Towards User-centered Corpus Development: Lessons Learnt from Designing and Developing MedTator

Huan He, PhD, Sunyang Fu, PhD, Liwei Wang, MD, PhD, Andrew Wen, MS, Sijia Liu, PhD, Sungrim Moon, PhD, Kurt Miller, MS, and Hongfang Liu, PhD Department of Artificial Intelligence and Informatics, Mayo Clinic, Rochester, MN, USA

Abstract

A gold standard annotated corpus is usually indispensable when developing natural language processing (NLP) systems. Building a high-quality annotated corpus for clinical NLP requires considerable time and domain expertise during the annotation process. Existing annotation tools may provide powerful features to cover various needs of text annotation tasks, but the target end users tend to be trained annotators. It is challenging for clinical research teams to utilize those tools in their projects due to various factors such as the complexity of advanced features and data security concerns. To address those challenges, we developed MedTator, a serverless web-based annotation tool with an intuitive user-centered interface aiming to provide a lightweight solution for the core tasks in corpus development. Moreover, we present three lessons learned from the designing and developing MedTator, which will contribute to the research community's knowledge for future open-source tool development.

Introduction

Natural language processing (NLP) has been widely applied in healthcare-related domains for clinical practice and research field ^{1,2}. While a high-quality annotated corpus is usually indispensable for developing NLP models, building such a corpus requires considerable time and expertise to go through the whole annotation process with a comfortable annotation tool³. Consequently, there have been many comprehensive annotation tools developed in the past to facilitate the annotation process ³, and they can provide many advanced features to help users reduce the burden of annotation in many ways. Most of the tools are open-source that can be obtained freely.

For example, brat⁴ is a popular web-based tool with advanced visualization functionalities. PubTator Central integrates PubMed to provide many features such as literature search and biological entity identification for biomedical literature. ezTag⁵ supports annotating PubMed abstracts and PubMed Central full-text articles in BioC format with customized lexicons. INCEpTION ⁶ is a general-purpose annotation tool that incorporates weak labels generated from existing NLP systems. TeamTat⁷ provides the collaborative annotation ability to help the researcher to manage a team annotation project. MyMiner supports the biomedical document annotation task with pre-defined biological entities such as proteins and genes⁸. LabelStudio⁹ provides application programming interfaces (APIs) for developers to integrate the annotated data with other systems. In addition to the above-mentioned open-source tools, there are also a lot of commercial web-based tools, such as prodigy, tagtog, LightTag, etc. Those tools can provide more advanced features such as multi-media data annotation, assistance based on machine learning, online team collaboration, etc.

However, while existing tools provide many powerful features, it is still challenging for research teams to leverage them in their own research task due to three main factors. First, while existing tools bring many advanced corpus annotation features, annotators must spend extra effort to learn, evaluate, and apply those features to their own tasks. Secondly, existing web-based tools often need a correctly configured web server environment as one of the prerequisites, which requires extra technical support. Finally, and most importantly, due to patient privacy concerns, clinical documents containing protected health information (PHI) cannot be sent to external servers, and users usually can only annotate the documents in a restricted environment (e.g., no internet and no permission of tool installation). As a result, the ability to run in a restricted machine is required for an annotation tool to be adopted in clinical settings.

To address these challenges, we developed MedTator¹⁰, a serverless web-based annotation tool aiming to provide an intuitive and interactive user interface that focuses on the core steps of corpus annotation. More importantly, there are three lessons we learned during the development and maintenance, which are about the needs of users, the non-functional requirements, and the community participants. These lessons we learned are applicable to a range of similar tools and thus benefit the clinical research community.

We have released the source code on GitHub under the Apache 2.0 license: <u>https://github.com/OHNLP/MedTator</u>. The documentation, such as the user manual and sample annotations can be found in the MedTator wiki in the same GitHub repository. In addition, a public online demo of MedTator is also available with several sample datasets at <u>https://ohnlp.github.io/MedTator</u>.

System Design

Motivation, methods, and system features

Clinical document annotation was always demanding in our past clinical NLP research projects, and one of the most valuable assets of these projects is the annotated corpus. To fully utilize those annotated corpora, a tool is needed to explore the documents and the annotated tags for corpus analysis. Therefore, MedTator was initially developed as a web-based visualization tool for exploring the annotated corpus.

Over time, additional new features such as revising annotated tags, tag statistics, and loading customized schema, were added to meet the needs of different projects. The newly added features were welcomed by researchers, which suggests that MedTator can be further improved into a text annotation tool. Then, we began the design process to shape the functional requirements and non-functional requirements for MedTator as follows.

Firstly, we collected and installed popular annotation tools from literature reviews to understand the workflow and user experience of existing tools to form initial design ideas.

Secondly, we invited researchers who have extensive experience in NLP and clinical research as domain experts and interviewed them to assess our initial ideas and the prototype. With their suggestions, we summarized the goal of MedTator as "a lightweight, easy-to-use text annotation tool" and the core value is simplicity, which guides the design decisions in the development process.

We iterated through weekly design and development cycles to discuss the latest progress and gather feedback from our domain experts and early users. Each cycle began with a prototype demonstration of the latest features developed (e.g., new functionalities or visual designs), and the latest prototype will be uploaded to our internal server to be tested by domain experts and users. The comments on the prototype will be collected to form new functional and nonfunctional requirements. Those new requirements were evaluated by our domain experts to decide whether to be included in the development plan or not. In addition, we also held training sessions and development meetings with annotators and researchers to review the features and issues as needed. Their comments were also utilized in the design and development. To manage the workload and roadmap, we used GitHub Project and Microsoft Teams to track the design and development tasks.

After about 3-4 months of iterative design and development cycles, the prototype was upgraded to a fully functional annotation tool with the following core features implemented to meet the functional requirements:

- **Text annotation**: this is the core task in the whole annotation process. MedTator provides core functionalities to meet the needs of this task, such as annotated file parsing and saving; text span annotation at levels of characters, words, sentences, and documents; semantic relationship annotation between multiple text spans; multiple document annotation; and customizable annotation schema for various tasks.
- Annotation statistics: the statistics on annotated tags are demanded by researchers. MedTator supports statistics on the annotated corpus from multiple aspects, which can be used in all stages of the annotation process to understand the annotated corpus. For example, at the initial stage of the annotation guideline development, the overall statistical results on the small size sample corpus can be used to revise the schema and guideline; during the iterative annotation, the results can be useful for tracking the annotation process; and in the final evaluation stage, the results can help to evaluate the annotation quality.
- Adjudication: to resolve the inconsistencies among annotations from different annotators, MedTator supports adjudication on multiple annotation versions to help researchers to create the gold standard corpus. For example, the measurement of the inter-annotator agreement (IAA) is needed to help researchers estimate the overall annotation quality. Moreover, the agreements among multiple annotation versions need to be measured from multiple levels (e.g., overall agreement, concept level, document level, tag level, attribute level, etc.) to help researchers locate and investigate the disagreements.
- Annotation export: MedTator supports the conversion of the annotated corpus to different formats. The annotated corpus itself is usually not the final goal of our research projects but for the corpus analysis and system development. According to our previous experience in clinical research, the annotated corpus can be used in many ways, including but not limited to: corpus analysis to understand the cohort¹¹, knowledge discovery ¹², training machine learning or deep learning models¹³, and building NLP systems ¹⁴, etc. Therefore, multiple formats, such as CSV, BioC ¹⁵, and MedTagger rulesets ¹⁶, are supported by MedTator for various downstream tasks or other software.



Figure 1. The MedTator architecture based on a serverless design. MedTator can be accessed as a web page from (A) public internet using GitHub pages or other public server, or (B) any static HTTPs server in the intranet. (C) It can also run as a standalone program locally without any network access. All functionalities are implemented based on web frontend techniques within the web browser environment. MedTator can only access local storage for read / write annotated corpus without sending / receiving out any data outside of local machine.

In addition to the above features for the functional requirements, we summarized the non-functional requirements, such as data security, maintainability, and easy-to-install. To meet these requirements, we adopted a serverless architecture to design MedTator. The details are described in the following sections.

Serverless architecture

Based on the above-mentioned requirements and our expertise in software engineering, we adopted a serverless architecture to design MedTator. Serverless is an emerging application design that removes much of the need for a traditional always-on server and incorporates third-party services¹⁷. Using this design, MedTator works like a single-page application, which is assembled as one HTML file with several external JavaScript libraries. Users can access it from any web server through HTTPs protocol or locally run it as a standalone program.

As shown in Figure 1, both the user interface and core modules of MedTator are implemented as a single-page web application within a web browser environment based on pure frontend web techniques, including JavaScript, HTML5, web APIs, and many JavaScript libraries such as Vue.js, D3.js, and CodeMirror. The user interface contains customized visualization and interactivity designs that collect user input and present the processed data. The core modules support the core functionalities described in the previous section. The data exchange between the user interface and the core modules is completely restricted to the user's local web browser, and all information of the annotated corpus is saved in the user's local storage to avoid any leakage of PHI.

