

NLP文本标注工具对比分析

钟旭红, 刘 伟

北京交通大学语言与传播学院, 北京

收稿日期: 2025年6月1日; 录用日期: 2025年7月1日; 发布日期: 2025年7月16日

摘 要

大语言模型是人工智能算法在自然语言处理领域的具体应用。数据标注作为训练大语言模型的关键环节, 其质量直接决定大语言模型的效能。在数据标注的完整体系中, 文本标注占据核心模块的地位。文本标注旨在将自然语言环境中广泛存在的非结构化文本, 按照既定的标注规范和语义逻辑, 处理为结构化的数据形式。这一过程所产生的结构化数据, 是机器学习算法有效运行以及深度自然语言处理任务高效开展的关键支撑要素。传统人工标注模式存在效率低下、成本高昂及质量参差不齐等固有缺陷。本文基于系统文献分析法, 选取13个具有代表性的文本标注工具, 在技术架构、数据处理能力、功能三个维度进行对比研究, 揭示了现有文本标注工具在可用性、可配置性、标注效率、预标注等方面的优势特征与技术瓶颈。本文相关发现有望为下一代文本标注工具的构建提供部分理论依据, 并为其技术发展提供一定的思路借鉴, 对自然语言处理领域标注范式的探索提供新的方法参考。

关键词

文本标注, 语料库构建, 大数据模型, 自然语言处理

A Comparative Analysis of NLP Text Annotation Tools

Xuhong Zhong, Wei Liu

School of Languages and Communication Studies, Beijing Jiaotong University, Beijing

Received: Jun. 1st, 2025; accepted: Jul. 1st, 2025; published: Jul. 16th, 2025

Abstract

Large language models (LLMs) represent the specific application of artificial intelligence algorithms in the field of natural language processing (NLP). As a critical component in training LLMs, the quality of data annotation directly determines the effectiveness of these models. Within the complete framework of data annotation, text annotation occupies a core position. Text annotation aims to process the unstructured text widely present in natural language environments into structured

data forms according to predefined annotation specifications and semantic logic. The structured data generated through this process serves as a key supporting element for the effective operation of machine learning algorithms and the efficient execution of deep natural language processing tasks. Traditional manual annotation models suffer from inherent drawbacks such as low efficiency, high costs, and inconsistent quality. This paper employs a systematic literature analysis approach to select 13 representative text annotation tools for comparative study across three dimensions: technical architecture, data processing capabilities, and functional features. The study reveals the advantageous characteristics and technical bottlenecks of existing text annotation tools in aspects such as usability, configurability, annotation efficiency, and pre-annotation functionality. The findings of this research are expected to provide partial theoretical foundations for the development of next-generation text annotation tools, offer innovative insights for their technological advancement, and provide new methodological references for exploring annotation paradigms in the field of natural language processing.

Keywords

Text Annotation, Corpus Construction, Big Data Models, Natural Language Processing (NLP)

Copyright © 2025 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

近年来,大语言模型已成为计算机科学领域备受瞩目的研究方向。此类模型对数据的依赖程度极高,数据不仅是其训练与优化的基石,更是决定其应用效能的关键因素。然而,互联网上大量的低质量数据的存在减缓了大语言模型的训练进程,因此,数据标注作为构建大语言模型必不可缺的环节,其重要性日益凸显。数据标注是指对原始数据(如文本、图片、音频、视频等)添加标签的过程,有助于为机器学习和人工智能模型提供高质量的训练数据。标注后的数据能够有效辅助算法理解数据特征,从而提高模型的准确性和性能。

根据标注者类型,数据标注可划分为人工标注、机器标注和半自动标注。当前,为了保障优质的语料库数据,人工标注仍是企业的首选。尽管传统人工标注在标注质量上具有显著优势,但其高昂的标注成本与低下的标注效率成为了制约其发展的瓶颈。随着大语言模型的兴起,一大批数据标注工具应运而生,极大地缓解了人工标注的效率问题。但工具的可用性对标注过程有着直接的影响,因此,在众多可选工具中选择最合适的标注工具是一项极具挑战性的任务。由于可用的工具数量众多且新工具不断涌现,系统性地分析并归纳代表性标注工具的优缺点具有重要的实践价值。

