Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Applications Note

Data and text mining

MedTator: a serverless annotation tool for corpus development

Huan He, Sunyang Fu, Liwei Wang, Sijia Liu, Andrew Wen and Hongfang Liu*

Department of Artificial Intelligence and Informatics, Mayo Clinic, Rochester, MN 55901, USA.

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on October 7, 2021; revised on December 23, 2021; accepted on December 26, 2021

Abstract

Summary: Building a high-quality annotation corpus requires expenditure of considerable time and expertise, particularly for biomedical and clinical research applications. Most existing annotation tools provide many advanced features to cover a variety of needs where the installation, integration, and difficulty of use present a significant burden for actual annotation tasks. Here we present MedTator, a serverless annotation tool, aiming to provide an intuitive and interactive user interface that focuses on the core steps related to corpus annotation, such as document annotation, corpus summarization, annotation export, and annotation adjudication.

Availability: MedTator and its tutorial are freely available from https://ohnlp.github.io/MedTator. MedTator source code is available under the Apache 2.0 license: https://github.com/OHNLP/MedTator.

Contact: Liu.Hongfang@mayo.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

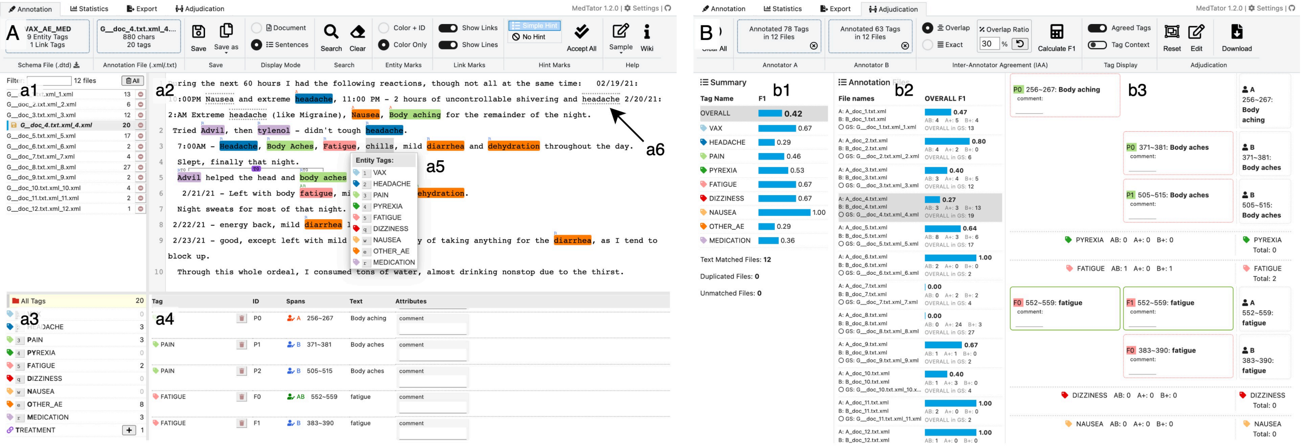
1 Introduction

The creation of gold standard annotation corpora is a critical task in natural language processing (NLP), an essential approach for extracting information from unstructured text that has been extensively employed in bioinformatic and healthcare research (Wei *et al.*, 2019; Neves and Ševa, 2021). Many tools have been developed for facilitating the annotation process, for which the installation and deployment process can broadly be categorized into two categories: stand-alone or web-based. The stand-alone tools can be installed locally and allow users to create annotations. For example, eHOST is a Java-based cross-platform annotation tool that provides machine-assisted semi-automated annotation to support large-scale annotation in clinical settings (South *et al.*, 2012). MAE is another widely used stand-alone tool that supports inter-annotator agreement (IAA) calculation to evaluate the annotation quality (Stubbs, 2011; Rim, Kyeongmin, 2016).

