# 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



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

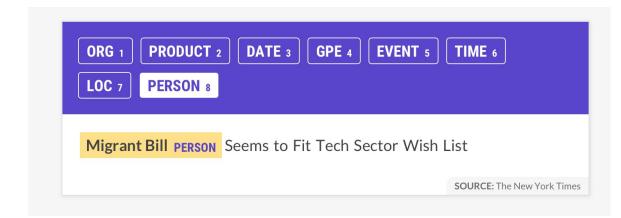
数据标注的过程需要大量重复劳动

需要一个智能化的标注工具,减少重复劳动,提高效率



### Prodigy

- 界面舒适,操作便捷
- 智能化标注
- 不支持关系标注

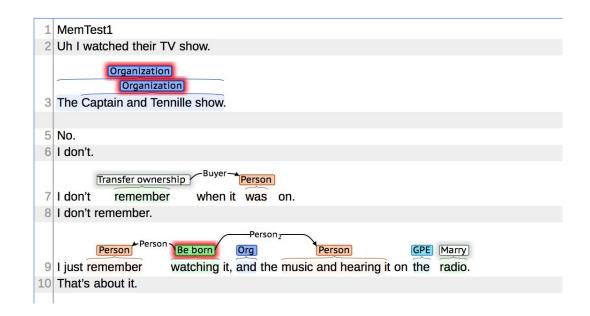


Kleiner Perkins, Disrupted Person

SOURCE: The New York Times

#### brat

- 可以自定义标注结构
- 界面不简洁
- 不支持智能化标注



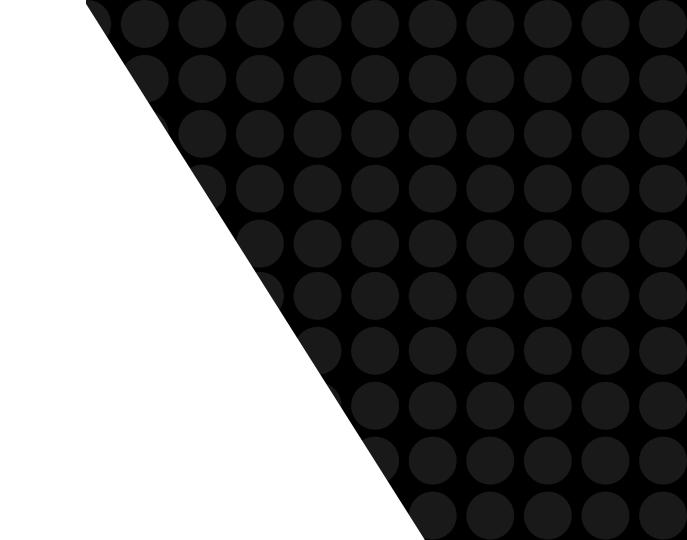
#### **SUTDAnnotator**

- 界面简洁
- 不支持智能化标注
- 不支持关系标注

File: /Users/Jie/Dropbox/Research/SUTDAnnotator/demotext/Jie_resumeSample.txt.ann		Shortcuts map Labels
刘凤山[@先生#Location*],1965年11月出生,[@新加坡南洋理工大学#Organization*]公共关	Open	A: Artifical
系学硕士。1987年7月毕业于[@昆明工学院#Organization*]选矿专业,获工学学士学位。1987年至1998年8月,先后任[@大冶有色金属公司#Organization*]赤马山矿技术员、车间主任、副	RMOn	B: Event
矿长、矿长、党委书记; 1998年8月至1999年8月,任[@大冶有色金属公司#Organization*]铜录山矿矿长、党委书记; 1999年8月至2002年1月,任[\$大冶有色金属公司#Organization*]党委	RMOff	C: Fin-Concept
副书记、纪委书记;2002年1月至2004年4月,任黄石市纪委副书记;2004年4月至2006年10月	ReMap	D: Location
,任大冶市委副书记、市长; 2006年10月至2006年11月,任黄石市委副秘书长; 2006年11月 至2011年9月,任民政局局长、党组书记; 2011年9月起,任[@华新集团有限公司#Organizatio	(Format )	P. Occasionica
n*]董事长、总经理, [\$华新集团有限公司#Organization*]、 <mark>华新水泥股份有限公司</mark> 党委副书记。2012年4月起, 任本公司董事。2012年6月, 出任公司副总裁。	Export	E: Organization
此。2012年7月起,在平台内里手。2012年6月,山区台内附心城。	Quit	F: Person
吴伟钢[\$先生#Location*], 1967年出生,中国国籍,无境外永久居留权,本科学历。2001年至2014年7月任公司副总经理、财务总监,2012年至2014年7月任力源应用服务董事。已于201		G: Sector
4年7月1日正式辞去公司副总经理、财务总监,力源应用服务董事职务。		H: Other
闵勇: 男,生于1963年,清华大学教授、博士生导师。1984年清华大学电机系本科毕业,199	Cursor:	
0年在清华大学电机系获博士学位。曾任清华大学电机系柔性输配电系统研究所所长、电力系   统研究所所长、清华大学电气工程学位分委员会主席、清华大学电机系系主任,现任清华大	row: 1	
学低碳能源实验室副主任、电力系统及发电设备安全控制和仿真国家重点实验室副主任,任	col: 474	
广西桂冠电力股份有限公司、北京双杰电气股份有限公司独立董事。	RModel:	
黄颂恩[\$先生#Location*],1943年3月出生,中共党员,副研究员、高级工程师。1965年毕业	True	
于成都电讯工程学院二系,1968年毕业于北京工业学院研究生部,1981年毕业于中国科技大   学研究生院,获硕士学位。曾任国防科工委第七○五研究所技术员、助工,中国科学院电子	1100	
Note that the religion of the	Enter	