从标注对象的角度出发,数据标注可进一步细分为文本标注、语音标注与图像标注[1]。鉴于文本数据类型的普遍性以及标注任务的多样性,本研究着眼于文本标注工具。本文选取了三个对比维度(即技术、数据、功能),对13个代表性的文本标注工具进行对比分析,旨在为标注人员在标注过程中选择最优工具提供决策依据。

2. 文本标注工具选取与评估体系

2.1. 选取标准

本文所选取的文本标注工具均源自谷歌学术、知网等学术平台论文中提及的工具。鉴于学术平台上提及的文本标注工具数量繁多,本文设定了若干筛选标准,将符合这些标准的工具纳入对比列表。经过严格筛选后,最终保留了13个文本标注工具,包括BRAT、Doccano、Djangology、GATE、LightTag、

MedTAG、Prodigy、POTATO、tagtog、TeamTat、WAT-SL、WebAnno、YEDDA。

上述工具满足的筛选标准为:

(1) 工具应具备易获取性,无论是可直接在线使用,还是可安装使用,且无需联系开发者。这一标准排除了部分无法找到或者已停止开发的工具。

(2) 工具应具备一定的知名度,此标准可确保工具的安全性和可靠性,并便于获取足够的安装和使用资料。

(3) 工具应具备可安装性,此要求仅针对需通过安装才能进行标注的工具。安装文件应足够全面。据此,部分安装过程过于繁琐或者经常安装失败的工具被排除在外。

(4) 工具应具备可操作性,必须有充分证据表明工具便于使用,而不会因为频繁的操作失误导致标注效率下降。

2.2. 对比维度

2.2.1. 技术维度

该维度聚焦于工具本体特性分析,包含四项标准:

(1) 平台兼容性

据 Musabeyezu Fortunee (2019) [2], GATE、LightTag、WAT-SL、YEDDA 四款工具在平台兼容性维度存在显著局限性,其 API 接口尚未实现对主流众包服务平台的兼容性支持。这种架构缺陷可能导致其在分布式标注场景下存在数据交换障碍。

(2) 工具架构

据 Mariana Neves 和 Jurica Seva (2021) [3], 基于 Web 技术架构的工具(如 BRAT、Doccano)展现出显著优势。此类无客户端部署模式有效降低了用户认知负荷,避免了安装额外软件,使标注人员可专注于标注任务,提高了标注效率。

(3) 安装简便性

据 Mariana Neves 和 Jurica Seva (2021) [3], Djangology、POTATO 和 WebAnno 三款工具的安装流程复杂度较高,安装时间多在两小时以上,其可能引入潜在的时间成本风险。

(4) 收费标准

据官方资料表明[4], LightTag 采用学术研究豁免收费模式,但商业应用需承担未披露的许可费用; Prodigy 实施年度订阅制商业授权模式,收费很高[2]; Tagtog 提供基础功能模块的有限制免费访问权限,深度功能扩展需额外支付运营成本[5]。

图 1 展示了 13 种标注工具在技术维度上的对比结果:

工具名称	平台兼容性	工具架构	安装简便性	收费标准
BRAT	跨平台兼容	客户端-服务器架构	需Python环境,中等难度	免费
Doccano	跨平台兼容	开源工具	Docker一键部署,简便	免费
Djangology	跨平台兼容	web-based, 框架	需Django配置,较复杂	免费
GATE	不能与众包平台兼容	web-based	需Java环境,中等难度	免费
LightTag	不能与众包平台兼容	commercial tool	无需安装,浏览器使用	仅学术使用免费
MedTAG	跨平台兼容	web-based	需Docker,简便	免费
Prodigy	跨平台兼容	web-based	Python包安装,简便	商业授权(付费)
POTATO	跨平台兼容	开源工具	需Node.js环境,较复杂	免费
Tagtog	跨平台兼容	web-based	Python包安装,简便	部分功能免费
TeamTat	跨平台兼容	web-based	需服务器部署,简便	免费
WAT-SL	不能与众包平台兼容	web-based	Python包安装,简便	免费
WebAnno	跨平台兼容	web-based	需Tomcat部署,较复杂	免费
YEDDA	不能与众包平台兼容	开源工具	需Python环境,简便	免费

Figure 1. Comparison results of 13 tools in the technical dimension
图 1. 13 种工具在技术维度上的对比结果

2.2.2. 数据维度

本维度旨在分析标注工具的数据格式，主要依据以下两项标准进行评估：

(1) 导入文件格式(标准格式如 JSON，XML 等)

