arXiv:2301.03095v1 [cs.HC] 8 Jan 2023

MEGAnno：计算笔记本中的自然语言处理探索式标注

Dan Zhang*，**Hannah Kim***，**Rafael Li Chen**、**Eser Kandogan**、**Estevam Hruschka**Megagon Labs{dan_z,hannah,rafael,eser,estevam}@megagon.ai

摘要

我们介绍 MEGAnno，这是一个面向自然语言处理研究人员和从业者的新型探索性标注框架。与现有仅专注于数据标注的工具不同，我们的框架旨在支持更广泛、迭代式的机器学习工作流程，涵盖数据探索和模型开发。借助 MEGAnno 的 API，用户可以通过高级搜索和自动建议功能以编程方式探索数据，并随着项目的发展逐步更新标注方案。结合我们的小组件，用户可以在承载自然语言处理项目其余内容的同一笔记本中，交互式地对多个项目同时进行排序、筛选和分配标签。我们通过一个情感分析用例，展示 MEGAnno 灵活、探索式、高效且无缝的标注体验。

1 引言

数据标注是机器学习生命周期中的重要步骤，因为训练数据的质量和数量会直接影响模型性能（Geiger 等人，2021年）。遗憾的是，现有标注工具往往将数据标注步骤与更广泛的机器学习生命周期割裂开来，忽视研究人员和从业者的迭代工作流程。然而，数据选择、探索性数据分析、数据标注、模型训练和评估等活动并非按顺序依次发生（Rahman 和 Kandogan，2022年）。相反，持续迭代是改进数据、标注和模型所必需的（Hohman 等人，2020年）。

为进一步探究这一差距，我们考察了机器学习模型开发生命周期中的数据标注实践。基于对本组织六位研究人员开展的一项形成性研究，我们将他们的标注实践概括为图1所示的“双循环”模型。完成数据预处理后

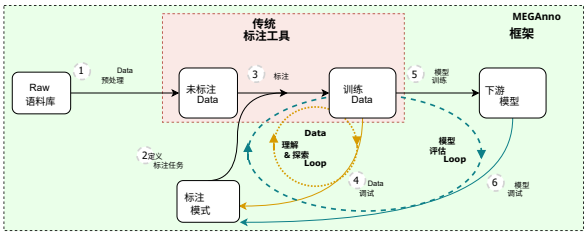

图1：用于数据标注的双循环模型：(1) 数据理解/探索循环（黄色）：在探索和标注数据的过程中，反复更新标注模式；(2) 模型评估循环（绿色）：通过调试数据，在多轮迭代中训练并改进下游模型。大多数工具只关注标注步骤（红框）。MEGAnno 旨在在该框架（绿框）中无缝涵盖这两个循环。

过程中（图 1 ①），研究人员会根据需要收集哪些标签、需要多少数据点等来定义标注模式（图 1 ②）。在探索和标注数据时（图 1 ③），他们通常会回过头来完善标注模式，因为他们对数据的理解有所加深，且对任务的心智模型也有所更新（图 1 ④）。例如，在文档分类任务中，用户可能会从定义较为宽泛的类别标签开始，并随着发现相关文档而增加更多选项（Felix 等人，2018年）。在本文中，我们将这一循环称为 **数据理解/探索循环**。接下来，标注数据会从标注工具中导出，并用于训练模型（图 1 ⑤）。然而在实践中，机器学习模型训练很少能一次完成，通常需要经过多轮标注、训练、数据调试和模型调试（Pustejovsky 和 Stubbs，2012年）。图 1 ⑥ 展示了研究人员可能需要收集更多数据（例如，由于下游模型的预测性能不理想，需要为代表性不足的类别补充数据）或进一步完善模式的情况。我们将这一循环（即模型训练、评估和调试，

* 贡献相同。

收集更多数据并再次训练）称为**模型评估循环**。

我们发现，现有的大多数工具很少支持自然语言处理研究人员和从业者采用的迭代式双循环工作流。具体而言，我们确定了支持完整标注生命周期面临的三项主要挑战：

- **机器学习工具之间的差距**。大多数现有工具都是独立的，且针对特定的机器学习步骤而设计，因此研究人员在日常工作流中需要频繁进行上下文切换和数据迁移，造成额外开销。