Recently, web-based tools have become popular as they can better facilitate large-scale annotation. For example, brat is a widely adopted web-based tool with advanced visualization functionalities for corpus annotation (Stenetorp *et al.*, 2012). PubTator Central becomes a comprehensive annotation system that integrates many features such as literature search, pre-annotation, and biological entity identification for biomedical literature (Wei *et al.*, 2019). INCEpTION is a recent general-purpose annotation tool which incorporates weak labels generated from existing NLP systems (Klie *et al.*, 2018).

Although these existing tools provide many powerful features to cover a variety of needs for corpus annotation tasks, it is challenging for users without technical expertise to implement these tools. First, those existing tools and their complex features typically require a correctly configured runtime environment, requiring extensive technical expertise to setup and maintain. Secondly, while additional features may bring benefits for certain needs such as project management and collaborative annotation, it also requires extra efforts to evaluate, learn, and apply those additional features. Moreover, if some additional features are too complicated or not suitable for the user's annotation task, they may become a burden to the user and affect the annotation efficiency.

To address these challenges, we propose and implement MedTator, a serverless annotation tool aiming to provide an intuitive and interactive user interface that focuses on the core steps related to corpus annotation. MedTator is freely available at https://ohnlp.github.io/MedTator. Comprehensive documentation, tutorials, and sample data are also available at both the MedTator website and the supplementary material.



VAX_AE_MED
9 Entity Tags
1 Link Tags

G__doc_4.txt.xml_4....
880 chars
20 tags

1 During the next 60 hours I had the following reactions, though not all at the same time: 02/19/21: 10:00PM Nausea and extreme headache, 11:00 PM - 2 hours of uncontrollable shivering and headache 2/20/21: 2:AM Extreme headache (like Migraine), Nausea, Body aching for the remainder of the night.

2 Tried Advil, then tylenol - didn't tough headache.

3 7:00AM - Headache, Body Aches, Fatigue, chills, mild diarrhea and dehydration throughout the day.

4 Slept, finally that night.

5 Advil helped the head and body aches

6 2/21/21 - Left with body fatigue, mi… ehydration.

7 Night sweats for most of that night.

8 2/22/21 - energy back, mild diarrhea l…

9 2/23/21 - good, except left with mild … y of taking anything for the diarrhea, as I tend to block up.

10 Through this whole ordeal, I consumed tons of water, almost drinking nonstop due to the thirst.

Tag	ID	Spans	Text	Attributes
	P0	A 256~267	Body aching	comment
PAIN	P1	B 371~381	Body aches	comment
PAIN	P2	B 505~515	Body aches	comment
FATIGUE	F0	AB 552~559	fatigue	comment
FATIGUE	F1	B 383~390	fatigue	comment

Annotated 78 Tags in 12 Files

Annotated 63 Tags in 12 Files

Overlap Ratio 30 %

Tag Name	F1
OVERALL	0.42
VAX	0.67
HEADACHE	0.29
PAIN	0.46
PYREXIA	0.53
FATIGUE	0.67
DIZZINESS	0.67
NAUSEA	1.00
OTHER_AE	0.29
MEDICATION	0.36