据 Mariana Neves 和 Jurica Seva (2021) [3]，Djangology、TeamTat、WAT-SL 及 YEDDA 在导入文件格式上采用了非标准格式，这可能会给标注员在数据导入过程中带来一定的挑战。

(2) 导出文件格式(标准格式如 JSON，XML 等)

据 Mariana Neves 和 Jurica Seva (2021) [3]，BRAT、Djangology、POTATO、WAT-SL 及 YEDDA 在导出文件格式上采用了非标准格式，这可能给标注员在标注数据存储方面造成不便。

图 2 展示了 13 种标注工具在数据维度上的对比结果：

工具名称	导入文件格式	导出文件格式
BRAT	标准格式	独立文件格式或标准矢量图格式
Doccano	CSV, JSON, 纯文本, CoNLL	CSV, JSON
Djangology	非标准格式	非标准格式
GATE	HTML, XML, PDF, Word, CoNLL, CSV, JSON, plain text	XML, HTML
LightTag	标准格式	标准格式
MedTAG	允许所有格式	XML, JSON, CSV
Prodigy	标准格式	JSONL
POTATO	标准格式JSON或者分隔符文件	PDF, CSV, TXT
Tagtog	PDF, CSV, plain text, source code, HTML, 标准格式	TXT, HTML, XML
TeamTat	PDF, BioC, plain text	BioC, XML
WAT-SL	非标准格式	非标准格式
WebAnno	标准格式	标准格式
YEDDA	UTF-8编码的TXT文件	BIO, BIOES

Figure 2. Comparison results of 13 tools in the statistical dimension
图 2. 13 种工具在数据维度上的对比结果

2.2.3. 功能维度

本维度聚焦于各项标注工具的标注功能属性分析，包含以下 13 项精细化指标：

(1) 标注界面

根据调研结果，多数标注工具的界面设计遵循直观性原则并注重用户体验。其中 Doccano [2]和 POTATO [6]采用纯键盘快捷键操作模式，摒弃传统鼠标交互方式。WebAnno [7]通过引入分页加载机制实现性能优化，该机制通过控制可显示句子数量，有效消除文档规模对系统响应速度的影响。相比之下，Prodigy [2]的界面复杂度较高，其多层菜单结构和功能嵌套设计可能对标注效率产生负面影响。

(2) 可配置性

WebAnno 在配置维度上展现出独特优势：除支持工具参数自定义和算法模块修改外，还提供标注层可见性配置功能[7]。而 Djangology、LightTag、Prodigy、Tagtog 等四款工具在开源协议方面存在显著局限，或未提供修改许可(Djangology、LightTag 采用专有许可)，或完全封闭配置接口(Prodigy 仅开放部分 API，Tagtog 采用 SaaS 模式限制用户权限)。

(3) 是否支持序列标注

仅 Djangology 和 WAT-SL 在此维度存在功能缺失，据 Kiesel J, Wachsmuth H, Al-Khatib K, Stein B (2017) [8]，WAT-SL 作为专用片段标注工具，其设计定位决定了其无法进行序列标注(命名实体识别、词性标注等任务)。

(4) 是否支持关系标注

Djangology、MedTAG、WAT-SL 三款工具未实现该功能，其中 MedTAG 作为生物医学领域专用工具，其功能局限性可能源于特定领域标注规范的约束[9]。

(5) 是否支持多标签标注

MedTAG、Prodigy、Tagtog、WAT-SL 在此维度存在功能缺陷, 其中 Prodigy 的多标签标注缺失与其界面设计复杂度存在潜在关联。

(6) 是否支持本体或术语

仅 Djangology 和 WAT-SL 未提供本体导入接口, 这限制了其在知识库构建类任务中的应用潜力。

(7) 是否拥有数据检索功能

仅有 BRAT、Djangology、LightTag、Tagtog、TeamTat 五款工具具备数据检索模块。值得注意的是, 据 Juan Miguel Cejuela (2014) [5], Tagtog 创新性地实现了语义概念检索, 而其他四款工具均采用关键词匹配的基础检索策略。

