基于众包训练数据的中文实体标注研究

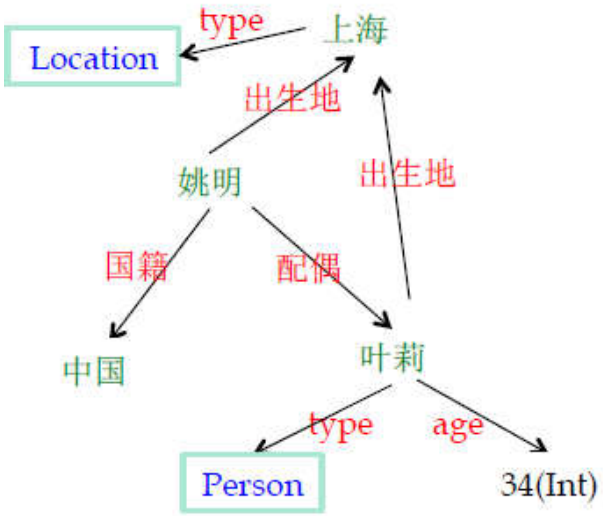
陈文亮

苏州大学人类语言技术研究所(SUDA-HLT)

2017-12

知识图谱

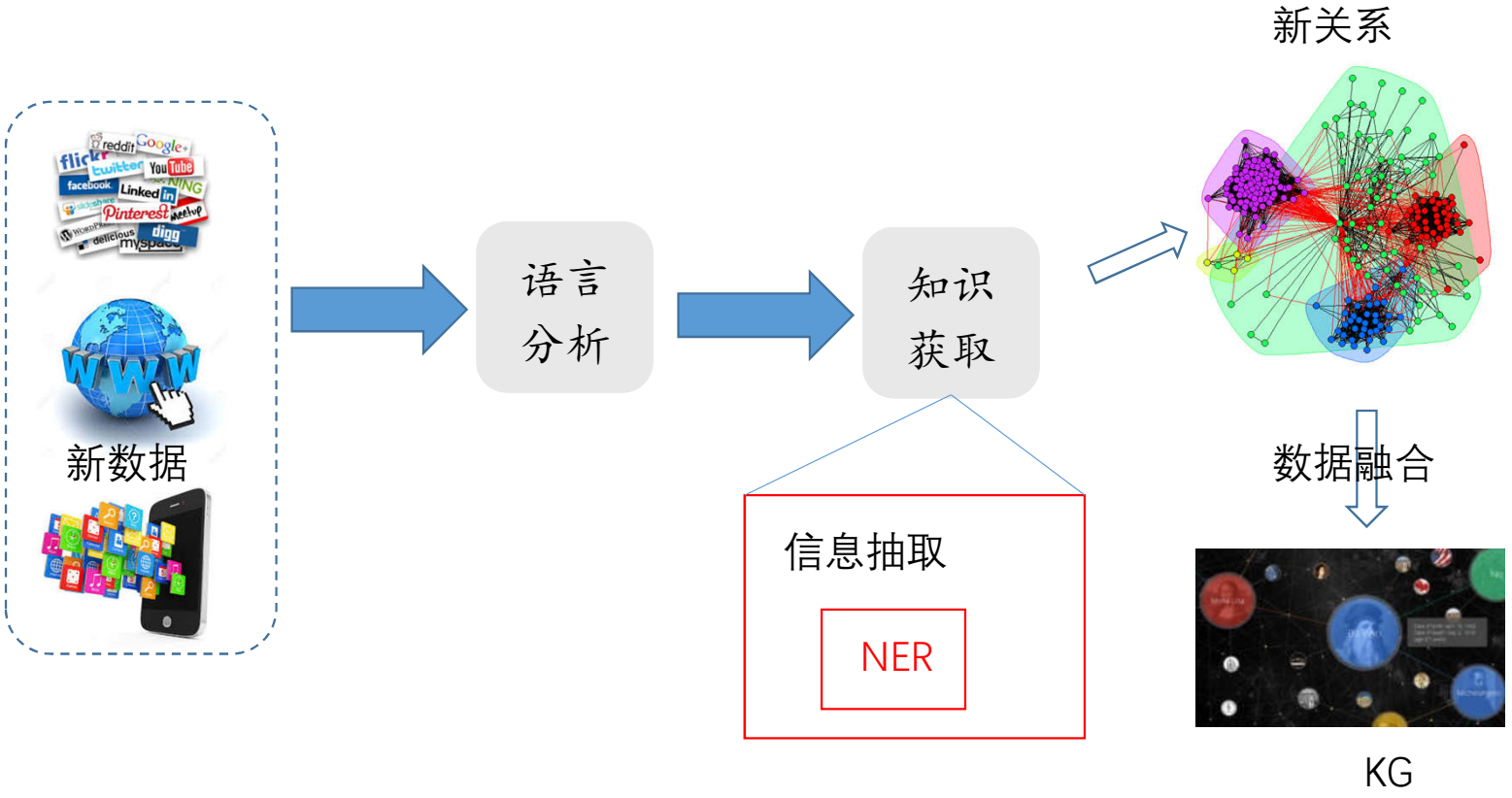
- 知识图谱本质上是一种语义网络。其结点代表实体(entity)或者概念(concept)，边代表实体/概念之间的各种语义关系。