- **缺乏可定制的精细控制**。并非所有数据点都同等重要。用户往往希望优先处理特定批次（例如，为了更好地覆盖类别或领域，或聚焦于下游模型无法准确预测的数据点）。尽管一些近期基于主动学习的工具（Montani 和 Honnibal，2018年；hua）可以为下一批次提供建议，但大多数工具无法提供可定制的精细控制，也无法结合下游模型进行控制（即覆盖两个循环）。

- **缺乏对项目演进的支持**。当前的标注工具通常假定数据收集任务已明确定义且不会改变，却忽视了随着探索开展、见解不断积累，标注项目也可能随之演进。因此，这些工具无法帮助用户做出演进决策；其不可变的特性也使应用这些变更变得困难。

为应对这些挑战，我们推出 MEGAnno，这是一个灵活、探索式、高效且无缝的标注框架，面向自然语言处理研究人员和从业者。它提供了流畅无缝的体验，让用户能够在同一笔记本中完成数据预处理、标注、分析、模型开发和评估；笔记本是数据科学从业者日常工作中广泛使用的环境。MEGAnno 提供可定制界面，帮助用户通过丰富的基于启发式的搜索、自动建议，以及基于主动学习的下一数据批次建议，推动项目朝预期方向发展。考虑到项目演进，MEGAnno 采用灵活的任务模式，并内置分析仪表板以辅助决策。为据我们所知，MEGAnno 是首个灵活的探索式标注框架，能够在计算笔记本中无缝支持机器学习工作流程（图 1：绿色方框）。

2 MEGAnno

2.1 框架概述

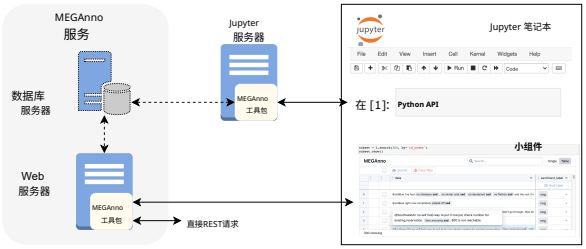

图 2：MEGAnno 框架通过工具包提供探索性标注服务。该工具包可作为 Python 库安装，并提供编程接口；此外，框架还包含提供语言无关 REST API 的 Web服务器，以及用于存储数据、标注和相关产物的内部数据库。实线表示通过 Python API 调用和 REST 调用进行的程序化交互；这些调用由笔记本小组件代为发起，或由经过身份验证的应用程序直接发起。虚线表示内部接口：MEGAnno 工具包通过这些接口与数据库通信，这些接口对用户隐藏。

MEGAnno 通过以下组件提供服务：1）用于存储数据、标注以及标注过程中产生的各种产物的内部数据库；2）用于以编程和可视化方式进行数据探索和标注的 MEGAnno 工具包，提供 Python API；以及 3）提供与语言无关的 REST API 的 Web 服务器（图 2）。在 Jupyter服务器上安装该工具包后，用户即可通过任意已连接的笔记本使用我们的 Python API 和基于 React 的组件，管理项目、探索并标注数据。

数据模型 数据记录是指预处理数据语料库中用于标注的条目。它可以是一个句子、一段文字、一篇文档，也可以是由多个文本展平而成的文本，例如一组问答。标签是用户标注输出的最小单位。MEGAnno 目前支持记录级标签（例如文档或句子的主题）和文本片段级标签（例如命名实体）。标注是标注员为一条数据记录赋予的一组标签。元数据是指与数据记录内容（例如外部生成的词性标签、嵌入）或标注（例如标注所花费的时间、标注员间分歧）相关的附加信息。