Text Matched Files: 12

Duplicated Files: 0

Unmatched Files: 0

File names	OVERALL F1
A: A_doc_1.txt.xml B: B_doc_1.txt.xml GS: G__doc_1.txt.xml_1.xml	0.47 AB: 4 A+: 5 B+: 4 OVERALL in GS: 13
A: A_doc_2.txt.xml B: B_doc_2.txt.xml GS: G__doc_2.txt.xml_2.xml	0.80 AB: 4 A+: 2 B+: 0 OVERALL in GS: 6
A: A_doc_3.txt.xml B: B_doc_3.txt.xml GS: G__doc_3.txt.xml_3.xml	0.40 AB: 3 A+: 4 B+: 5 OVERALL in GS: 12
A: A_doc_4.txt.xml B: B_doc_4.txt.xml GS: G__doc_4.txt.xml_4.xml	0.27 AB: 3 A+: 3 B+: 13 OVERALL in GS: 19
A: A_doc_5.txt.xml B: B_doc_5.txt.xml GS: G__doc_5.txt.xml_5.xml	0.64 AB: 8 A+: 3 B+: 6 OVERALL in GS: 17
A: A_doc_6.txt.xml B: B_doc_6.txt.xml GS: G__doc_6.txt.xml_6.xml	1.00 AB: 2 A+: 0 B+: 0 OVERALL in GS: 2
A: A_doc_7.txt.xml B: B_doc_7.txt.xml GS: G__doc_7.txt.xml_7.xml	0.00 AB: 0 A+: 2 B+: 2 OVERALL in GS: 4
A: A_doc_8.txt.xml B: B_doc_8.txt.xml GS: G__doc_8.txt.xml_8.xml	0.00 AB: 0 A+: 24 B+: 3 OVERALL in GS: 27
A: A_doc_9.txt.xml B: B_doc_9.txt.xml GS: G__doc_9.txt.xml_9.xml	0.67 AB: 1 A+: 1 B+: 0 OVERALL in GS: 2
A: A_doc_10.txt.xml B: B_doc_10.txt.xml GS: G__doc_10.txt.xml_10.x...	0.40 AB: 1 A+: 3 B+: 0 OVERALL in GS: 4

Fig. 1 Screenshots for MedTator user interface. (A) The annotation tab, including (a1) the file list view, (a2) the tagging view, (a3) the concept list view, (a4) the tag list view, (a5) popup menu showing the entity concepts to be tagged, and (a6) the annotation hints based on annotated tags. (B) The adjudication tab for (b1) computing the overall IAA, (b2) checking the IAA of each document, (b3) comparing annotations, and creating adjudication copy.

2 System Description

MedTator is designed to be an easy-to-install and easy-to-use corpus annotation tool by adopting a serverless design, i.e., no need of any server and any installation of a runtime environment. It leverages pure frontend web techniques (e.g., HTML5 and JavaScript) and open-source JavaScript libraries to implement features needed for corpus annotation that can run on a web browser. MedTator can be accessed via network as a web page or launched locally as a stand-alone tool. The serverless design ensures all the data processing is handled through the web browser from the local machine, addressing potential data security and privacy concerns (e.g., clinical documents with patient health identifiers). Technical details are available in the supplementary material, as well as on the wiki page of the source code repository.

MedTator contains the following tabs with customized visualization and interactive designs to facilitate core annotation tasks, i.e., document annotation, corpus summarization, annotation export, and annotation adjudication (Fu *et al.*, 2020).

Annotation tab. This tab allows the user to annotate documents according to a schema through four coordinated views (Figure 1(A)), including: (1) the file list view (Figure 1(a1)), which shows a summary of annotation documents; (2) the tagging view (Figure 1(a2)), which shows the document content and visualized annotated tags, and annotation hints in said document; (3) the concept list view (Figure 1(a3)), which shows all concepts in the schema and the count of each concept annotated; and (4) the tag list view (Figure 1(a4)), which shows the detailed information of the annotated tags, such as spans, text, and attributes.

The tagging view is the main interface for annotation, which allows the user to create or revise annotations on a selected document with customizable settings. The interface with default settings has provided necessary features for annotation, the user can customize the settings by using the menu, such as display modes and annotation hints. To reduce the repetitive workload, MedTator is capable to optionally shows hints on potential words of interest based on real-time statistics of the annotated tags (Figure 1(a6)). In addition, the real-time statistical results (e.g., the number of annotated entities per document, per concept, etc.) are displayed on the file list view (Figure 1(a1)) and concept list view (Figure 1(a3)) to show annotation progress.

Statistics tab. The detailed information of the statistics on the annotated tags is shown in this tab, which allows the user to analyze both the overall status and the annotated tags. The user can also trace back to the source document of each annotated tag in the annotation tab through hyperlinks.

