

关系抽取智能化语料标注系统

报告人:张敏学

合作人:袁禹

刘亚

The background of the slide is split diagonally from the top-left to the bottom-right. The upper-left portion is white, and the lower-right portion is dark gray with a repeating pattern of lighter gray circles.

1. 背景

关系抽取

从文本中识别实体并抽取实体之间的语义关系

例句:比尔盖茨是微软公司的创始人

抽取结果:(比尔盖茨, 创始人, 微软公司)

现存的问题

主流方法是监督学习模型,但是需要大量标注好的数据

数据标注的过程需要大量重复劳动

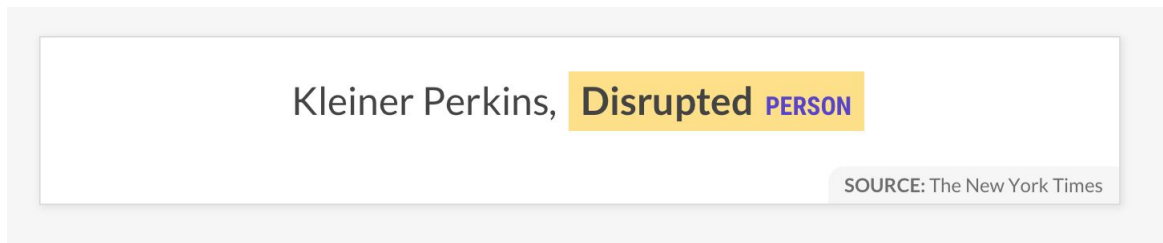
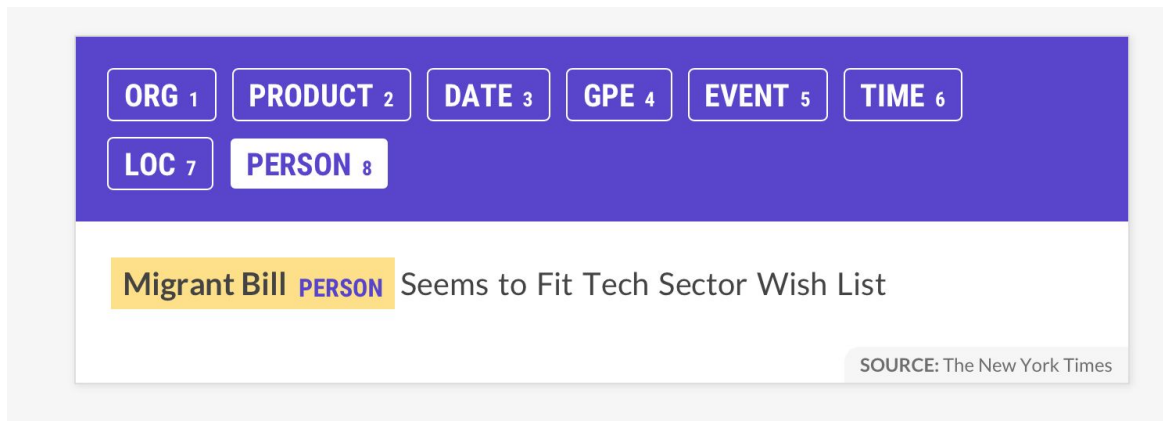
需要一个智能化的标注工具,减少重复劳动,提高效率