(8) 是否支持协作标注

12 款工具中有 11 款支持多用户协同标注。Prodigy 在此维度的功能缺失, 与其采用的单线程任务处理架构存在直接关联。

(9) 是否支持标注者间一致性(IAA)测算

Doccano、Prodigy、Tagtog 三款工具未集成 IAA (Inter-Annotator Agreement) 测算模块。此功能缺失可能影响标注质量控制流程的完整性。

(10) 是否可以主动学习

主动学习能力分布呈现显著分化: 据 Mariana Neves 和 Jurica Seva (2021) [3], Prodigy、POTATO、Tagtog、TeamTat 四款工具具备主动学习能力, 其余工具虽未实现完整主动学习流程, 但都内置预测功能或者支持上传或下载机器学习模型。

(11) 是否支持团队登录

BRAT、Prodigy、POTATO、Tagtog、WAT-SL、WebAnno 六款工具缺乏团队登录管理功能。据 Juan Miguel Cejuela (2014) [5], Tagtog 采用完全匿名登录机制, 其余五款工具仅支持个人用户登录体系。

(12) 是否支持多语言

11 款工具中有 9 款支持多语言标注。Djangology 和 WAT-SL 的语言局限性可能影响其多语言标注任务的适应性。

(13) 是否能与 Medline 或者 Pubmed 集成

领域专用性功能模块方面, 据 Mariana Neves 和 Jurica Seva (2021) [3], 仅 MedTAG、Tagtog、TeamTat 三款工具可以与 Medline 或者 Pubmed 集成。此功能设计针对生物医学文本标注需求, 体现了工具在垂直领域的应用优化。

图 3 展示了 13 种标注工具在功能维度上的对比结果:

工具名称	BRAT	Doccano	Djangology	GATE	LightTag	MedTAG	Prodigy	POTATO	Tagtog	TeamTat	WAT-SL	WebAnno	YEDDA
标注功能	直观, 用户友好, 点选式	直观, 用户友好, 不用鼠标	直观, 用户友好	直观, 用户友好	直观, 用户友好, 点选式	直观, 用户友好, 点选式	复杂	直观, 用户友好, 不用鼠标	直观, 用户友好	直观, 用户友好	直观, 用户友好	直观, 用户友好, 分页机制	直观, 用户友好
可配置性	✓	✓	×	✓	×	✓	×	✓	×	✓	✓	✓	✓
是否支持序列标注	✓	✓	×	✓	✓	✓	✓	✓	✓	✓	×	✓	✓
是否支持关系标注	✓	✓	×	✓	✓	×	✓	✓	✓	✓	×	✓	✓
是否支持多标签标注	✓	✓	✓	✓	✓	×	×	✓	×	✓	×	✓	✓
是否支持本体或术语	✓	✓	×	✓	✓	✓	✓	✓	✓	✓	×	✓	✓
是否有数据检索功能	✓	×	✓	×	✓	×	×	×	✓	✓	×	×	×
是否支持协作标注	✓	✓	✓	✓	✓	×	×	✓	✓	✓	✓	✓	✓
是否支持IAA测算	✓	×	✓	✓	✓	×	×	✓	×	✓	✓	✓	✓
是否可以主动学习	×	×	×	×	×	×	×	✓	✓	✓	×	×	×
是否支持团队登录	×	✓	✓	✓	✓	✓	×	×	×	✓	×	×	✓
是否支持多语言	✓	✓	×	✓	✓	✓	✓	×	✓	✓	×	✓	✓
是否能与Medline或者Pubmed集成	×	×	×	×	×	✓	×	×	✓	✓	×	×	×

Figure 3. Comparison results of 13 tools in the functional dimension
图 3. 13 种工具在功能维度上的对比结果

3. 代表性文本标注工具的优缺点分析

3.1. 各工具存在的优势与不足

(1) BRAT

据 Pontus Stenetorp (2012) [10], BRAT 工具的核心优势体现在三方面: 其一, 支持统一资源标识符 (URI) 的集成, 可通过超链接直接关联本体资源, 实现语义规范化操作; 其二, 内置预测模块采用置信度评分机制, 优先采纳高置信度标注结果; 其三, 协同标注功能支持实时进度同步, 增强团队协作效率。然而, 该工具在数据处理维度存在显著局限性: 导出格式采用非标准化设计, 虽支持学术论文矢量图输出, 但不利于跨平台标注格式统一[11]。此外, 其标注文件采用独立存储格式, 复杂标注场景(文本与图像叠加)支持不足, 长文档标注效率较低, 实时协同用户上限设定为 50 人, 且存在非拉丁语系字符处理缺陷[12]。当前版本不能在线使用。