As a serverless tool, this architecture allows users to use the tool without managing servers. The whole tool is packaged into a single HTML file, which can be accessed in three ways: 1) as a public webpage served by GitHub Pages or any other public server with the third-party libraries provided by public content delivery network (CDN) (Figure 1A); 2) as an internal webpage served by a HTTPs server on intranet website with cached third-party libraries (Figure 1B); and 3) as a static webpage that can be opened by local web browser directly (Figure 1C). All the dependent third-party libraries can be loaded through local disk, and all the traffics of the network can be monitored in the web browser console to ensure data security.

The serverless architecture brings the following advantages to MedTator. First, there is no need to manage a server to host the tool and install any programming runtime environment at all, such as Java, Python, or Node.js on the user's local machine. Secondly, end users can easily access the program in a secure way. Last and most important, no bit of data would be sent out from the user's local machine which is usually behind firewall to address security concerns.

VAX_AE_MED 9 Entity Tags 1 Link Tags	A_doc_4.txt. 880 chars 9 tags	xml 🐻 🕻 Save Save		~ ~ ~ ~		ID Show Link	No Hint	ept All Sample Wiki	Ciear All	Annotated 78 Tags in 12 Files	Annotated 70 Tags in 12 Files	± Overlap 2 Over Exact 30	Nap Ratio	-	Agreed Tags Tag Context	Reset Edit	Download	
Schema File (.dtd) 🛦	Annotation File (.x	emi/.txt) Save	Display Mode	Search	Entity Mar	s Link Marks	Hint Marks	Help		Annotator A	Annotator B	Inter-Annot	stor Agreement (IAA)	Та	ag Display	Adjudica	ation	
r: 12 files	14 B	1 During the ne:	t 60 hours I h	ad the follo	wing reaction	s, though not all	at the same time	: 02/19/21: 10:00PM	I≣ Summary		IE Annotation Files		04 406-411 faint	certainty:				A 406-411:
o_1.tet.ami	9 0	Nausea and	extreme mbead	ache 11:00	PH = 2 hours	of uncontrollabl	e shivering and	headache 2/20/21: 2:AM	Tag Name	F1	File names	OVERALL F1	comment:	Positive				faint
	6 0								OVERALL	0.43	A A doc 1tst ami	0.47	may merge to diza	ty .				
e_aloc_4.tet.aml	9 0	192					remainder of the	night.			B: B_doc_1.txt.xml	AB:4 A+:5 B+:4						
5.tet.orni	12 0	2 Tried Advi	🛛 , then 🔢 Tyle	anol - didn'	t tough 回 hea	dache .			VAX	0.67	O 05: 0doc_ltxt.xml_l.xm				00 1	1407~1424: face w	as very red	▲B 1407~14
c_6.tat.oml	2 0	3 7:00AM - H	headache, 🔯 Bo	dy aches,	fatique, chi	lls, mild 🖸 diar	chea and 🚺 dehydra	ation throughout the day.	HEADACHE	0.29	B: B_doc_2.txt.xml	0.80				ainty: comment: aitive or red?		face was very
c_7.txt.aml c_8.tat.xml	2 0		ly that night.		Entity T				PAIN	0.46	O 05: 0_doc_2.txt.xml_2.x	01000001000.0			POS	answe or red /		
c_9.txt.xml	2 0								PYREXIA	0.53	A: A_doc_3.txt.xml B: B. doc_3.txt.xml	0.53						* B 1430-14
ic_10.tst.oni	4 0		ped the head as		2 HE	ADACHE	_		FATIGUE	FATIGUE 0.67	0 05: 0_doc_3.txt.xml_3.xm	OVERALL IN OS: 14			O1 1 certs	1430-1456: eyes were builging slightly certainty: comment:	B 1430-1456: eves were bulois	
ins.tot.eni			eft with body	ro fatigue ,	J PA		on .		DIZZINESS	0.67	A: A_doc_4.txt.xml B: 8_doc_4.txt.xml	0.26			Positive or just bulging?		ng?	slightly
_doc_12.txt.xml	1 0	7 Night sweat	Night sweats for most of that night. I 2/22/21 - energy back, mild I diarthea I 4 I 2/22/25 I 2/22/21 - energy back, mild I diarthea I 2/22/25 I 2/22/21 - energy back, mild I I I I I I I I I I I I I I I I I I I						NAUSEA	0.80	0 05: 0_dot_4.txt.xml_4.x	0.64					MEDICATION Total: 4	
		8 2/22/21 - en							OTHER_AE	OTHER_AE 0.27	A: A_doc_5.txt.xml		MEDICATION		TION AB: 2	AB: 2 A+: 0 B+: 2		
		9 2/23/21 - good, except left with mild 0 diari > NAUSEA taking anything for the 0 diarrhea, as I tend to						MEDICATION	0.41	B: B_doc_S.txt.xml O GS: Gdoc_S.txt.xml_S.xml	AB: 8 A+: 3 B+: 6 0 OVERALL IN CO. 17							
		block up.	-,,		• • 01		,,		Text Matched Fil		A: A_doc_6.txt.xml	1.00	MT 1568-1574: Ep	comment:		1568-1574: EpiPen	comment	A 1568~153
All Tags	9 1		ID Spar	n Text	© z M	DICATION					8: 8_doc_6.txt.xml O 05: 0doc_6.txt.xml_6.x	AT 2 As 0 To 0	M1 1568-1574: Ep	Pen	M1 1	1568-1574: EpiPen		EpiPen
VAX		-			-				Duplicated Files:	0	A: A_doc_7.txt.xml	0.00						
HEADACHE	2	OTHER_AE	02 5	76-587 dehyd		comment			Unmatched Files	E 0	R B doc 7.txt.xml	ARC 8 441 9 1041 9	M2 470-478; ice s	comment.	112	470~488: ice pack	and water	▲A 470~478:
PAIN	1				Positiv						0 65: 6_doc_7.txt.xml_7.xt		Ma 470-470.1001		com			pack
PYREXIA	0 9	MEDICATION	M0 3	26-311 Advil	commer						A: A_doc_8.txt.xml B: 8_doc_8.txt.xml	0.00						
FATIGUE	1										O 05: 0doc_8.txt.xml_8.x						comment:	
DIZZINESS	0	HEADACHE	<u>в</u> н1 з	41-349 heada	commer						A: A_doc_9.txt.xml B: B_doc_9.txt.xml	0.67			MD 1	1143~1149: EpiPen		▲ B 1143~114 EpiPen
NAUSEA																		
											0 05: 0_doc_9.txt.xml_9.x	ni OVERALL in GS: 2						
	3	P TREATMENT	10 L	NK	link_AE	link_MED					A: A_dec_10.txt.xml	0.40					comment.	
OTHER_AE MEDICATION	1		10 L	NK		DICATION - AI H1 HEA		MedTator 1.2.5 🖗 Settings	O DAnnotation	Leff Statistics		0.40 AB:1 A:3 B:0 - OVERALL is OS:4			83	643~651: ice pack		nack
Amotation Lef Statis	1		9 TO L	NK.				MedTator 12.5 Ø Settings	12 Documents	2	A A_doc_30.txt.xml B: E_doc_30.txt.xml O 05: 0_doc_30.txt.xml_30 Export HA Adjudicatio	AB 1 Ar 3 Br 0 AB 1 Ar 3 Br 0 Ar 0 VERALL in OS 4	J.			643-651: ice pack • to use the exporte	Med	pack
OTHER_AE MEDICATION Introtation Let Statist Ether Summary (cov)	1		9 TO L	NK				MedTator 1.2.5 Ø Settings	an ourmary:	S Tag Text Tag & Se	A A_dec_30.txt.anl B_B_dec_30.txt.anl O 05: 6_dec_30.txt.anl_30 Export H Adjudicatio IOB2/BIO (5xv)	AB:1 Ar:3 Br:0 AB:1 Ar:3 Br:0 ModTagger (. BioC (.xm)	zip) spaCy (.jsoni)	How to use			Med	pack
OTHER_AE MEDICATION Summary (cov) Download	a 1	t 😽 Adjudication	10 L	NK				MedTator 1.2.5 Ø Settings	12 Documents 81 Annotated Tay	2 ¹⁵ Tag Text Tag & So Text	A A_doc_30.txt.xml B: E_doc_30.txt.xml O 05: 0_doc_30.txt.xml_30 Export HA Adjudicatio	AB:1 Ar:3 Br:0 AB:1 Ar:3 Br:0 ModTagger (. BioC (.xm)					Med	▲8 643-651 nack
OTHER_AE MEDICATION anotation Lef Statis Summary (cov) Download ummary	a 1 ics b Export	t H Adjudication			MOIN	DICATION - A	DACHE - hear		Export Prev	25 Tag Text Tag & Sv Text	A Adde: 33 to a mill B Adde: 33 to a mill O GHS - 35 to a mill	0.40 AB 1 A 3 B+ 0 AB 1 A 3 B+ 0 B+ 0 Overhald in OS 4 BioC (xmi) BioC (xmi) MedTagger (zip) spaCy (.jsoni)	How to use			Med	pack
OTHER_AE MEDICATION Ltd Statis each Summary (osv) Download kummary documents:	ics B-Export	t 😽 Adjudication	6 *2 Body act	es ×1 Br	no i M	I achy •	2 aches •1	aching #1 pain	es dummary: 12 Documents 81 Annotated Tay Export Prev concept text	2 ²⁵ Tag Text Tag & Sv Text Text tiew t doc_span anxiety 84,93	A A_dec_30.txt.anl B_B_dec_30.txt.anl O 05: 6_dec_30.txt.anl_30 Export H Adjudicatio IOB2/BIO (5xv)	0.40 A8:1 A:3 8:0 A:0 A8:1 A:3 8:0 A:0 A:0 <td>zip) spaCy (.jsoni) Ruleset</td> <td>How to use Help</td> <td>xamples of how</td> <td>r to use the exporte</td> <td>t Med</td> <td>flator 12.5 Q Set</td>	zip) spaCy (.jsoni) Ruleset	How to use Help	xamples of how	r to use the exporte	t Med	flator 12.5 Q Set