知识图谱


Google
Search
超过5.7亿实体
超过18亿条事实（关系）
规模庞大
ProBase
2,653,873概念
Market
European Markets
Emerging Markets
China
Developing countries
Newly Industrialized Countries
India
百度知心
搜狗知立方

本报告涉及内容



主要内容

- 噪音训练数据
- 众包NER数据
- 基于众包数据的NER研究进展
- 总结

噪音训练数据

- NLP系统构建
  - 给定一个NLP任务
  - 通常需要训练语料，理想是100%正确语料
- 专家语料
  - LDC分词语料/北大分词语料一致性都低于99%
- 常见人工语料
  - 一致性更差

场景1：多快糙省构建人工语料

- 任务：新领域/新任务
- 例子：互联网文本处理
- 文本种类多、数量大
  - 微博、微信。。。
  - 论坛帖子，如百度贴吧、水木社区
  - 用户评论文本
  - 博客
  - 。。。

场景1：多快糙省构建人工语料

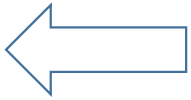
- 在处理互联网文本面临的挑战
  - 现有语言分析工具性能下降的很快
  - 互联网文本通常没有人工标注语料
- 专家标注
  - 代价高，速度慢
  - 在新领域中，有时候不得不标一些新语料
  - 少、慢、好、贵
- 众包数据
  - 非专家标注员快速完成语料标注，包含大量噪音
  - 多、快、糙、省

场景2：现有一个列表，如何构建新系统

- 任务：有一个实体表/KB关系表，构建能识别类似实体的系统
- 例子：识别歌名，现有一个歌名表，要求识别句子中的歌名
- 问题：一般都缺乏标注语料
- 远程监督数据
  - 使用现有KB自动生成训练语料，也包含大量噪音

NER系统构建场景

- 研究课题
  - 有一定规模的人工标注训练语料
  - 常见领域：新闻领域
  - 常见类别：人名、地名、组织机构名等
  - 目标：构建在测试集上表现很好的系统
- 实际应用
  - 新领域：电商领域、对话领域、金融领域等
  - 新类别：产品、品牌、歌名等
  - 目标：构建在新领域里面还算能用的系统
  - 问题：通常无人工标注训练语料



招人，标语料

标注数据

- 专家标注员（适用于不计成本的主）
  - 对标注规范了如指掌，且有耐心标注
  - 优点：标注质量高
  - 缺点：难找且贵
- 普通标注员（适用于精打细算的主）
  - 对标注规范粗通（能花15分钟阅读规范就是好标注员）
  - 优点：数量多，便宜
  - 缺点：标注质量较低