(2) Doccano

据 Musabeyezu Fortune (2019) [2], Doccano 的创新价值在于支持表情符号(Emoji)标注这一非常规标注类型。但功能维度评估显示: 该工具缺乏数据检索模块, 未集成标注者间一致性(IAA)测算机制, 且长文档标注性能表现平庸。上述功能缺失可能限制其在大型标注项目中的适用性。

(3) Djangology

据 Emilia Apostolova (2010) [13], Djangology 的显著优势在于其零配置特性——用户无需进行本地化安装与参数设置即可通过浏览器直接访问标注界面。但功能维度的分析表明, 其在 13 项评估标准中仅满足 5 项。数据方面, 文件导入与导出均采用非标准格式, 可能增加数据转换成本。

(4) GATE

据 Musabeyezu Fortune (2019) [2], GATE 工具实现了本体资源的超链接直接访问功能, 但在系统集成维度存在双重制约: 其一, 与主流众包平台不兼容[14]; 其二, 官方提供的安装文件不够全面, 可能影响普通用户的使用体验。

(5) LightTag

据 Tal Perry (2021) [4], LightTag 在关系标注领域表现出显著优势, 其实体拖拽交互机制有效提升了关系标注效率。该工具内置双重机制: 预测结果自动采纳算法结合冲突检测系统, 支持同类型标注的批量验证处理, 并采用系统置信度评分机制。但其许可协议存在使用限制: 仅面向学术研究提供免费版本, 且采用专有许可模式, 可能制约其应用场景扩展。

(6) MedTAG

据 Fabio Giachelle (2021) [9], MedTAG 在功能维度评估中满足 9 项核心标准, 展现出较强大的标注功能。但其仅支持纯文本格式的输入文档, 且标注规范严格限定于生物医学文本处理范畴, 跨领域扩展性受限。

(7) Prodigy

Prodigy 通过一次只专注于一个标注任务的模式提升标注质量, 实时进度可视化功能有助于管理员监督标注进度[2]。但经济性分析显示, 该工具使用成本较高, 可能增加项目预算压力。且功能维度评估表明, 标注界面的多层交互设计可能影响标注员的标注效率。

(8) POTATO

据 Jiaxin Pei (2023) [6], POTATO 的创新性设计体现在: 通过上下文关键词高亮与条件性随机高亮相结合, 显著降低标注人员对关键词的过度依赖, 大幅提升标注准确率。该工具内置标注质量管控机制, 包括注意力测试与标注时间追踪, 有助于提升标注质量。但标注界面采用纯键盘快捷键设计, 可能给依

鼠标交互的用户带来操作适配困难。

(9) Tagtog

据 Juan Miguel Cejuela (2014) [5], Tagtog 在实体规范化处理方面表现突出, 但其基础功能免费而高级功能需付费解锁。操作界面缺乏键盘快捷键配置选项, 且 SaaS 交付模式导致配置接口完全封闭, 限制了专业用户的定制化需求。

(10) TeamTat

据 Rezarta Islamaj (2020) [15], TeamTat 独创的匿名任务分配机制在协作标注场景中展现出独特优势: 管理员可离散化分配标注任务, 确保标注人员独立完成任务而不受群体决策干扰。但数据维度存在短板, 非标准导入格式可能增加跨平台数据迁移成本。

(11) WAT-SL

WAT-SL 作为片段标注专用工具, 填补了该细分领域的工具空白, 且标注操作每次都会自动保存, 便于标注员或管理者发现标注问题[8]。但功能维度分析显示, 该工具仅满足 13 项评估标准中的 4 项, 且缺失上下文关键词高亮这一基础功能。导入和导出文件格式采用非标准格式, 可能造成数据转换困难。

(12) WebAnno

据 Seid Muhie Yimam (2013) [7], WebAnno 内置预测机制提供自动标注建议, 但是只适用于跨度标注任务。团队协作标注时, 标注员可以进行独立标注, 确保协作标注的效率。但子词元跨度标注功能缺失构成技术瓶颈, 且安装复杂度较高, 文献资料表明其安装时间超过两小时, 可能增加运维成本。