2. 已有工具调研

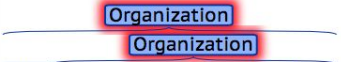

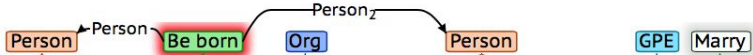
Prodigy

- 界面舒适, 操作便捷
- 智能化标注
- 不支持关系标注



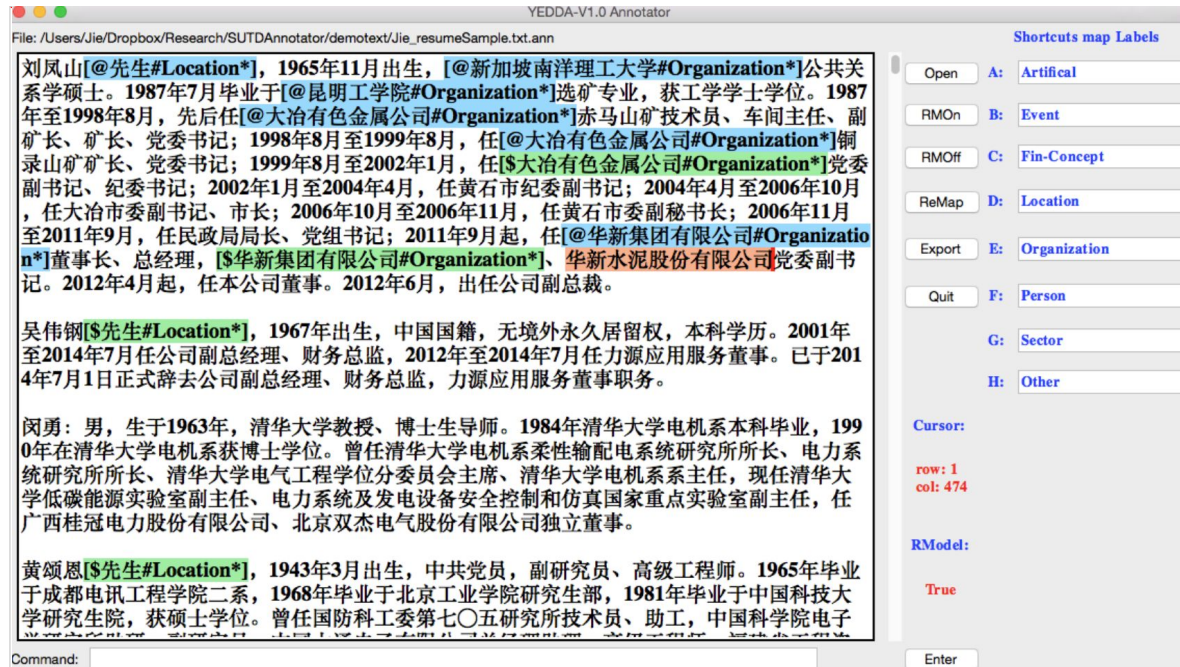
brat

- 可以自定义标注结构
- 界面不简洁
- 不支持智能化标注

1	MemTest1
2	Uh I watched their TV show.
	
3	The Captain and Tennille show.
5	No.
6	I don't.
	
7	I don't remember when it was on.
8	I don't remember.
	
9	I just remember watching it, and the music and hearing it on the radio.
10	That's about it.

SUTDAnnotator

- 界面简洁
- 不支持智能化标注
- 不支持关系标注



标注工具小结