标注数据

- 有些缺钱但希望拥有高质量语料的研究者

- 中间路线：N名普通标注员+1~2名专家
  - 完美结合：专家负责解决难题，普通人解决简单题
  - 预算合适：一群便宜的+几位贵的
  - 标注速度：应该是很快的
  - 多、快、好、省

- 为了这个美好路线，苏大设计SNAP系统

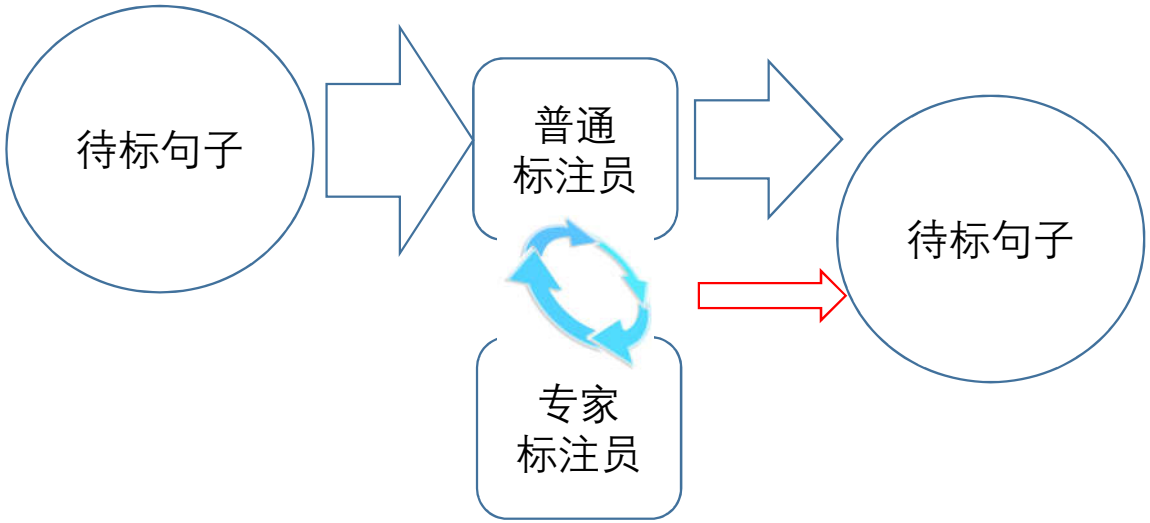
SNAP标注系统

- 苏州大学SNAP标注系统
  - 任务类型：分类任务、序列标注任务和句法标注任务
  - 序列标注任务：NER、分词、词性
  - 浏览器模式：支持多人同时标注
  - 质量控制：
    - 随机多人普通标注员标注
    - 专家审核标注不一致
    - 投诉机制
    - 权威专家确定答案
  - 标注员评价
    - 地雷审核
    - 反馈学习专家意见