OTHER_AE MEDICATION Menotation Let Statist Commence Summary documents: tags in schema:	3 1 1 1 1 1 1 1 2 9	t H Adjudication	6 *2 Body act	es ×1 Br	no i M	1 achy * A doc_Statami 1	DACHE - hear 2 aches #1 > A_doc_Ztst.ami 1>	aching #1 pain	Export Prev Concept Lexit	2 ^s Tag Text Tag & S Text tiew t doc_span anxiety 86,93 pervousness	A A Addr:33tra anti B B.der:33tra anti B B.der:33tra anti D D3: 0_me_33tra anti D D4: 0_me_33tra anti D D2: 0_	0.40 A8:1 A:3 8:0 A:0 A8:1 A:3 8:0 A:0 A:0 <td>zip) spaCy(json) Ruleset</td> <td>How to use Help</td> <td>xamples of how</td> <td>to use the exporte</td> <td>t Med ad ruleset files.</td> <td>Tator 12.5 Ø Se</td>	zip) spaCy(json) Ruleset	How to use Help	xamples of how	to use the exporte	t Med ad ruleset files.	Tator 12.5 Ø Se
OTHER_AE MEDICATION Lanstation Let Statist Summary (cov) Deveload Bummary documents: Lags in schema: annotations:	3 1 1 12 9 81	t H Adjudication	6 *2 Body act	es ×1 Br	no i M	1 achy * A doc_Statami 1	2 aches •1	aching #1 pain	Barter State State State	Tag Test Tag & So Teg Test Tag & So Test tiew t doc_span anxiety 86,93 nervourness sociare.	A A ARC-Stanumi B R_stanumi O 69: 0det_35 int am(32 Export H Adjudication intence IOB2/BIO (truy) Dataset 94:0,32 A_doc_1.txt. 94,105 94,105 A_dot	A 40 Mai 1 Aug 2 are 0 Mai 1 Aug 2 are 0 Mai 1 Aug 2 are 0 Med Tagger (. a Med Tagger (. a ment sentence mai 1 rule to receivin c_l.txt.xzl Prior to	z(p) spaCy(json) Ruleset g her COVID-19 vac receiving her COVID	How to use Help cine, the patien D-19 vaccine, th	xamples of how it was alread se patient wa	to use the exporte dy having sympto as already having	d nuleset files.	flator 12.5 Ø Set
orteR_AE INEDICATION Internation Being Summary (oor) Download Nummary documents: tags in schema: annotations per tags	3 1 1 1 1 1 1 1 1 3 1	t H Adjudication	6 *2 Body act	es •1 Br Istami 1 > A	no i M	1 achy * A doc_Statami 1	DACHE - hear 2 aches #1 > A_doc_Ztst.ami 1>	aching #1 pain	Barten State State State State State State State State State State State State State	t doc_span arxiety 86,93 nervourseas disy 199,200	A 4.64/2 Storent B 5.4/4/2 Storent O 68 0_dex_35.4/4/2 D 68 0_dex_35.4/4/2 D 80 0_0 (164) D 802/80 (164)	AD 2.0 AD 2.0 AD 2.0 AD 2.0 BioC (JMD) BioC (JMD) ModTagger (A ModTagger (A AD 2.0 ModTagger (A AD 2.0 AD 2.0	zip) spaCy(jsoni) Rudeset g her COVID-19 vac- receiving her COVI en her injection inj	How to use Help	xamples of how at was alread to patient was instee, she k	to use the exporte dy having sympto as already having es, she began feeling di	Med of ruleset files.	stator 12.5 Ø Se enervoussets abo anklety/nervous ded."
OTHER_AE MEDICATION Instation Let Statist Semmary (onv) Download Download ummary documents: tags in schema: annotations per tags annotations per tags	3 1 1 1 1 1 2 8 1 1 3 3 1 1 2 8 1 1 3 1 1 1	t HAgudication	6 • 2 Body ad 2 > A_doc,2 2 +3 fever >	es •1 Br Istami 1 > A	no i M	1 achy * A doc_Statami 1	DACHE - hear 2 aches #1 > A_doc_Ztst.ami 1>	aching #1 pain	Concept text Concent Concept text Concept text Concept text Concep	Tag Text Tag & S Tag Text Tag & S Text tiew t doc_span arxiety 86,93 nervourness scctrs. tipipen 1566,1 at me and eaid **17	A A LAC State with B Exec. State with D 68 C _mex. 35 Mick and 38 B Exec. State with D 68 C _mex. 35 Mick and 38 D 60 C _mex. 3	ALC LAS ALC LAS ALC AND ALC A	zip) spaCy(jsoni) Ruleset g her COVID-19 vac- receiving her COVI en her injection an was given her inj ted her with the R	How to use Help Dine, the patien D-19 vaccine, th ad within 1-2 mi sortion and withi piPen immediatel	xamples of how the was alread to patient we instee, she b in 1-2 minutes	by to use the exporte dy having symptons as already havin began feeling di se, she began in roughly 5 see	Med ad ruleset files.	sTator 12.5 \$ Se stator 12.5 \$ Se nervousness abo anxiety/nervous ded." ishtheaded." ck to conscious
e OTHER_AE MEDICATION Annotation Let Statistic tesh Summary (cov) Devertised Summary documents: Inanotations: annotations per docume Isentences:	3 1 1 1 1 1 1 1 1 3 1 1 1 3 1 1 3 1 1 3 1	t HAquidication	6 *2 Body act 2 > A_doc_2	es •1 Br Istami 1 > A	no i M	1 achy * A doc_Statami 1	DACHE - hear 2 aches #1 > A_doc_Ztst.ami 1>	aching #1 pain	Aummary: 12 Documents 81 Annotated Ta Concept text Concept text Concept text dots	te doc span articley 64,93 sectors be doc span articley 64,93 sectors	A A. A. C. Stander B E. B. C. Stander D ER C	0.0 0.0 0.01 0.0	zip) spaCy(jsoni) Ruleset g her COVID-19 vac- receiving her COVII en her injection a was given her inj ted her with the R The vaccine admin	How to use Help Dise, the patien D-19 Vaccine, th dwithin 1-2 mi evicion and withi piPen immediatel Latrator came to	wamples of how it was alread so patient was instead, she b in 1-2 minute y, and within o me (pharmaco	dy having sympto dy having sympto as already havin began feeling di ss, she began f in roughly 5 sec cist on daty) to	Med ad nuleset files. ons from anxiety/r ng aymptona from a isny and lighthese weiling dizzy and i conds abe came bas o notify me that s	STator 12.5 Ø Set stator 12.5 Ø Set enrvoussess abo anxiety/nervous ded." lightheaded." k to conscious she was not fee
THELAE MEDICATION Menication E Summary (cov) Deveload Summary documents: Lags in schems: annotations per docume sentences:	3 1 1 1 1 1 2 2 8 1 1 3 3 1 2 9 8 1 1 1 3 5 1 2 5 1 1 2 5 1 1 1 1 1 1 1 1 1 1 1 1	t HAgudication	6 • 2 Body ad 2 > A_doc,2 2 +3 fever >	es •1 Bc Listurni 1 > A 3 feverish	ody athing •	1 achy * A doc_Statami 1	DACHE - hear 2 aches #1 > A_doc_Ztst.ami 1>	aching #1 pain	 adammary: 12 Documents 81 Annotated Ta (a) Export Prev concept. text concept.	Tag Text Tag & St Tag Text Tag & St Text text	A A.G.C.Stanut B E.G.C.Stanut D ER C., MIL, SHARE, SH B E.G.C.S.H.C.S.M.C.S.M.C.S. B C.C.M.C.S.M.C.S.M.C.S. D B C. MIL, SH C.S. D B C.S. MIL, SH C.S. MIL, SH C.S. D B C.S. MIL, SH C.S. MIL, SH C.S. MIL, SH C.S. MIL, SH C.S. MIL, SH C.S. MIL, SH C.S. MIL, SH C.S. MIL, SH C.S. MIL, SH C.S. MIL, SH C.S. MIL,	0.40 0.40 0.11 0.15	<pre>zbp) spaCy(jsoni) Ruleset g her COVID-19 vac- receiving her COVID- in her injection as was given her injection the R in her injection as was given her</pre>	How to use Help -19 Vaccine, th dwithin 1-2 mi estion and withi piPen immediatel istrator came to 3 her name out 1	wamples of how it was alread se patient wa instea, she b in 1-2 minut y, and withi o me (pharmac oud and mudg	to use the exporte dy having symptons as already havin began feeling di may able began for in roughly 5 me cist on daty) to ged her shoulder	Med ad nukeset files. ons from anxiety/r ng appretung from a sirry and lightbeak exiles dizry and 1 coolds mic case has coold in the case has coold in the that is c serveral times, b	Banck STator 12.5 \$ Se servoussess abs anxiety/nervous ded." lightheaded." k to conscious she was not fer
orretes_AE MEDICATION which are a set of statistic set Summary (only) bound and the set of statistic documents: tags in schems: annotations per docume sentences:	3 1 1 1 1 1 2 2 8 1 1 3 3 1 2 2 8 1 1 1 3 5 1 1 2 5 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1	t HAdjudication	6 ≠2 80dy ad 2 > A_doc. 2 ×3 fever > 1 ×1 fetgue 3 ×2 dizzy 1 > A_doc.	es • 1 Bo tst.sml 3 > A 3 feverish • 1 lig tst.sml 1 > A	ody athing •	1 achy A_doc_6tstam 3 dizenees	DACHE - hear 2 aches #1 > A_doc_Ztst.ami 1>	aching #1 pain	Export Prev Concept Least Concept L	22 Tag Text Tag & C Tag Text Tag & C Text	A A. A. C. Stander B E. B. C. Stander D ER C. M. S. Stander D ER C. M. S. Stander D E. S.	BOC AO m A BCC Amod Tangare (a MadTangare (b Amod (a MadTangare (b Amod (<	200) spoCy (joon) Ruleset y ber COVID-19 van receiving her COVID m her injection as was given her inj ced har with the B The vaccine admin et.sml * I reparte	How to use Help Delte, the patien Delte vaccine, th delthin 1-2 mi socion and within Differ immediatel Listrator came to d her name out 1 d stated that she	xamples of how at was alread to patient was inites, she b inites, she b	dy having sympto dy having sympto as already havin began feeling di ss, she began fe list on daty) to ped her shouldes	I Med one from anxiety/r ng symptom from tigy and lighthese tealing dirzy and l o colify me that a c reveral lime, b	Theor 125 (0 Se rervousness ab antiety/nervou ded, " (k to contcious ahe was not fe hovever the par