Export tab. The export tab allows the user to convert the annotations to different formats for downstream tasks. For example, the annotated tags and context sentences can be converted to a tab-separated values file for corpus analysis, a keyword dictionary for building rule-based system (e.g., MedTagger (Liu *et al.*, 2013)), an inside-outside-beginning format file for training models for named-entity recognition, or a BioC format file for other NLP tasks (Comeau *et al.*, 2013).

Adjudication tab. This tab allows the user to compute IAA, compare annotations of two annotators, and create an adjudication copy for final annotations (Figure 1(B)). The tool currently uses F1-score to assess IAA (Hripcsak and Rothschild, 2005) and the comparison of annotations can be visualized at different levels for interactive comparison (Figure 1(b1 and b2)). An adjudication copy can be created by accepting or rejecting each annotator's annotation which can then be used for creating final annotations (Figure 1(b3)).

3 Usage

MedTator enables users to start annotation rapidly in just two steps: upload a schema file (".dtd") that defines the annotation task and start annotation by opening raw text files (".txt") or annotated files (".xml"). Once the annotation is finished, users can save the annotation or export annotation for downstream applications such as building rule-based NLP system, training model for named-entity recognition, or other tasks. More details of each tab of MedTator are available via wiki pages and usage manual (see supplementary material).

4 Conclusion

We developed a serverless annotation tool, MedTator, that enables rapid corpus development for a variety of research needs without any installation of a runtime environment. MedTator aims to provide an intuitive and interactive user interface with customized visualization and interactive designs, which offers researchers a lightweight solution to facilitate the core steps in corpus annotation, i.e., document annotation, corpus summarization, annotation export, and annotation adjudication. The tool has been piloted for multiple annotation tasks with very positive feedback. We plan to further improve the usability with a formal usability study planned.

Acknowledgements

The authors wish to thank members in the OHNLP community for their insights on the tool design and valuable feedback.

Funding

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

Conflict of Interest: none declared.

References

Comeau,D.C. *et al.* (2013) BioC: a minimalist approach to interoperability for biomedical text processing. *Database J. Biol. Databases Curation*, **2013**, bat064.

Fu,S. *et al.* (2020) Assessment of the impact of EHR heterogeneity for clinical research through a case study of silent brain infarction. *BMC Med. Inform. Decis. Mak.*, **20**, 60.

Hripcsak,G. and Rothschild,A.S. (2005) Agreement, the F-Measure, and Reliability in Information Retrieval. *J. Am. Med. Inform. Assoc. JAMIA*, **12**, 296–298.

Klie,J.-C. *et al.* (2018) The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. In, *Proceedings of the 27th COLING : System Demonstrations*. Santa Fe, New Mexico, pp. 5–9.

Liu,H. *et al.* (2013) An Information Extraction Framework for Cohort Identification Using Electronic Health Records. *AMIA Summits Transl. Sci. Proc.*, **2013**, 149–153.

Neves,M. and Ševa,J. (2021) An extensive review of tools for manual annotation of documents. *Brief. Bioinform.*, **22**, 146–163.

Rim, Kyeongmin (2016) MAE2: Portable Annotation Tool for General Natural Language Use. In, *Proceedings of the 12th Joint ACL-ISO Workshop on Interoperable Semantic Annotation*. Portorož, Slovenia.

South,B. *et al.* (2012) A Prototype Tool Set to Support Machine-Assisted Annotation. In, *Proceedings of the 2012 BioNLP Workshop*. ACL, Montréal, Canada, pp. 130–139.

Stenetorp,P. *et al.* (2012) brat: a Web-based Tool for NLP-Assisted Text Annotation. In, *Proceedings of the EACL 2012: Demonstrations*. ACL, Avignon, France, pp. 102–107.

Stubbs,A. (2011) MAE and MAI: Lightweight Annotation and Adjudication Tools. In, *Proceedings of the 5th Linguistic Annotation Workshop*. Association for Computational Linguistics, Portland, Oregon, USA, pp. 129–133.

Wei,C.-H. *et al.* (2019) PubTator central: automated concept annotation for biomedical full text articles. *Nucleic Acids Res.*, **47**, W587–W593.