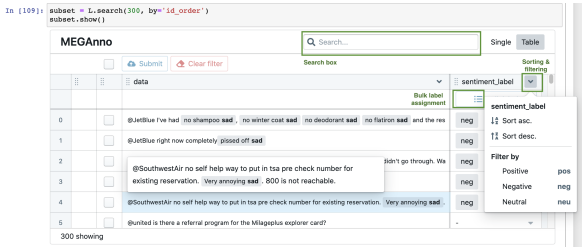

图3：表格视图用于展示多条数据记录。将鼠标悬停在某条数据记录上，会在弹出窗口中显示其完整文本。此视图支持通过搜索、排序和筛选标签，以及进行单条/批量标注来探索数据。

这类信息有助于机器学习迭代的各个步骤。子集是数据库中数据记录的一部分。子集可以由随机数据记录组成，也可以通过搜索查询生成，以匹配特定特征。

任务模式 我们通过可自定义的 JSON 格式模式支持多种任务。要收集标签，用户需要指定级别（即记录或文本片段），并提供一个可供选择的选项列表。对于一项任务，MEGAnno 支持任意数量的这两种类型的标签。我们将在第3节看到用于情感预测和抽取任务的具体模式示例。在任何阶段，用户都可以随时更新模式，以反映项目的演变。为保持数据的一致性，模式更新受到一些限制。始终允许添加新标签或新标签选项。删除标签或选项会触发数据库查询，如果存在已标注的此类实例，MEGAnno 将向用户发出警告。 [1]

2.2 MEGAnno Jupyter 笔记本小部件

MEGAnno 的交互式笔记本小部件具有以下功能：1）我们新颖的表格视图，可促进探索性标注和批量标注；2）单条视图，与传统标注界面类似。

表格视图 表格视图（图 3）显示子集 中的数据项及其标注（如有）。每个记录级标签都显示为一列，而片段级标签则与数据列中高亮显示的文本片段一同呈现。用户可以将鼠标悬停在某个数据项上，在弹出窗口中查看全文。搜索框支持三种搜索方式（精确、模糊和基于正则表达式），以进一步筛选数据子集。用户可以根据任意记录级标签使用下拉菜单对行进行排序和筛选。

[1]MEGAnno 提供了清理旧标签并自动重试的选项。

要分配记录级标签，用户可以点击单元格中的箭头按钮，并从下拉选项中选择。或者，用户也可以选中多个记录，再点击批量标注按钮，同时为这些记录分配相同的标签。

单条视图 表格视图适合探索，但空间有限，使得片段级标注变得繁琐。因此，我们还提供了单条视图，让用户能够更专注地查看数据。点击右上角的“单条”按钮，或双击一条数据记录，小部件就会切换到单条视图（图 4）。在此视图中，用户可以在右侧分配记录级标签，并在左侧通过选择或拖动目标片段，再从下拉选项中选择标签来分配片段级标签。用户可以根据表格视图中指定的顺序，使用“上一条”/“下一条”按钮遍历子集。用户可随时点击右上角的按钮在两种视图之间切换；切换视图时，小部件会保留所有尚未提交的标注。

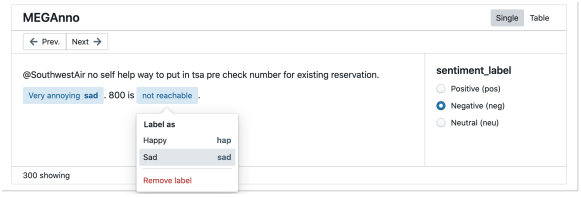

图4：逐一将单视图转换为标注数据。在此视图中，用户可以拖动文本片段并为其添加标签，以完成抽取任务。

与多位标注人员协作 标注很少由一个人独自完成。作为迈向协作式标注的初步步骤，MEGAnno为每位标注人员提供彼此隔离的命名空间。用户连接服务时通过唯一的身份验证令牌表明身份，并且只能通过小组件更新自己的标签。 MEGAnno 提供协调视图（图5），以查看不同人员添加的标签并解决可能出现的冲突。

仪表板 MEGAnno还提供了一个内置的可视化监控仪表板（图 6）。随着项目不断发展，用户需要了解项目的状态，以决定接下来的步骤，例如收集具有特定特征的更多数据点，或为