orhele, AE MEDICATION Anotation ME Statist Summary (only) Downsad documents: anotations per docume sentences:	3 1 1 1 1 1 2 2 8 1 1 3 3 1 2 2 8 1 1 1 3 5 1 1 2 5 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1	t HAdjudication	6 • 2 Body ad 2 > A_doc. 2 +3 fever + 1 +1 fatigue 3 • 2 dizzy	es • 1 Bo tst.sml 3 > A 3 feverish • 1 lig tst.sml 1 > A	ody acting dec_Atatami 1	1 achy A_doc_6tstam 3 dizenees	DACHE - hear 2 aches #1 > A_doc_Ztst.ami 1>	aching #1 pain	the Jammary: 12 Documents 81 Annotated Te 12 Documents 81 Annotated Te 12 Documents	Tag Test Tag & Tag Test Tag & Tag Test Tag & Set Test Tag & Set	A A LAUSELANDER B CLAUSELANDER D CLE LANSIE CLEAR CLAUSE CLEAR CLAUSE CLEAR CLEAR CLAUSE CLEAR	A0	point space (source) Rulesst g her COVID-19 vac- recestving her COVII en her injections and was given bars inj ted her with the R The vectore addin at.sel * 1 repeate ed kon to stay new from to stay new from to stay new	How to use Help Dine, the patien >-19 Vaccine, th ad within 1-2 mi sction and within jefen immediatel latrator came to a her name out 1 stated that she used with a stated that she	xamples of how at was alread to patient was in 1-2 minute, she b in 1-2 minute y, and within me (pharmac ioud and nudg r felt as the the would like the would like	dy having sympto dy having sympto as already havin began feeling di ss, she began fe list on daty) to ped her shouldes	I Med one from anxiety/r ng symptom from tigy and lighthese tealing dirzy and l o colify me that a c reveral lime, b	Theor 125 (0 Se rervousness ab antiety/nervou ded, " (k to contcious ahe was not fe hovever the par
orhele, AE MEDICATION Anotation ME Statist Summary (only) Downsad documents: anotations per docume sentences:	3 1 1 1 1 2 0 81 1.33 1 2 2 0 81 1.33 1 2 2 1.25 2 1.25	t HAdjudication	6 ≠2 80dy ad 2 > A_doc. 2 ×3 fever > 1 ×1 fetgue 3 ×2 dizzy 1 > A_doc.	es ¥1 B tst.sml 1 > A tst.sml 1 > A tst.sml	ody acting dec_Atatami 1	1 achy A_doc_6tstam 3 dizenees	DACHE - hear 2 aches #1 > A_doc_Ztst.ami 1>	aching #1 pain	Export Prev Concept: Last Concept:	22 Tag Test Tag A S Tag Test Tag A S Text Text tex	A A. A. C. Stander A A. B. C. Stander A C. S.	0.0 0.0 0.0	part (Jonn) Retest g bar COVID-19 vac receiving her COVI en her injection a was given har (COVI en her injection a was given har (I) to a set the l to a set the l to a set the l en her in an she dher to a set set for in army instawe intended her	How to use Help -19 vaccine, th how the patien >-19 vaccine, th is vaccine, th is vaccine, the isological second present of the patient is her name out 1 is	xamples of how to take a line of the second	to use the exporte dy having sympt- as already havin bases of the symptonic states of the symptonic for the symptonic constant of the symptonic symptonic of the symptonic of the symptonic symptonic of the symptonic of the symptonic symptonic of the symptonic of the sympton	Most of ruleset files.	Theor 125 0 Se Theor 125 0 Se anxiety/nervous ded," ck to conscious ahe was not fee however the part
THELAE MEDICATION Menication E Summary (cov) Deveload Summary documents: Lags in schems: annotations per docume sentences:	3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	t HAquistation E Annotated Tags © PAN © PYREXIA © PYREXIA © FATTOLE © INLUSEA	6 12 Body Add 2 3.4, dec. 2 2 8.3 feerer 1 3 8.2 dezy, dez, 1 1.4, dec. 1 3.4, dec. 1.2 Addressed 2 8.3 feast 1.4, dec. 3 8.2 dezy, dezy, 1 1.4, dec. 2 8.4 Nasses 1.4	es v1 Bc tst.srd 1 > A 3 foverish tst.srd 1 > A tst.srd 1 > A	Mo I M Mo I M Mo Lating Asso, Atatami Mheaded Soc, Itelami I	DICATION - A int HEA 1 actry ■ 3 A_doc_Bot.wni	E aches ≥1 > A_doc_Batami	aching XI pan A_doc_Bot.ani	Concept text Concept text Co	12 Tog Text Tog & S Tog Text Tog & S Text Text Text Text Text Text Text Text	A A LANCASSON OF A LA	0.0 0.0 0.0	sport (Jonn) Rulesst g ber COVID-19 vacc ceesiving her COVI en her injection a was given her injection en her injection a was given her injection en her injection a the particular of the particular the particular the particular in annu int annu	How to use Help Time, the patien C-19 vaccine, th nd within 1-2 mi otion and within 10 fer inwesilatel istrator came to 1 her name out 1 stated that she tod, asked if ah absprevation/mafe prevass in arm ad body	xamples of how at was alread us patient was in 1-2 minutes, who is in 1-2 minutes y, and within ord and mode to explain the would like typ	dy having sympto dy having sympto sa already havin began feling di m roughly 5 set in roughly 5 set in roughly 5 set in to pack as e an ice pack as e an ice pack as e as on 15-ban2021.	to a series of the series of t	nack dieter 12.5 0 de natiety/nervouses abo natiety/nervous ded." (1) to conscious the was not fee however the pat to yet
e OTHER_AE MEDICATION Annotation Let Statistic tesh Summary (cov) Devertised Summary documents: Inanotations: annotations per docume Isentences:	3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	t HAdjustation EAnnotated Tags PAIN PYREXA PYREXA FATIOUE OCZENESS	 *2 Body ad 2 > A, doc, J 2 *3 fever 1 1 *1 falgue *2 ditzy 1 > A, doc, J *2 ditzy 1 > A, doc, J 2 *1 Nacesa 14 Maniety 	es •1 Be tuttani 3 > A 3 feverish tuttani 1 > A 1 Bg tuttani 1 > A 1 asses +1 nousces	ody athing side statements of the second statement of	I achy II in iteration in the second	E aches File A doc_Backard completely Fil faint	aching El pain A_dou_Backand	Administry: 12 Documents 14 Annotated Te 15 Annot	22 Tag Test Tag & S Tag Test Tag & S Test test 2 doc_span anxiety 85,93 bett 2 doc_span anxiety 85,93 bett 2 doc_span anxiety 85,93 bett 2 doc_span anxiety 85,93 bett 2 doc_span 1 giptheaded 2 fait doc so 1 giptheaded 2 fait doc 1 giptheaded	A ALEXANDER A ALEXANDER D 4 CLANDER D 4 C	a.do c.do b.d. (r.d. b) c.do b.d. (r.d. b) c.do b.d. (r.d. b) c.do n b.do b.do b.do b.do b.do c b.do c b.do c c.do c c.do c	second seco	How to use Help -19 vaccine, th totion and within 1-2 min totion and within iften immediatel istrator came to stated that she tota and body between vation/safe prevents in arm and body all done. for 000	xamples of how tt was alread te patient was instee, she b in 1-2 minute y and within o me (pharmac could and nudg to fait as the te would like try) -boy achieves an 353 NT14282 (T)	to use the exporte dy having sympt. as already havin began feeling di s, abe began feeling di began feeling to be the source of the source of the s, abe began for the source of the source of the source of the source of the source of the ped har shoulder output the source of the s	term and release these one from anxiety/r og syngtung for any setting and the setting and on solify me that a or several these, b of setting the fail. If we found the setting of setting the setting the setting the setting of setting the setting the setting the setting the setting of setting the setting the setting the setting the setting of setting the setting the setting the setting the setting the setting of setting the setting the setting the setting the setting the setting of setting the s	nack dieter 12.5 Ø Se eervousses ab anziety/servou ded," ightheaded." ck to contciou she was not fe however the par