(13) YEDDA

据 Jie Yang (2018) [16], YEDDA 通过键盘快捷键与命令行标注的双模操作, 实现了高效标注范式。其轻量级架构设计使得安装流程极简, 显著降低了工具使用门槛。但数据兼容性存在不足, 非标准文件格式可能增加预处理工作量。

3.2. 各工具对比评估结果

分析结果表明, 功能特性维度呈现出最显著的标准满足差异:

1. 高频满足标准(≥ 11 项工具达标):

- 标注界面设计直观且注重交互体验(12 项工具, Prodigy 未实现)
- 支持多用户协同标注功能(同上)
- 具备序列标注能力(11 项工具, Djangology 与 WAT-SL 未支持)
- 支持本体/术语库集成(同上)
- 支持多语言处理(同上)

2. 低频满足标准(≤ 5 项工具达标):

- 与 Medline/PubMed 医学文献库集成(仅 3 项工具)
- 具备主动学习能力(仅 4 项工具)
- 内置数据检索模块(仅 5 项工具)

在跨维度综合评估中, 工具的标准满足度呈现显著分层:

- 最优表现组: TeamTat 满足 18 项标准(仅缺失导入文件标准格式支持), BRAT、GATE、LightTag、POTATO 均满足 14 项标准;
- 最低满足组: WAT-SL 仅满足 7 项标准, Djangology 与 Prodigy 各满足 9 项标准。

4. 结论与展望

4.1. 结论

由于纯人工标注效率低下, 标注者们逐渐转向自然语言处理(NLP)标注工具以提升标注效率。鉴于文本数据的广泛性及标注任务以文本标注为主, 本研究从众多工具中遴选出 13 个符合预设标准的代表性工具进行对比分析, 从技术实现、数据适配和功能特性三个维度系统评估了各工具的标注效能和操作便捷性, 并剖析了各工具相较其他候选工具的独特优势与潜在局限。

本研究的工具对比结果揭示, 功能特性维度在标注工具标准满足度上呈现显著分层。高频满足标准(≥ 11 项工具达标)集中于交互设计(Prodigy 例外)、多用户协同标注、序列标注(Djangology 与 WAT-SL 缺失)、本体集成及多语言支持; 而医学文献库对接、主动学习、数据检索等低频标准(≤ 5 项工具达标)则凸显功能扩展性局限。多工具评估结果显示, TeamTat 以 18 项标准满足度领跑(仅缺失文件格式兼容), BRAT、GATE、LightTag、POTATO 构成次优梯队(14 项达标), 而 WAT-SL (7 项)、Djangology 与 Prodigy (9 项)因功能覆盖不足位列末席。通过多维度量化对比, 本研究为不同标注需求场景下的工具选择提供了决策依据。

4.2. 展望

基于前述研究中揭示的低频满足标准(即医学文献库集成、主动学习能力、数据检索模块), 未来 NLP 标注工具的发展需着重优化以下四个方向:

现有工具已呈现两极分化特征: BRAT 等工具通过模块化设计实现跨场景通用性, 而 MedTAG、WAT-SL 等则针对生物医学等垂直领域或片段标注等标注任务进行深度适配。未来工具开发应构建“基础功能 + 领域插件”的混合架构, 既保证核心标注能力的普适性, 又支持通过插件扩展专业标注范式。

尽管现有工具普遍内置预测模型以辅助标注, 显著降低了人工标注负担, 但需重点研发工具的主动学习模型。通过构建动态反馈循环, 使系统能基于标注员的标注操作和验证行为实时优化预测模型, 实现从“被动纠错”到“主动引导”的范式转变, 这是提升标注效率的关键突破口。

针对长文本及复杂事件结构标注需求, 应研发多维度搜索引擎, 集成语义索引与可视化导航功能, 解决跨段落信息关联难题, 还可以增加分页机制, 限制可显示的句子数量, 使标注员更专注于标注任务。

基于 Web 的工具虽具备零安装优势, 但需建立分级安全机制: 对公开数据集采用云端协作模式, 对敏感数据提供本地化部署选项, 并支持两种模式间的无缝切换。同时应采用联邦学习等技术, 实现多方数据联合建模时的隐私保护, 满足医疗、金融等领域的合规要求。