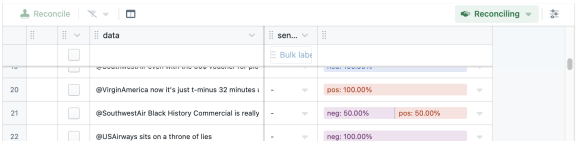

图5：用于显示数据点现有标签分布的协调视图。

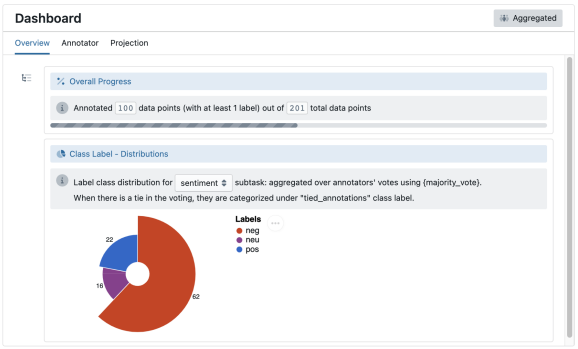

图6：用于监控项目进度和统计信息并辅助决策的仪表板小组件。

任务定义添加新的类别。为辅助此类分析，仪表板小组件根据对试点用户的调查，汇集了常见统计数据和分析可视化内容。“概览”面板展示了整体进度和各标签类别分布的统计信息。如果涉及多位标注人员，分布将反映标注人员之间的多数投票 [2]。由于空间有限，其余的“标注员”和“投影”面板默认隐藏。为帮助识别存在问题的标注员，标注员面板会显示总体个人贡献以及与其他人意见不一致程度等统计信息。投影面板提供可自定义的可视化方式，将数据点投影到二维可视空间中。默认情况下，我们展示 Sentence-BERT（Van der Maaten 和 Hinton，2008年）嵌入的 t-SNE（Reimers 和 Gurevych，2019年）投影。

2.3 MEGAnno API

项目管理 管理模块通过 Project 类提供各种接口，用于配置和监控标注项目。import_data 可将 CSV 或 JSON 文件中的数据记录加载到数据库中。随着项目的发展，set_config 可更新项目配置。set_meta 可通过用户定义函数为每条数据记录分配元数据（例如词性标签、文档嵌入）。get_status 返回项目状态，例如已标注的数据记录数量以及每个标签的详细统计信息。

MEGAnno 的一项关键功能是选择有趣的数据子集并在小组件中展示。用户可以通过搜索引擎主动生成子集，也可以通过自动建议以数据驱动的方式生成子集。

搜索子集 MEGAnno 支持通过 Project.search API 对数据记录、标注和用户定义的元数据进行复杂搜索。用户可以通过关键词（例如，提及“客户服务”的文档）或正则表达式搜索数据记录，以表达更复杂的模式。用户还可以根据已分配的标签搜索数据库（例如，带有正向情感标签的记录）。稍后将会解释，MEGAnno 认可辅助信息对机器学习项目的价值，并提供针对元数据的高级搜索功能。例如，用户可以使用结合正则表达式和词性标签的模式进行查询，如 project.search("(best|amazing)<ADJ> <NOUN>", by="POS").

自动子集建议 用户发起的搜索有助于以受控的方式探索数据集，但搜索质量取决于用户对数据和领域的了解或启发式判断。MEGAnno 提供自动子集建议引擎，协助进行探索。用户可根据需要接入外部建议模型，自定义该引擎。目前，该引擎提供两类技术：

- **基于嵌入的建议** 根据用户提供的数据嵌入向量生成建议。Subset.suggest_similar 会建议查询子集中数据的近邻。Project.suggest_coverage 会以无监督的方式检查嵌入空间中的所有数据记录，并从标注较少的区域建议数据点，以提高语料库的标注覆盖率。

- **主动建议** 利用主动学习技术，为下游模型推荐信息量最高的数据。借助 ModAL 等库（Danka 和 Horvath），

[2]用户可以提供自己的聚合函数，以解决标注人员之间的冲突

用户可以根据模型不确定性、熵等因素，从多种选择策略中进行选择。由于 MEGAnno 在笔记本中提供了无缝体验，涵盖从标注到模型训练和调试的整个循环，因此用户可以主动选择一个子集、为其添加标注、更新模型并再次测试，而无需在不同环境之间切换。