e OTHER_AE MEDICATION Annotation Let Statistic tesh Summary (cov) Devertised Summary documents: Inanotations: annotations per docume Isentences:	3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	t 64 Adjudication E Annotated Tags © PRIX © PRIEXA © RATIOUE © DIZZNESS © NAUSEA © ONRELAE	 C 2 Body acd C 2 A 3 fever A 4 3 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5	es #1 Bc tattani 3 > A 3 feverish tattani 1 > A 5 feverish 1 austos 1 nervousness 1 darnes	ody aching dige_dist_dist_dist_dist_dist_dist_dist_dist	1 Adv × Internet	2) adres (21) 2) adres (21) 3 Agregations (1) 3 Agregations (1) 4	lading EX pan A,An,Littered	Constant Consta	22 Tag Text Top & St Tag Text Top & St Text thew thew to doc.span anticity 84.9 anticity	A ALEXANDER A ALEXANDER D 4 CLANDER D 4 C	CONTRACTOR OF A CONTRACTO	second seco	How to use Help The the patient >-19 vaccing, th the the patient provide the the the the the the the testing and the the the testing the the the the testing the the the the provide the the the provide the the the the testing the the the testing the the testing the testing the testing the testing the testing the testing the testing	xamples of how tt was alread te patient was notes, she k notes, she	to use the exporte dy having sympti- tics as already havin began facilised di in roughly 5 are clist on daryh tr g an ice yet and origh hat was gol origh hat	bed ad ruleset Bles. and from anxiety//r og symptoms from a maxiety/ durry and b on anxiety/ any any and on anxiety methant of a sectory to fails. Maker, and step ad water, and step	nervousness ab antiety/nervou ded. Uptotended. Extra particular pred out to ger lot number and
orretes_AE MEDICATION which are a set of statistic set Summary (only) bowelsad which are a set of set	3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	t HAquistation E Annotated Tags © PAN © PYREXIA © PYREXIA © FATTOLE © INLUSEA	 C 2 Body acd C 2 A 3 fever A 4 3 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5	es #1 Bc tattani 3 > A 3 feverish tattani 1 > A 5 feverish 1 austos 1 nervousness 1 darnes	ody aching dige_dist_dist_dist_dist_dist_dist_dist_dist	1 Adv × Internet	2) adres (21) 2) adres (21) 3 Agregations (1) 3 Agregations (1) 4	aching El pain A_dou_Backand	Adammary: Documents	The Tend Tay & S The Tend Tay & S Tend Tay & S	A	(1) (second seco	How 50 use Help Help Time, the patient >-19 warcine, thi within 1-2 all within 1-2 all	<pre>xamples of how st was alread spatient wa spatient wa spatient wa spatient wa spatient wa spatient was spatient was spatient was spatient was spatient was spatient was spatient sp</pre>	dy having sympt dy having sympt has already havin began feeling di expansion of the sympt set of the sympt set of the sympt set of the sympt set of the sympt set of the sympt set of the sympt set of the sympt s	te drukestfile. om from anisty// og symptom from anisty// og symptom from anisty// og symptom from anisty try and lightback of the try and lightback of the order back of the or	finer 12.5 0 Se nervousness ab nakiet/varvou ded." 10pthaeded." k to concious abe was not fee however the pa pped out to ge lot number and
THELAE MEDICATION Menication E Summary (cov) Deveload Summary documents: Lags in schems: annotations per docume sentences:	3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	t 64 Adjudication E Annotated Tags © PRIX © PRIEXA © RATIOUE © DIZZNESS © NAUSEA © ONRELAE	6 ¥2 Body and 2 A.disc, J. A.disc, J. 1 ¥3 Body and 3 ¥4 Body and 1 ¥3 Body and 2 ¥3 Body and 3 ¥4 Body and 1 ¥3 Body and 2 ¥3 Body and 2 ¥4 Body and 1 ¥4 Body and 21 ¥4 Body and	es v1 Be Ltt.smi 3 > A 3 feverish tst.smi 1 > A 5 feverish v1 Be v1 se v1 se v	ddy aching ddy aching ddsc_Atat.emi 3 dsc_Late.emi 3 8 ¥1 not freeling ¥1 datydration ¥3 Dappefer ¥	A day a strange of the second	2 adm 12 adm 12 adm	lading EX pan A,An,Littered	externary in Document Te in Accounter Te in Accou	The Tend Tay & S The Tend Tay & S Tend Tay & S	Аландантон Аландантон Останизация	(1) (spacy (gont) Rulest The second	How to use How to use How to use the patient to see the patien	xamples of how it was alread up attent was in 1-2 minutes in 1-2 minutes	to use the exporte dy having sympti- tas already havin began failed the began fa	been and the second sec	Ther 12.5 0 Se Ther 12.5 0 Se anticty/nervous deal d
	3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	t 64 Adjudication E Annotated Tags © PRIX © PRIEXA © RATIOUE © DIZZNESS © NAUSEA © ONRELAE	6 #2 Body and 2 3 Active 2 2 2 #3 fewer 3 3 12 Active 1 3 4 3 12 Active 1 3 4	es x1 00 ititized 10 A 3 feveriak titized 10 A 10 A 11 A 11 A 11 A 11 A 11 A 11 A	sdy aching s disc_stations 1 disc_stations 1 disc_stations 1 s1 distydration 13 distydration 1 3 discyther 1 13 distydration 1	ladry Advaction - A Intrinsic Advaction - A	2 artes 22 artes 22 artes 24 a	Lating XI per Autor_dataset	Community Decommunity	Constraints of the second			Backy (down) Relevant Relevant The CVID-19 vac record vis many record vis m	How to use inter, the patient >-19 vaccine, the descent patient -19 vaccine, the set of the set of the set of the set of the present patient set of the set of the present patient 1 age (age) 19, 1 ag	xamples of how at was alread se patient was notes, she b not contained to the se optimum of the se would like typ	dy having sympt- dy having sympt- sa already havin began feeling di sa da began d	Mod ad ruleset files. The second seco	Ther 125 \$ \$ \$ Ther 125 \$ \$ Ther 125 \$ \$ Ther 125 \$
orhele, AE MEDICATION Anotation ME Statist Summary (only) Downbad ummary documents: anotations per docume sentences:	3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	t 64 Adjudication E Annotated Tags © PRIX © PRIEXA © RATIOUE © DIZZNESS © NAUSEA © ONRELAE	6 12 Body and 2 3 State 1 3 State 3 State State 1 3 Adsc, 1 2 State State 3 State State 1 Adsc, 1 Adsc, 1 2 State State 3 State State 4 State State 5 State State 5 State State 5 State State 5 State State	es vi B tetturi 3 A 3 foverish beturi 3 A tetturi 1 A tetturi 1 A tetturi 1 A 1 A 1 A 1 A 1 A 1 A 1 A 1 A	ady acting disc_Atstant 3 disc_Atstant 3 disc_Atstant 1 disc_Atstant 1 di	Adv A	a sense SI ante > American SI ante > American SI ante dennes SI ante SI antendete SI ante SI sensember SI ante SI sensember SI ante SI an	ang XI on Adridation XI forenitam to XI wing XI sha XI forent hypothesis	Constraints Constrain	Constraints of the second	A Cardination of the Construction of the Const	The second secon	control sector (sector) control sector) control s	How to use inte, the patient >-19 vaccing, thi didiking 1-2 million present the second second statator case to statator case to statator case to statator case to statator case to statator in arm and book presents in arm and statator case to statator c	xamples of how at was alread int was alread interest was interested by and interested by and interested by and adress on 152 NTI(1282 (PT) VUD-19 immer units; unknot ation, on 95 on of the fe the duration of	<pre>to use the exporte dy having sympt. as already having search as already having bean feeling in roughly 3 weights into a dary to interime the search as a search as a search as a search as a search as a search as a search as a search as a search as a search as a search as a search as a search as a search as a search as a search as a search as a search as a search as a sear</pre>	And of ruleast files. The second second second second second second second seco	Theor 12.5 0 S Theor 12.5 0 S nervousness ak anxiety/nervos ded." to conclose the was not for the