基金项目

教育部人文社会科学研究规划基金项目《基于动态句法学和分布语义学的汉语照应语动态研究》, 编号 24YJA740032。

参考文献

- [1] Cai, L., Wang, S.T., Liu, J.H. and Zhu, Y.Y. (2020) Survey of Data Annotation. *Journal of Software*, **31**, 302-320.
- [2] Fortune, M. (2019) Comparative Study of Annotation Tools and Techniques. Ph.D. Thesis, African University of Science and Technology.
- [3] Neves, M. and Ševa, J. (2019) An Extensive Review of Tools for Manual Annotation of Documents. *Briefings in Bioinformatics*, **22**, 146-163. <https://doi.org/10.1093/bib/bbz130>
- [4] Perry, T. (2021) Lighttag: Text Annotation Platform. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Punta Cana, November 2021, 20-27.

- <https://doi.org/10.18653/v1/2021.emnlp-demo.3>
- [5] Cejuela, J.M., McQuilton, P., Ponting, L., Marygold, S.J., Stefancsik, R., Millburn, G.H., *et al.* (2014) Tagtog: Interactive and Text-Mining-Assisted Annotation of Gene Mentions in PLOS Full-Text Articles. *Database*, **2014**, bau033. <https://doi.org/10.1093/database/bau033>
 - [6] Pei, J., Ananthasubramaniam, A., Wang, X., Zhou, N., Dedeloudis, A., Sargent, J., *et al.* (2022) POTATO: The Portable Text Annotation Tool. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Abu Dhabi, December 2022, 327-337. <https://doi.org/10.18653/v1/2022.emnlp-demos.33>
 - [7] Yimam, S.M., Gurevych, I., Eckart de Castilho, R. and Biemann, C. (2013) WebAnno: A Flexible, Web-Based and Visually Supported System for Distributed Annotations. *Proceedings of the 51th Annual Meeting of the Association for Computational Linguistics—System Demonstrations*, Sofia, 4-9 August 2013, 1-6.
 - [8] Kiesel, J., Wachsmuth, H., Al Khatib, K. and Stein, B. (2017) WAT-SL: A Customizable Web Annotation Tool for Segment Labeling. *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, Valencia, April 2017, 13-16. <https://doi.org/10.18653/v1/e17-3004>
 - [9] Giachelle, F., Irrera, O. and Silvello, G. (2021) Medtag: A Portable and Customizable Annotation Tool for Biomedical Documents. *BMC Medical Informatics and Decision Making*, **21**, Article No. 352. <https://doi.org/10.1186/s12911-021-01706-4>
 - [10] Stenetorp, P., Pyysalo, S., Topic, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). Brat: A Web-Based Tool for NLP-Assisted Text Annotation. *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Avignon, April 2012, 102-107.
 - [11] Strippel, C., Laugwitz, L., Paasch-Colberg, S., Esau, K. and Heft, A. (2022) BRAT Rapid Annotation Tool. *Medien & Kommunikationswissenschaft*, **70**, 446-461. <https://doi.org/10.5771/1615-634x-2022-4-446>
 - [12] Patil, C.S. (2022) NLP Assisted Text Annotation. *International Journal of Scientific Research in Engineering and Management*, **6**, 1-10. <https://doi.org/10.55041/ijssrem14651>
 - [13] Apostolova, E., Neilan, S., An, G., Tomuro, N. and Lytinen, S. (2010) Djangology: A Light-Weight Web-Based Tool for Distributed Collaborative Text Annotation. *Proceedings of the International Conference on Language Resources and Evaluation*, Valletta, May 2010, 17-23.
 - [14] Bontcheva, K., Cunningham, H., Roberts, I., Roberts, A., Tablan, V., Aswani, N., *et al.* (2013) GATE Teamware: A Web-Based, Collaborative Text Annotation Framework. *Language Resources and Evaluation*, **47**, 1007-1029. <https://doi.org/10.1007/s10579-013-9215-6>
 - [15] Islamaj, R., Kwon, D., Kim, S. and Lu, Z. (2020) Teamtat: A Collaborative Text Annotation Tool. *Nucleic Acids Research*, **48**, W5-W11. <https://doi.org/10.1093/nar/gkaa333>
 - [16] Yang, J., Zhang, Y., Li, L. and Li, X. (2018) YEDDA: A Lightweight Collaborative Text Span Annotation Tool. *Proceedings of ACL 2018, System Demonstrations*, Melbourne, July 2018, 31-36. <https://doi.org/10.18653/v1/p18-4006>