选择引擎的输出是 Subset 类的实例，具有以下方法：show 返回一个笔记本小部件，用于交互式探索和标注。batch_annotation 为子集中的所有数据记录设置相同的记录级标签。suggest_similar 根据具有有效距离函数的某些元数据，为查询子集中的每条记录从数据库中返回包含最相似数据的新子集。

3 用例：情感分析

我们介绍一个用例，展示 MEGAnno如何支持自然语言处理研究人员和从业者的工作流程中的数据标注。Meggie 是一名数据科学家，她希望为自己的项目训练一个与情感相关的模型。她获取了一个 Twitter 数据集[3]，内容涉及美国航空公司，并决定使用MEGAnno对其进行标注。

导入数据 在一个空白笔记本中，Meggie 首先使用 MEGAnno Python API 初始化名为“Tweet Sentiment”的项目。她从产品经理那里获得一份数据副本。该经理经常使用 Google 电子表格；Meggie 则通过其已发布的链接导入数据。

```
来自meganno 导入Project2L = Project(<
au
dc
```

设置初始模式 由于对数据不太了解，Meggie 决定先收集二元标签，并为项目设置模式。Meggie 知道 MEGAnno 支持灵活且可编辑的模式，因此不担心第一轮就要制定出完美的模式，可以先开始探索和标注。

```
标签_模式 = [{2 "label_name":"
se
```

Kaggle 数据集（for Everyone 库）提供了真实情感标签。不过，为了演示，我们忽略这些标签，并假设 Meggie 只能获取原始推文。该数据集包含 2015 年 2 月抓取的 1.4 万条关于美国主要航空公司的推文。

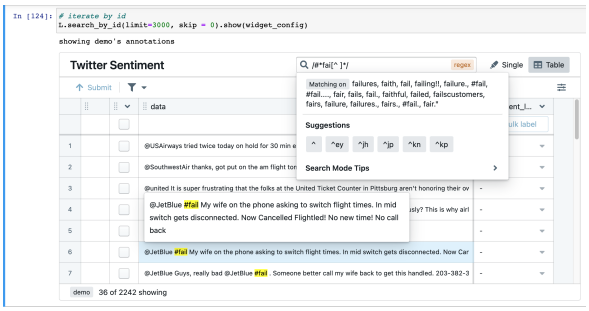

图 7：通过正则表达式进行 UI 搜索。匹配的关键词会高亮显示。

```
3 "level":"record",4 "options": ["正
面
```

探索并标注 她首先在小组件的表格视图中探索前 300 个数据点。她使用搜索框，筛选出包含关键词“amazing”的子集。不出所料，大多数数据记录都体现出正面情感，因此她使用“批量标注”按钮，为多个数据项赋予正面标签。接下来，她想查看带有与失败相关主题标签的推文，于是尝试使用正则表达式搜索`#fai[^ ]*` (图 7)。为了更好地理解数据集，Meggie 选择使用词性元数据进行更深入的探索。她从 spaCy（Honnibal 和 Montani，2017年）导入词性标注器，并检索符合有趣模式的推文，例如`best <名词> of <名词>`。

```
import spacy2标注器 = spacy.load("en_core_
w
容
L.
ta
w
```

通过这种探索式方法和批量标注功能，她可以更快地以更可控的方式进行标注。

候选项建议 又尝试了几种启发式方法后，她再也想不出新点子，于是利用建议功能。她选择 sentence bert（Reimers 和 Gurevych，2019年）编码器作为元数据生成函数。她首先使用相似度 策略发起查询，以收集与之前词性搜索所检索结果相似的数据点。最后，她想查看嵌入空间中覆盖较少区域的样本，以便

提高训练数据的多样性，并发起覆盖查询。

```
m = SentenceTransformer("all-MiniLM-L6")# 设
置
lis
数
si
的
co
```