Figure 2. Screenshots of the four tabs in MedTator, including (A) the annotation tab showing multiple documents, visualized tags, and annotated tag list with details; (B) the adjudication tab showing the inter-annotator agreement in different levels with detailed tag context from two annotators; (C) the statistics tab showing the overall status, and detailed annotated tags of each concept; and (D) the export tab for converting the annotations in different formats.

User interface design

As shown in Figure 2, for the four core tasks identified in the annotation process, we designed four tabs to support the activities needed for each task. All four tabs follow the same design style and layout, which splits the interface into three parts, the top bar, ribbon menu, and workspace panel, from top to bottom, respectively (large figures and online demo can be found at our GitHub repo <u>https://github.com/OHNLP/MedTator</u>). Users can switch different tabs in the top bar, then the ribbon menu will be updated to show the related functionalities, and the workspace panel will be updated to show specific panels for the corresponding task. The details of each tab are as follows:

Annotation tab. This tab allows the user to annotate multiple documents according to a schema through four coordinated views (Figure 2A), including the file list view, tagging view, concept list view, and tag list view. The tagging view is the main interface for annotation, which allows users to add or delete tags on the selected document. At present, MedTator supports three types of text annotation in this tab: 1) span-based entity annotation, 2) entity-based relation annotation, and 3) document-level annotation. To reduce the repetitive workload, MedTator can optionally show hints on potential words 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 displayed on the file list view and concept list view for tracking annotation progress.

Adjudication tab. This tab allows users to compare annotations of two annotators to create a gold-standard corpus (Figure 2B). F1-score is used for IAA calculation to assess the reliability of annotations¹⁸, and the detailed results are visualized at different levels for interactive exploration, which can be downloaded for further analysis. In addition, users could accept or reject each annotator's annotation to generate the gold-standard corpus for further tasks.

Statistics tab. The detailed information of the real-time statistics on the annotated tags is shown in this tab, allowing users to analyze both the overall status and the annotated tags from concept and document aspects (Figure 2C). Users can also trace back to the source document of each annotated tag in the annotation tab through hyperlinks.

Export tab. This tab allows users to convert the annotations to different formats for downstream tasks (Figure 2D). For example, the annotated tags and context sentences can be converted to a ruleset to build a rule-based system based on open-source tools (e.g., MedTagger¹⁶ and spaCy), an inside-outside-beginning format file for training models for named-entity recognition, a tabular format file, or a BioC format¹⁵ file for other NLP tasks.

Lessons learnt

During the process of design and development of MedTator, we have learnt several lessons. Those lessons would help the design and development of informatics solutions.

Lesson 1: Apply a user-centered design approach to meet the needs of various users

Our previous studies analyzed the activities in the annotation process and summarized the best practices^{11,19}. To go through those activities and build a high-quality corpus, considerable time and expertise of the annotators are necessary. However, annotators are not the only stakeholder in the whole annotation process. In addition to annotators, we found that researchers and administrators are also involved in the annotation process with different goals and needs. Therefore, rather than developing a huge "all-in-one" platform to cover all activities for every stakeholder, developing a lightweight tool that focuses on the core activities is more feasible.

As a result, MedTator was designed to address the four core activities: text annotation, statistics, adjudication, and annotation export. For these four activities, a multidiscipline team with various skillsets is usually needed to meet the requirements of each activity, and their user requirements for the tool are also different.

The power user: researchers

Researchers are deeply involved in each task of entire corpus development and thus have a strong demand for an annotation tool to accelerate the whole process. They have advanced skills in informatics and rich experience in applying NLP in clinical studies. They make the decisions for each task to meet the research needs, as well as what tool to use for annotation. As a result, researchers of our team were the initial users since the early beginning of the development. We interviewed them and other researchers from the community to understand the needs in corpus annotation and the gap between those needs and existing tools as described in the System Design section.

In addition to the functional requirements which have been addressed in the core features, the following findings are also considered to inspire the tool design. First, project needs vary. We find that most of the needs in annotation projects are similar, but each project has a few special needs to meet its own research goals, such as ontology mapping or normalization. Those special needs can be addressed by developing new features, but since the goal of MedTator is a lightweight annotation tool, we need to evaluate whether the new feature can be applied to other projects and how it affects the user experience. Secondly, the resource is limited. Resources, such as computation infrastructure, budget, workforce, time, etc., seem to be always highly demanded but limited for every project. Therefore, when designing MedTator, we aim to offer a lightweight solution built upon public services (e.g., public web server and open-source libraries, etc.) to save the cost for both the team and users.

The end users: annotators

Annotators are the direct end user of an annotation tool. Unlike those annotation tasks of public corpora such as Twitter posts and news reports, the growing length of the clinical documents and the complexity of the contextual knowledge pose challenges to the annotators. Therefore, clinical document annotation heavily depends on experienced annotators with rich expertise in specific domains. They spend considerable efforts on the annotation tasks with the annotation tool; thus, any improvement on the user interface can thus have a significant impact on their experience and work efficiency.

According to the feedback from the annotators of our team and the community, we found that ease of use seems to be the only need for the annotation tool. Especially in the clinical corpus annotation, the annotators with clinical domain knowledge usually are not trained as technical experts. They stated that the annotation tasks would be challenging and very stressful if the tool was hard to use. Therefore, it is necessary to put them in the design cycles and take their comments into the design considerations for an easy-to-use user interface.

However, the ease of use is not just about the user interface or the usability of the annotation tool. In fact, it involves many aspects throughout the annotation process. For example, when new annotators join a project, they may not have any experience in the annotation tool and the corpus development. The first impression of the ease of use comes from the learning materials, which we provided in the training sessions, for helping them learn the basic operations. We prepared several types of user documentation for different purposes, such as a quick start to get a basic understanding of what annotation is, tutorial, and technical specifications. During the annotation, annotators may encounter various issues, such as program bugs, document management, version upgrade, questions about the tool, etc. These issues need to be quickly addressed and explained to annotators to help them. Therefore, we held training sessions, online forum, and email groups to collect feedback and discuss solutions.

The invisible users: system administrators

Because of the critical concerns of the sensitive PHI data in clinical research projects, system administrators are often involved in handling the management of computation environments and data access permissions. During the design and development cycles, we received comments from domain experts and community users about the server requirements and tool installation with the concerns of data privacy, security, and server permissions. Those comments indicated that the needs of system administrators should also be taken into consideration for tool development, especially in the clinical setting.