### 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

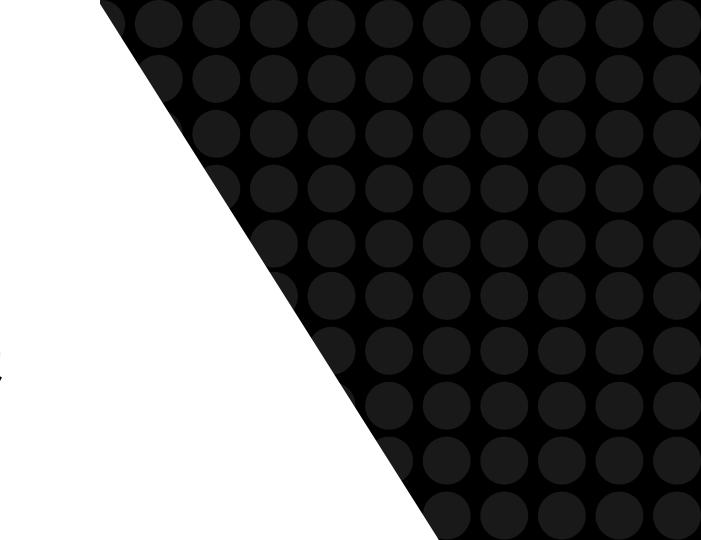


3. 动机

### 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

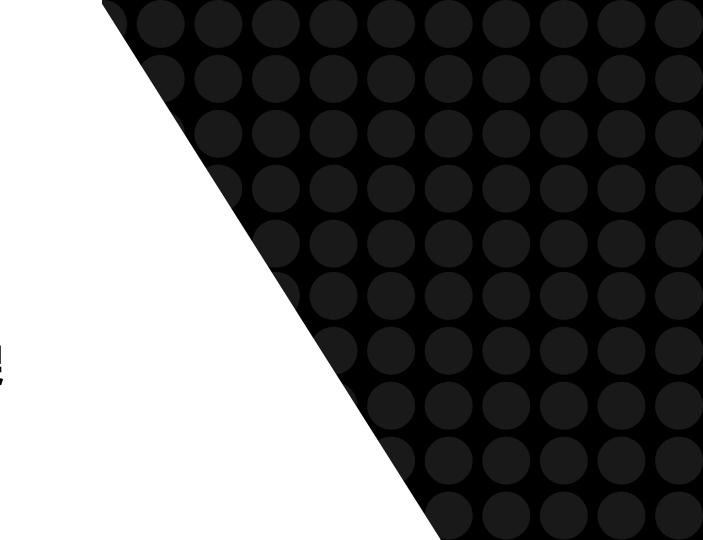
尽可能地减少标注者的工作量

#### 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

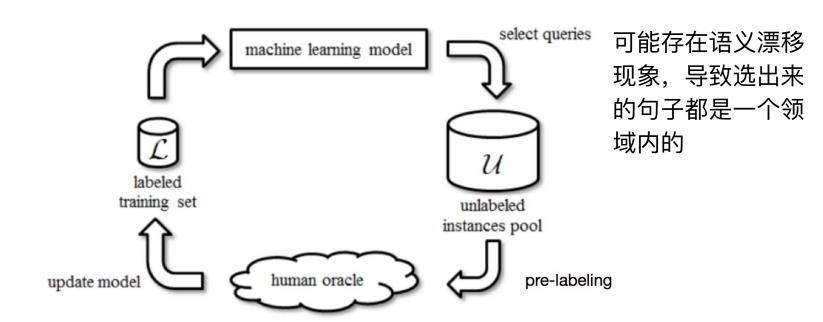
随着标注量的增加, 预标注算法可以迭代



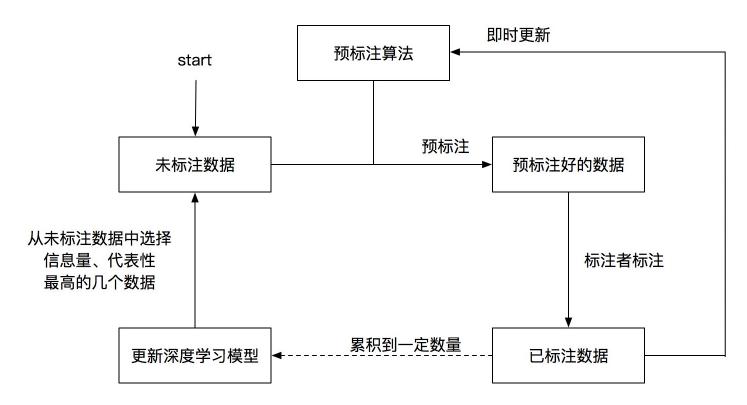
5. 框架

#### 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



#### 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

