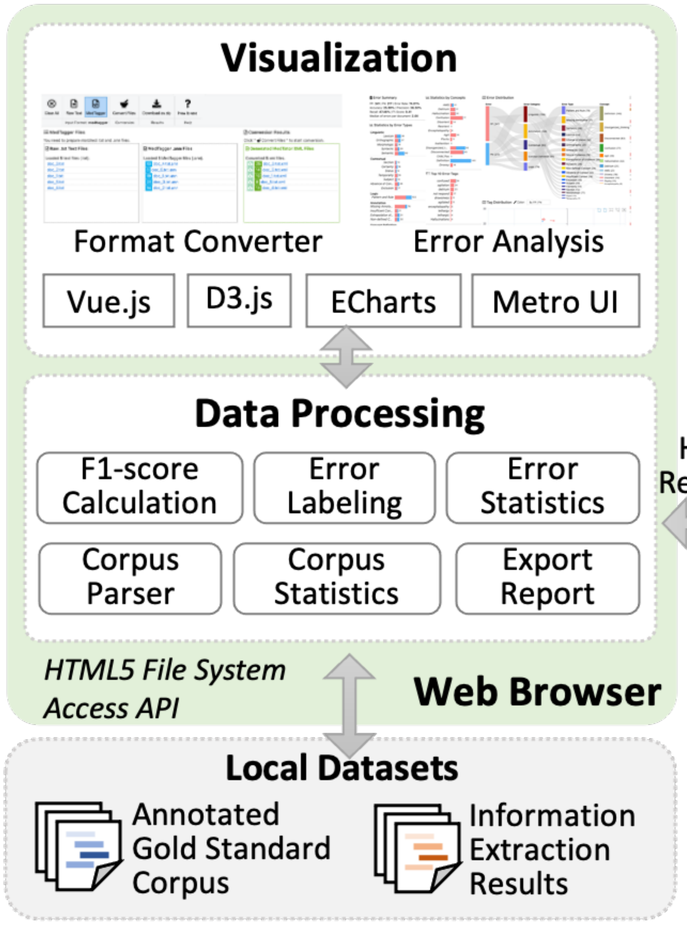
To explore the IE results, users can upload raw text files (a1) and IE result files (a2) by dragging and dropping them from local disk. Then the user can click on the file name to check the visualized IE system outputs with contextual text (a3) based on brat visualization.

To analyze the IE errors, users can upload the annotated gold standard, IE system output, and error definition schema into the tool. The errors are visualized from multiple aspects, such as error types (b1), categorized error counts (b2), and semantic distribution (b4). Moreover, users can click on any visual elements in any chart to examine the detailed information of error tokens (b3).

Visualization Design



System Design - A serverless architecture



A web-based ***serverless*** visual error analysis tool that facilitates the exploration of IE errors

- Completely run in the user's web browser
- All features are implemented within the browser
- A server is **NOT** required

Optional server support for advanced features:

- Error type classification based on ML
- Error token embedding and dimension reduction

Future Work

First, we plan to refine the error definition schema for general IE system errors. Users will customize the error definition schema based on their own needs.

Secondly, we plan to integrate online text embedding and dimensionality reduction algorithms to improve the performance of error token exploration.

Lastly, we plan to conduct usability test with end-users to evaluate the visualization and interactivity designs.

Acknowledgement

The authors thank Donna Ihrke, Heling Jia, Taylor Harrison, and other annotators for their expertise and valuable feedback throughout the development, the members of the AMIA, OHNLP, and N3C community for their insights on the tool design and assistance on testing, and the anonymous reviewers for their valuable comments. This work was supported by the National Center for Advancing Translational Sciences of the National Institutes of Health under award number U01TR002062.