Unlike researchers and annotators, administrators usually won't directly use the functional features of the annotation tool, such as text annotation and adjudication, but they need to figure out how to install and maintain the tool for researchers and annotators. They thus need to have the specifications of the technical requirements, such as installation manual, runtime requirements, database access, and tool configurations. Moreover, they need to address the data security concerns and the system permissions requirements (e.g., machine, network, software installation, data access, etc.) with the IT department. Therefore, we adopted the serverless architecture to design MedTator to ease the burdens of tool installation on administrators and other users.

Lesson 2: Non-functional requirements are critical for adoption

As described in the previous section, in addition to the functional requirements, non-functional requirements, such as easy to install, easy to use, documentation, and data security, are frequently mentioned in users' comments. We found that those non-functional requirements can form the basis of user experience and the evaluation of the tool, which are critical for the tool adoption decision. The main non-functional requirements we learned are summarized as follows:

Ease of installation as the first positive impression

The comments from the domain experts and administrators indicate that the first barrier to using an annotation tool is often its installation. We investigated existing tools and found that the tool installation also depends on the tool architecture, which determines how the frontend (i.e., the user interface that collects user inputs) and the backend (i.e., the core modules that handle the data processing and calculation) communicate and how the data are saved. Therefore, we plan to address this ease of installation from the perspective of system architecture.

Existing tools' architectures can be divided into standalone tools and web-based tools. For those standalone tools, both the user interface and the core modules run on the user's machine (Figure 3A), and users need to install a runtime (e.g., Java or Python) to use the tool. However, the versions of the runtime and libraries are often confusing and sometimes even conflict with the existing environment in the user's machine. Managing multiple versions of the runtimes and avoiding conflict of those runtimes are not easy, even for experienced users. It can be more difficult for non-technical users to solve those installation issues on their own.



Figure 3. The architecture differences among MedTator and existing annotation tools. (A) The architecture of standalone annotation tool that requires local runtime environment and won't send out data. (B) The architecture of typical web-based annotation tool that requires server runtime environment, and the data need to be sent out to server for processing and storage. (C) The serverless architecture of MedTator that doesn't require any installation of runtime and won't send out any annotation data to outside just like a standalone tool.

For those web-based tools, the user interface (i.e., webpages) runs in the web browser, while the core modules run on the server (Figure 3B). It is much easier for the annotators as they do not need to install the runtime, but it will be more complex for administrators due to the server-side installation.

Therefore, to address the installation issue, we adopted a serverless architecture to design MedTator as introduced in the System Design section. As shown in Figure 3C, both the frontend and the backend run in the web browser; therefore, they do not require any extra installation of a runtime environment on a local machine or server. Consequently, the installation of MedTator is as simple as unzipping a downloaded zip package without any further steps. Users can just double-click the HTML file or visit the public web page to start annotation.

Usability

The usability of the annotation tool directly impacts on the user experience, which may affect the workload and efficiency of annotators. Therefore, the usability of MedTator is always considered when we are developing any feature and making design decisions. We adopted core human-computer interaction (HCI) design principles to guide the tool design according to the comments on existing annotation tools by our domain experts and experienced annotators. The core HCI design principles are as follows:

Reducing cognitive load: the cognitive load imposed by the user interface is the amount of mental resources that is required to perform a certain task with the user interface²⁰. And the cognitive overload may be more likely to happen during annotation of the clinical documents because of the large number of clinical narratives and complex annotation guidelines. It is thus challenging for annotators to handle the annotation task comfortably without an easy-to-use annotation tool. We aim to apply this design principle in designing interfaces that decrease extraneous cognitive load so annotator's efforts can be devoted to the actual annotation tasks rather than other distracting tasks. For example, we decided to only include the core four features (e.g., text annotation, statistics, export, and adjudication) related to corpus annotation in MedTator to reduce the extra burdens on users of learning complex features. And for each feature included, we designed separated tabs to show the necessary information related to the annotation task with the same design language. Each tab only provides those necessary functionalities on one single webpage.

Low physical effort: users need to conduct several interaction steps to accomplish a task, and the tool should not involve making "unnecessary" interventions by the user. To minimize the number of steps in each task, we studied the existing workflow and optimized the interactive process for each task. Then, we use this design principle to guide the interactive design of specific operations. And sometimes, the first principle can also be applied together with the second principle to reduce the unnecessary interactions by reducing unnecessary cognitive load. For instance, new tags can be added by clicking shortcut key which is displayed with the tag name; when hovering any annotated tag, the detailed attributes will be displayed next to the hovered tag; the statistical results of the annotated tags are updated in real-time, so that users can track the annotation progress easily.

Flexibility in use: the tool should provide various functionalities to support different project preferences. Therefore, we designed and developed many features, including text span annotation at levels of characters, words, sentences, and documents; relationship annotation between multiple text spans for building semantic relationships; tag visualization; multiple document annotation; and customizable annotation schema for various tasks, etc. This design principle sometimes conflicts with the first principle, so we kept both principles in mind and tried to balance both of them when making design decisions. For example, according to the feedback on our prototypes, some users comment that the automatically generated annotation tags are useful because it helps the identification of all potential tags, while some users comment that these tags may lead to bias and confusion. From the perspective of the third principle, adding the automatic tag feature could be useful, but it also may sometimes confuse users from the perspective of the first principle. To balance the needs for this feature, we designed a group of radio buttons, which can turn on or off this feature in one click and update the tag visualization in real-time.

Documentation

As described in lesson 1, documentation is highly demanded by all users, but the creation and maintenance of documentation are often challenging for the team due to the massive workload of writing. To partially ease the workload, we summarized the comments from domain experts and our experience as follows.

First, identify the types of documentation and prioritize the most needed documents. A common consensus is that the documentation has great practical benefits in software development²¹, and there are different types of documentation that have different usage at different stage²². For example, technical documentation such as code comments, design principles, and API references, are only helpful for developers in programming and maintenance. In contrast, user documentation can be helpful to a broader range of readers. With the limited resources, focusing on the user

documentation rather than the technical documentation can be a feasible way to reduce the burden of writing. In our case, we paid more attention to the quick start, user manual, and annotation task-related documents, whose target readers are annotators, researchers, and system administrators.

Secondly, user documentation should be designed to accommodate different users for different needs. For example, when we introduce MedTator to other researchers and annotators, they usually ask for some "learning materials" for different learning goals. Some users want to check the features of interest, such as what annotation schemas are supported and how to use the annotation file for their downstream tasks, meanwhile some users prefer a short document that goes through the basic annotation process. In response, we prepared different types of user documentation: a quick start with a step-by-step tutorial for showing the basic workflow and necessary features for corpus development; a well-structured instruction manual that focuses on explaining the detailed usage of all functionalities; and several annotated datasets for demonstrating different annotation tasks.

Lastly, keep the documentation up to date. When features and functionalities are introduced or updated in new releases, the affected sections in the document should be updated to avoid confusing users. We are using the GitHub wiki to organize all user documentation, and everyone in our team can contribute to the content.

Lesson 3: Community participation makes it thrive

During the development, we introduced MedTator to the American Medical Informatics Association (AMIA), Open Health Natural Language Processing (OHNLP), and National COVID Cohort Collaborative (N3C) communities through workshops, online meetings, and training sessions. The valuable feedback from various groups convinced us that community participation is the key to making an open-source tool thrive.

First, MedTator benefits through, and because of, community collaboration. Software development involves research, design, thinking, testing documentation as well as communication with end-users. As an open-source project initiated by a small team, our limited resources have to be allocated to the most needed development tasks, such as architecture design and code-writing of main features. Then, the rest tasks, such as testing, training, and documentation, are supported by our collaborators from the community. They tested MedTator in their corpora and reported issues, which greatly improved the quality of MedTator.

Secondly, there are many different user groups in communities, such as students, computer science researchers, clinicians, informaticians, software developers, NLP experts from the industry, etc. They commented MedTator from many aspects, such as user experience, user interface design, visualization and interactivity, system architecture, annotation schema design, annotation workflow, deployment, platform dependency, integration with other tools, data format, metadata extension, and annotation file management, etc. Although our goals and limited resources do not allow us to include all these valuable suggestions in the design and development plan, their insights help us dive into the practical needs of various users and inspire future development.

Discussion

Comparison with other annotation tools

Many text annotation tools have been developed to help users improve their annotation efficiency by providing many advanced features, such as pre-annotation and online collaboration, which are not available in MedTator. However, in terms of prerequisites for installation, MedTator shows a significant advantage.