更新模式 Meggie现在对收集到的标签感到满意，但她希望更进一步，了解哪些词语或短语会导致某种情感判断。因此，她通过添加一个名为 sentiment_span 的新片段级标签来更新模式，其标签选项为“happy”或“sad”。每条数据记录可以包含任意数量的此类片段级标签。

```
标签_模式 = [{ "label_name":"
se
4
{#
```

更新模式后，她在一个小组件中获取所有带有中性标签的记录。为了突出显示并标注文本片段，她进入单条视图（如图 4 所示）。在迭代的任何阶段，她都可以查看仪表板小组件，以监控项目进度。经过几轮类似的迭代后，她觉得效果不错，便结束了探索。最后，Meggie 可以将标注后的数据导出为 JSON 或 CSV 格式，用于训练，也可以直接接入模型。

总而言之，借助 MEGAnno，Meggie 可以使用各种基于启发式的方法或自动搜索功能探索她的数据集，并在标注过程中更好地理解数据语料库。随着项目的推进，她可以灵活地迭代更新模式。借助小组件，Meggie 可以在同一个 Jupyter 笔记本中完成整个机器学习生命周期。

4 相关工作

存在许多可支持自然语言处理任务的标注工具，Neves 和 Seva (2019年、2020年) 对这些工具进行了全面综述。本节重点介绍与我们所提出的灵活、探索性、高效且无缝的框架更为接近的研究。

灵活的模式 遗憾的是，现有的大多数工具都不是为迭代开发模式而设计的，因此无法充分适应不断演进的项目。 Felix 等人 (2018年) 和 Kulesza 等人 (2014年) 允许用户通过检查分配给各类别的文档，逐步定义文档类别。但这些研究更接近交互式主题建模或交互式分类：用户将文档分配给类别；而非文档标注：用户为文档分配标签。我们的工具不止于文档标签细化，还支持逐步定义更广泛的标注模式任务（例如，额外收集片段级标签）。

探索式标注 这一概念由 Felix 等人(2018年) 提出，指运用计算技术帮助用户将文档归入不断演变的标签。在本文中，我们使用“探索式标注”一词，指在数据理解/探索循环中迭代进行探索性数据分析和数据标注。探索式标注之所以有益，是因为用户在标注数据的同时，可以加深对数据集的理解（Sun 等人，2017年）。

高效批次/批量标注 少数工具提供了同时为一条记录中的多个文本片段分配标签的功能。例如，YEDDA（Yang 等人，2018年）可以通过命令行为多个文本片段标注片段级标签。TALEN（Stephen Mayhew，2018年）是一款命名实体标注工具，具有实体传播功能，可一次性标注文档中某个实体的所有提及。相比之下，用户可以使用我们的 Python API 和图形用户界面小部件同时标注多条记录。

笔记本小部件数据分析师经常使用计算笔记本，迭代地编写和编辑代码，以理解数据、检验假设并构建模型（Head 等人，2019年；Randles 等人，2017年）。借鉴 Mage 的做法（Kery 等人，2020年），Mage 为特定任务扩展了 Jupyter 笔记本，加入了图形用户界面小部件。我们的组件旨在实现与其余机器学习开发代码的灵活交互。以 Jupyter 小部件形式实现的标注工具包括 Pigeon（pig）和 ipyannotate（ipy），但它们仅提供简单的标签分配功能。

5 结论

本文介绍 MEGAnno，这是一个面向自然语言处理研究人员和从业者设计的标注框架。借助 MEGAnno 的程序化接口和交互式小组件，用户可以在 Jupyter 笔记本中迭代地探索并搜索感兴趣的数据子集、标注数据、训练模型、评估和调试模型，而无需承担上下文切换或数据迁移的开销。

参考文献

humanloop.com。https://humanloop.com/。

ipyannotate。https://github.com/ipyannotate/ipyannotate。

pigeon. https://github.com/agermanidis/pigeon。

Tivadar Danka 和 Peter Horvath。modAL：一个适用于 Python 的模块化主动学习框架。可在 arXiv 上查阅：https://arxiv.org/abs/1805.00979。