SNAP标注系统（Demo）

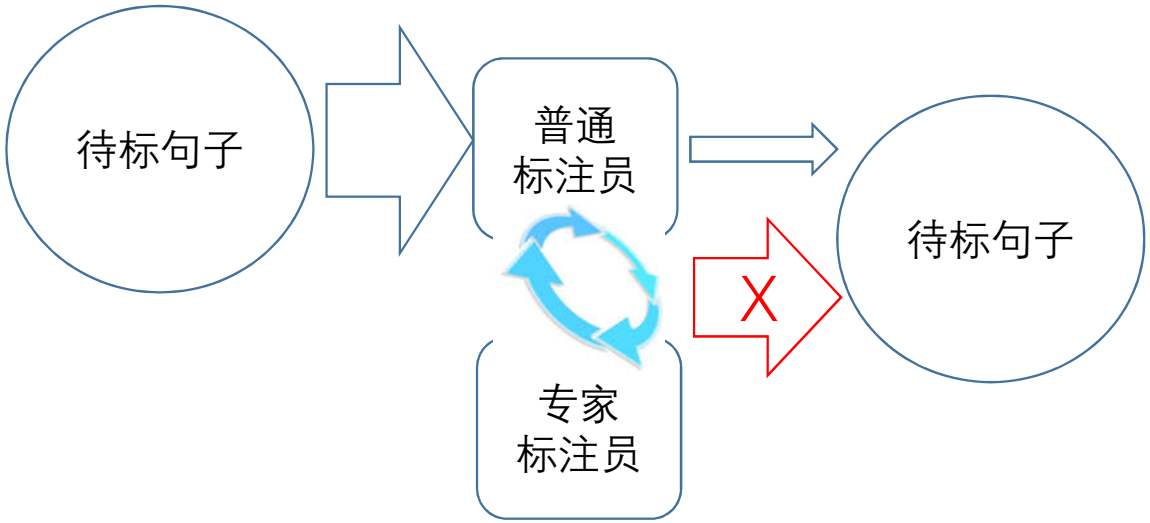
数据标注：理想 VS 现实

- 理想：句子-> 普通标注员 -> 偶尔求助专家标注员 ->完美收工



数据标注：理想 VS 现实

- 现实：普通标注员 不停的问专家 各种问 -> 专家崩溃



专家崩溃后。。。

- 普通标注员
  - 按照自己的理解标注完任务
  - 领着报酬，愉快地走了
- 留下众包标注数据
  - 数据规模是很大的
  - 有些标注挺好的，但有很多是有冲突的
  - 专家对这些结果是不满意的

例子（差别很大）

小白小白，我们玩个成语接龙吧！	
Annotations	**[小白小白]**，我们玩个成语接龙吧！ 小白小白，我们玩个成语接龙吧！ 小白小白，我们玩个成语接龙吧！
你说谢谢的诗意哥哥吗？	
Annotations	你说谢谢的诗意哥哥吗？ 你说**[谢谢]**的诗意哥哥吗？ 你说谢谢的**[诗意哥哥]**吗？

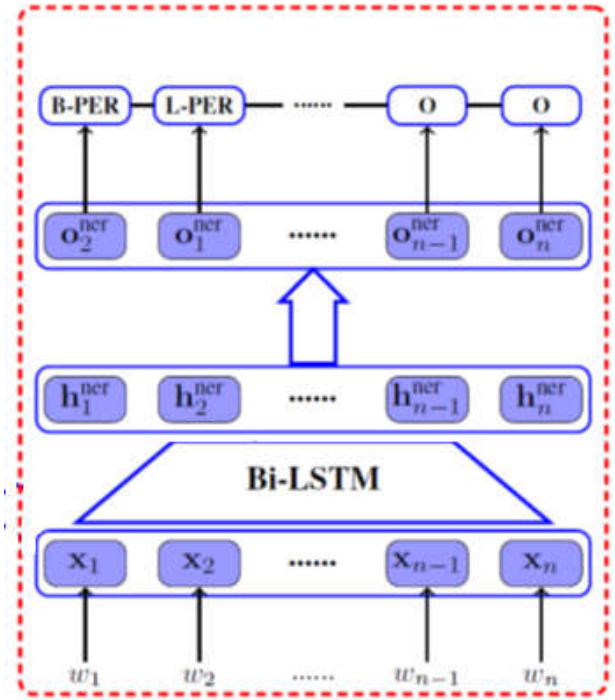
研究者的想法

- 三个臭皮匠可以抵得上一个诸葛亮
- 钱已经花了，留下了众包数据
- 果断抛弃专家，直接用众包数据
- 从众包数据学习一个可用中文NER系统
  - 学习众人取得的共识信息
  - 消解一些相互冲突的标注噪音

IBM的Jelinek：“每当我解雇一个语言学家，语音识别系统的性能就会改善一些。”
(“Every time I fire a linguist the performance of the recognizer improves”.)
某些研究者：当我解雇所有标注专家，在群众帮助下NER系统依然可以改善一些。

简单方法-直接使用

- 假装这个数据是专家标注的
- 直接使用LSMT-CRF训练



简单方法-投票

- 对众包语料采用少数服从多数原则再处理

	小白小白，我们玩个成语接龙吧！
Annotations	[**小白小白**]，我们玩个成语接龙吧！ 小白小白，我们玩个成语接龙吧！ 小白小白，我们玩个成语接龙吧！