Tool	Source Code	Prerequisites for server installation on server
BioQRator	https://github.com/dongseop/bioqrator	MacOS or Linux; Ruby, Ruby on Rails, MySQL
brat	https://github.com/nlplab/brat	Linux; Apache 2, GeniaSS, Python 2.5
Catma	https://github.com/forTEXT/catma	Java, Maven, Jetty
Djangology	https://sourceforge.net/projects/djangology/	Python, Django, database server
ezTag	https://github.com/ncbi-nlp/ezTag	Ruby, Ruby on Rails, MySQL
FLAT	https://github.com/proycon/flat	Apache / Nginx, Python, foliadocserve, pynlpl, Django
MAT	http://mat-annotation.sourceforge.net/	Python, Java
PDFAnno	https://github.com/paperai/pdfanno	Node.js
TextAE	https://github.com/pubannotation/textae	Node.js
WAT-SL	https://github.com/webis-de/wat	Java, Docker
WebAnno	https://webanno.github.io/webanno/	Java, Apache Tomcat, MySQL

Table 1. Open-source annotation tools and their prerequisites for installation on a server.

* Commercial tools (prodigy, tagtog, LightTag) and close-source tool (MyMiner) are excluded due to source code unavailability

In the latest review of annotation tools ³, only 15 out of 78 tools were selected for further evaluation, as other tools were excluded because they were either not installable or not workable. We further searched and summarized their prerequisites for installation on the server of those tools. As listed in Table 1, before the installation of those annotation tools, some specific software and libraries must be prepared, which usually requires:

- A stable server. A dedicated machine with a stable network connection is usually needed to host the serverside program, such as a physical computer or a virtual machine.
- **Technical expertise**. The installation, configuration, and maintenance of the server environment often require a series of technical expertise, such as the management of Linux server, network and firewall, virtual environment of programming language runtime, and databases administration.
- **Permissions**. Unlike annotating the public corpora (e.g., Wikipedia and PubMed metadata), the data privacy and security of clinical documents are critical concerns for clinical institutions. Administrators need to communicate with the IT department to get permissions such as software installation and network access.

By adopting the serverless architecture, MedTator can save the cost on the above items for researchers. The whole program is packaged into one HTML file, which can run locally or be served as a webpage through public web servers. Therefore, users do not need to configure any server and request extra permissions for using MedTator.

Usage scenarios

When a specific annotation tool is selected for a project, it usually would not be replaced during the annotation process to ensure consistent data quality and annotation experience. Therefore, the decision on the annotation tool should be made carefully based on project needs and tool features. We have been using MedTator in several annotation projects and we proposed a guideline of annotation best practice that aims to streamline the whole clinical document annotation process ¹⁹. According to our experience and comments from community users, we think the features provided by MedTator can help to meet the needs in the following scenarios:

- **Restricted environment**. Due to privacy and data security concerns, if the annotation has to be conducted in a specific local machine or secured network with limited permissions of changing environment, MedTator can provide a lightweight annotation solution and address those concerns.
- **Prototype validation**. If researchers plan to implement an NLP system based on an annotated corpus, MedTator provides a quick start to validate the workflow of the prototype without extra costs.
- Education. If educators plan to teach students about clinical NLP, MedTator can be used as it does not require any user registration, installation, or server/client configuration. All the inputs are just plain text files that are easy to manage and distribute. And the tutorial and manuals are also provided for users to follow.

Limitations

While having the above advantages, MedTator still has some shortcomings: First, unlike Java and Python, there are fewer JavaScript libraries for scientific purposes, so advanced techniques such as machine learning models, knowledge graphs, and NLP toolkits, are not able to be integrated easily to provide more features such as active learning ⁶. Secondly, due to the use of HTML5 APIs in the web browser, MedTator only supports specific modern web browsers. Lastly, the current user interface design is highly rated by our users, and it works well for annotating hundreds of documents, but a formal usability test and load test need to be conducted to evaluate the performance.

Conclusion

To facilitate the corpus development in clinical research, we developed a serverless web-based annotation tool, MedTator, to provide an intuitive and interactive user interface that focuses on the core annotation tasks. We summarized the lessons learned from designing, developing, and maintaining MedTator, which will be of value to the research community for future open-source tool development.

Acknowledgment

The authors thank Donna Ihrke 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.

References

- Wen A, Fu S, Moon S, El Wazir M, Rosenbaum A, Kaggal VC, et al. Desiderata for delivering NLP to accelerate healthcare AI advancement and a Mayo Clinic NLP-as-a-service implementation. Npj Digit Med. 2019 Dec 17;2(1):1–7.
- 2. Dreisbach C, Koleck TA, Bourne PE, Bakken S. A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data. Int J Med Inf. 2019 May;125:37.
- 3. Neves M, Ševa J. An extensive review of tools for manual annotation of documents. Brief Bioinform. 2021 Jan 1;22(1):146–63.
- 4. Stenetorp P, Pyysalo S, Topić G, Ohta T, Ananiadou S, Tsujii J. brat: a Web-based Tool for NLP-Assisted Text Annotation. In: Proceedings of the EACL 2012: Demonstrations. Avignon, France: ACL; 2012. p. 102–7.
- 5. Kwon D, Kim S, Wei CH, Leaman R, Lu Z. ezTag: tagging biomedical concepts via interactive learning. Nucleic Acids Res. 2018 Jul 2;46(W1):W523–9.
- 6. Klie JC, Bugert M, Boullosa B, Eckart de Castilho R, Gurevych I. The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. In: Proceedings of the 27th COLING : System Demonstrations. Santa Fe, New Mexico; 2018. p. 5–9.
- Islamaj R, Kwon D, Kim S, Lu Z. TeamTat: a collaborative text annotation tool. Nucleic Acids Res. 2020 Jul 2;48(W1):W5–11.
- 8. Salgado D, Krallinger M, Depaule M, Drula E, Tendulkar AV, Leitner F, et al. MyMiner: a web application for computer-assisted biocuration and text annotation. Bioinformatics. 2012 Sep 1;28(17):2285–7.
- 9. Maxim Tkachenko, Mikhail Malyuk, Nikita Shevchenko, Andrey Holmanyuk, Nikolai Liubimov. Label Studio [Internet]. 2021 [cited 2022 Jan 31]. Available from: https://github.com/heartexlabs/label-studio
- 10. He H, Fu S, Wang L, Liu S, Wen A, Liu H. MedTator: a serverless annotation tool for corpus development. Bioinformatics. 2022 Jan 4;btab880.
- Carlson LA, Jeffery MM, Fu S, He H, McCoy RG, Wang Y, et al. Characterizing Chronic Pain Episodes in Clinical Text at Two Health Care Systems: Comprehensive Annotation and Corpus Analysis. JMIR Med Inform. 2020 Nov 16;8(11):e18659.
- 12. Shen F, Wang L, Liu H. Phenotypic Analysis of Clinical Narratives Using Human Phenotype Ontology. Stud Health Technol Inform. 2017;245:581–5.
- 13. Peterson KJ, Jiang G, Liu H. A corpus-driven standardization framework for encoding clinical problems with HL7 FHIR. J Biomed Inform. 2020 Oct 1;110:103541.
- 14. Fu S, Leung LY, Wang Y, Raulli AO, Kallmes DF, Kinsman KA, et al. Natural Language Processing for the Identification of Silent Brain Infarcts From Neuroimaging Reports. JMIR Med Inform. 2019;7(2):e12109.
- 15. Comeau DC, Islamaj Doğan R, Ciccarese P, Cohen KB, Krallinger M, Leitner F, et al. BioC: a minimalist approach to interoperability for biomedical text processing. Database J Biol Databases Curation. 2013;2013:bat064.
- Liu H, Bielinski SJ, Sohn S, Murphy S, Wagholikar KB, Jonnalagadda SR, et al. An Information Extraction Framework for Cohort Identification Using Electronic Health Records. AMIA Summits Transl Sci Proc. 2013 18;2013:149–53.
- 17. Castro P, Ishakian V, Muthusamy V, Slominski A. The rise of serverless computing. Commun ACM. 2019 Nov 21;62(12):44–54.
- 18. Hripcsak G, Rothschild AS. Agreement, the F-Measure, and Reliability in Information Retrieval. J Am Med Inform Assoc JAMIA. 2005;12(3):296–8.
- Fu S, Leung LY, Raulli AO, Kallmes DF, Kinsman KA, Nelson KB, et al. Assessment of the impact of EHR heterogeneity for clinical research through a case study of silent brain infarction. BMC Med Inform Decis Mak. 2020 Mar 30;20(1):60.
- Oviatt S. Human-centered design meets cognitive load theory: designing interfaces that help people think. In: Proceedings of the 14th ACM international conference on Multimedia. New York, NY, USA: Association for Computing Machinery; 2006. p. 871–80. (MM '06).
- Aghajani E, Nagy C, Vega-Márquez OL, Linares-Vásquez M, Moreno L, Bavota G, et al. Software Documentation Issues Unveiled. In: 2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE). 2019. p. 1199–210.
- Aghajani E, Nagy C, Linares-Vásquez M, Moreno L, Bavota G, Lanza M, et al. Software Documentation: The Practitioners' Perspective. In: 2020 IEEE/ACM 42nd International Conference on Software Engineering (ICSE). 2020. p. 590–601.