Cristian Felix、Aritra Dasgupta 和 Enrico Bertini。2018年。探索式标注助手：对大型文档集合进行混合主动式标签整理。载于 第 *31* 届 *ACM* 用户界面软件与技术年度研讨会论文集，UIST '18，第 153–164 页，美国纽约州纽约市。计算机协会。

Crowdflower Data for Everyone 数据库。Twitter 美国航空公司情感数据集（第 4 版）。

R. Stuart Geiger、Dominique Cope、Jamie Ip、Marsha Lotosh、Aayush Shah、Jenny Weng 和 Rebekah Tang。2021年。重新审视"垃圾进，垃圾出"：机器学习应用论文报告了哪些关于人工标注的训练数据的信息？ *QuantitativeScienceStudies*，第 1–33 页。

Andrew Head、Fred Hohman、TitusBarik、Steven M. Drucker 和 Robert DeLine。2019年。管理计算笔记本中的混乱，第 1–12 页。计算机协会，美国纽约州纽约市。

Fred Hohman、Kanit Wongsuphasawat、Mary Beth Kery 和 Kayur Patel。2020年。理解并可视化机器学习中的数据迭代。载于《*2020*年计算机协会人因计算系统会议论文集，第 1–13 页。

Matthew Honnibal 和 Ines Montani。2017年。spaCy 2：借助 Bloom 嵌入、卷积神经网络和增量解析实现自然语言理解。即将发表。

Mary Beth Kery、Donghao Ren、Fred Hohman、Dominik Moritz、Kanit Wongsuphasawat 和 Kayur Patel。2020年。Mage：在计算笔记本中，让代码与图形化工作流畅切换。Todd Kulesza、Saleema Amershi、Rich Caruana、Danyel Fisher 和 Denis Charles。2014年。结构化标注，助力机器学习中的概念演化。载于《*SIGCHI* 人机交互系统中的人为因素会议论文集》，CHI '14，第3075–3084页，美国纽约州纽约市。计算机协会。

Ines Montani 和 Matthew Honnibal。2018年。Prodigy：一种全新的标注工具，助力实现极高效的机器教学。 人工智能，即将发表。

Mariana Neves 和 Jurica Seva。2019年。 文档人工标注工具综述。《生物信息学简报》，22(1):146–163。

Mariana Neves 和 Jurica Seva。2020年。 标注工具：可搜索的标注工具目录。

James Pustejovsky 和 Amber Stubbs。2012年。自然语言机器学习标注：应用领域语料库构建指南。" O'Reilly Media, Inc."。

Sajjadur Rahman 和 Eser Kandogan。2022年。剖析文本信息抽取工作流的实践、局限与机遇：人在回路中的视角。CHI '22，美国纽约州纽约市。计算机协会。

Bernadette M. Randles、Irene V. Pasquetto、Milena S. Golshan 和 Christine L. Borgman。2017年。将 Jupyter 笔记本用作开放科学工具：一项实证研究。载于 *2017*年 *ACM/IEEE* 数字图书馆联合会议（*JCDL*），第 1–2 页。

Nils Reimers 和 Iryna Gurevych。2019年。Sentence-BERT：使用孪生 BERT 网络的句子嵌入。载于 *2019*年自然语言处理实证方法会议论文集。计算语言学协会。

Dan Roth、Stephen Mayhew。2018年。TALEN：用于标注低资源实体的工具。载于*ACL*系统演示。

Yunjia Sun、Edward Lank 和 Michael Terry。2017年。Label-and-learn：可视化数据标注过程中机器学习分类器的成功可能性。载于 第*22*届智能用户界面国际会议论文集，IUI '17，第 523–534 页，纽约，美国纽约州。计算机协会。

Laurens Van der Maaten 和 Geoffrey Hinton。2008年。使用 t-SNE 可视化数据。《机器学习研究杂志》，9(11)。

Jie Yang、Yue Zhang、Linwei Li 和 Xingxuan Li。2018年。YEDDA：一款轻量级协作式文本片段标注工具。收录于 *ACL 2018* 论文集，

系统演示，第31至36页，墨尔本，澳大利亚。
计算语言学协会。