小白小白，我们玩个成语接龙吧！

	你说谢谢的诗意哥哥吗？
Annotations	你说谢谢的诗意哥哥吗？ 你说[**谢谢**]的诗意哥哥吗？ 你说谢谢的[**诗意哥哥**]吗？

你说谢谢的诗意哥哥吗？

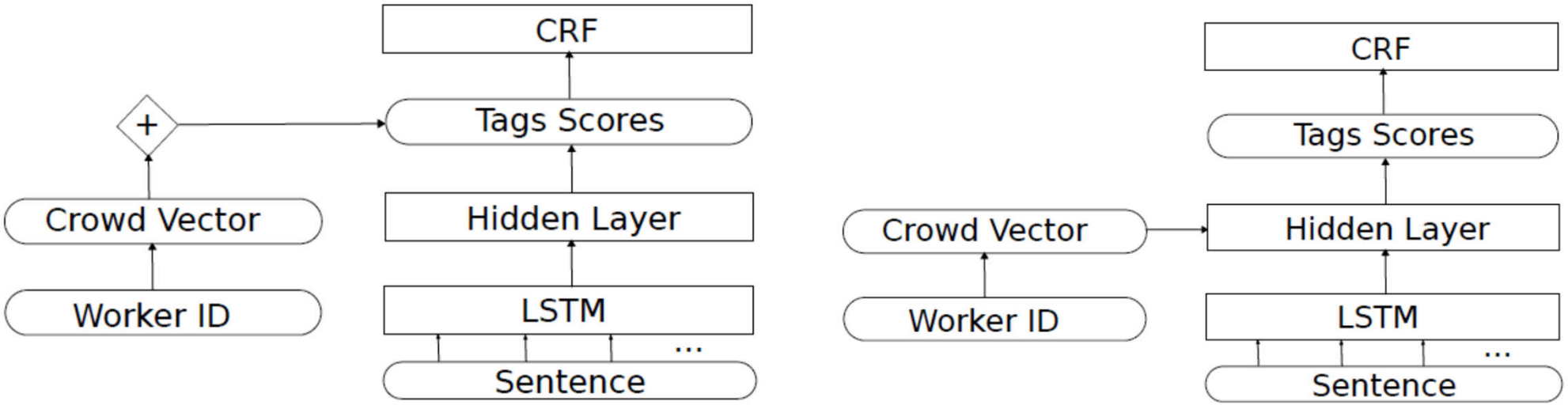
简单方法-投票

- 得到投票后的训练语料
- 直接使用CRF或者LSTM-CRF训练
- 但是效果不好

Model	P	R	F1
CRF	89.48	70.38	78.79
CRF-VT	85.16	65.07	73.77
LSTM-CRF	**90.50**	79.97	84.91
LSTM-CRF-VT	88.68	75.51	81.57

LSTM-crowd

- 把每个标注员都表示为向量
- 问题：测试时无法获得标注员信息



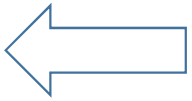
方案一

方案二

(Nguyen et al., 2017)

普通标注员的“特点”

- 快速看实体定义规范（15分钟），比如电商的产品
  - 和自己脑海里面的产品概念进行拟合
- 直接开工
  - 可以快速标注句子
- 每个人由于背景/知识面不同，对规范理解会不同
- 标注员的共性
  - 有些人对鞋子了解的多一些
  - 有些人对衣服了解的多一些
  - 。。。