	界面友好程度	是否支持关系标注	是否包含智能化
Prodigy	友好	不支持	有
brat	不友好	支持	无
STUDAnnotator	友好	不支持	无



3. 动机

动机

我们想要做一个标注系统，它拥有以下几个特点：

- 支持关系标注
- 标注界面友好
- 标注过程背后含有智能算法，将人工重复劳动降到最低



4. 需求

目标

尽可能地减少标注者的工作量

需求

有选择性地标注

每个句子的标注, 有算法预标注的功能

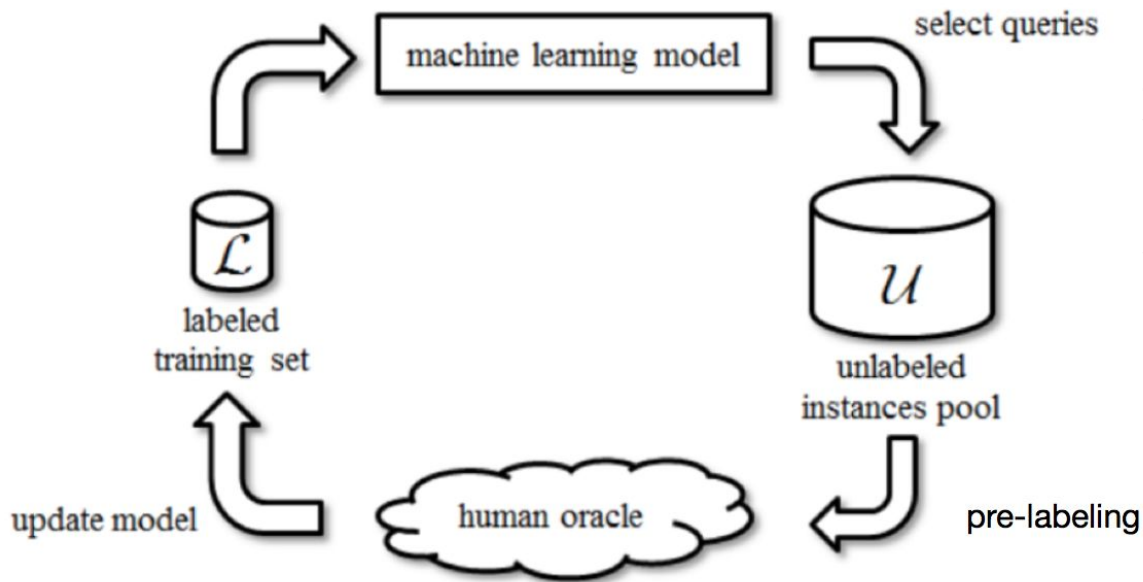
随着标注量的增加, 预标注算法可以迭代



5. 框架

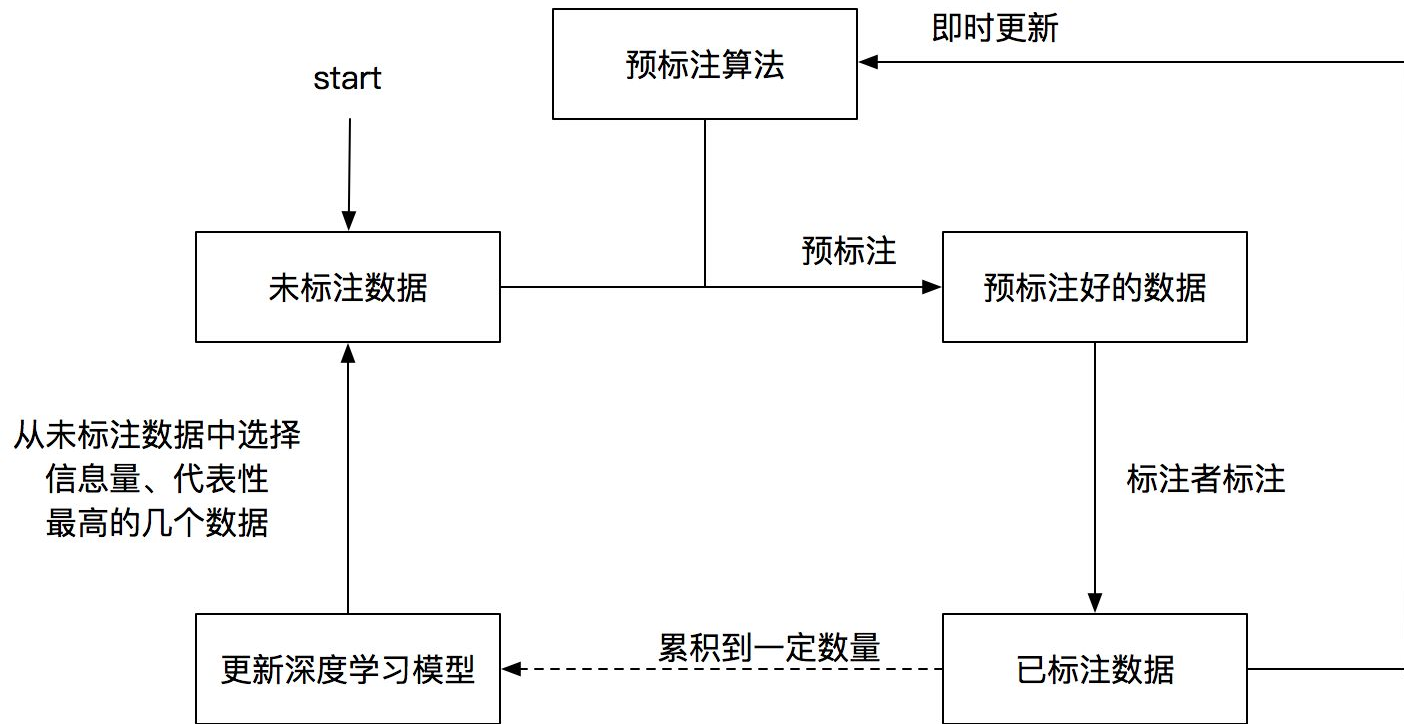
标注算法框架

假设有一个专家，可以为模型提出的 query 提供真实的标签



可能存在语义漂移现象，导致选出来的句子都是一个领域内的

模型迭代过程



要选出足够多的数据让人来标注，要是数据标完模型还没更新完就尴尬了

算法框架

Online 部分: 更新速度快、准确率较高

- 结构化 SVM 模型

Offline 部分: 准确率高

- 深度学习模型

模型要具有鲁棒性

系统框架