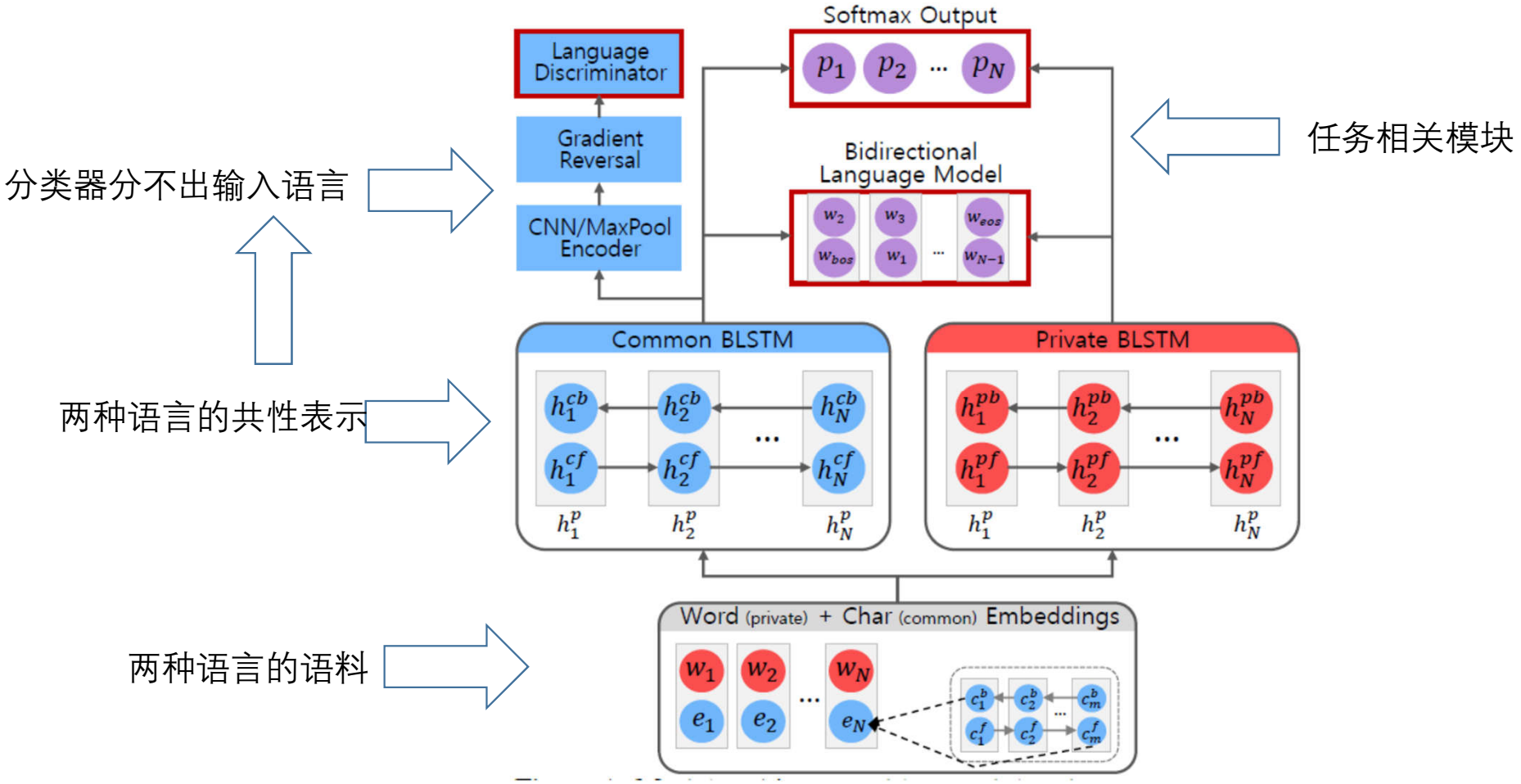


我们专注的对象

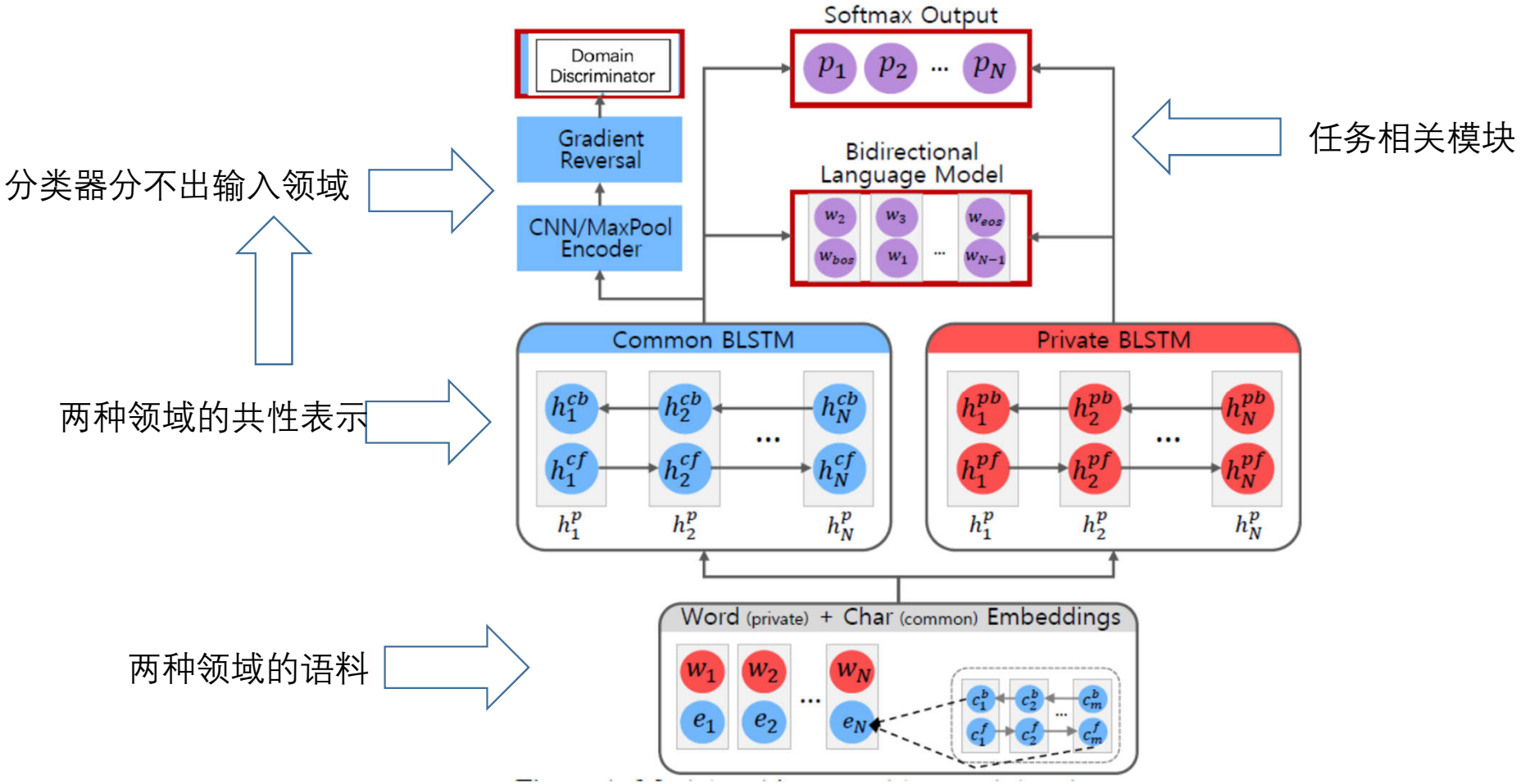
对抗网络

- 适用任务
  - 跨语言(Cross-Lingual)
  - 跨领域(Domain Adaptation)
  - 多任务(Multi-task)
- 通过对抗网络学习
  - 学习不同语言共性
  - 学习不同领域共性
  - 学习不同任务共性

跨语言


Softmax Output
Language Discriminator
Gradient Reversal
CNN/MaxPool Encoder
Bidirectional Language Model
Common BLSTM
Private BLSTM
Word (private) + Char (common) Embeddings
分类器分不出输入语言
两种语言的共性表示
两种语言的语料
任务相关模块

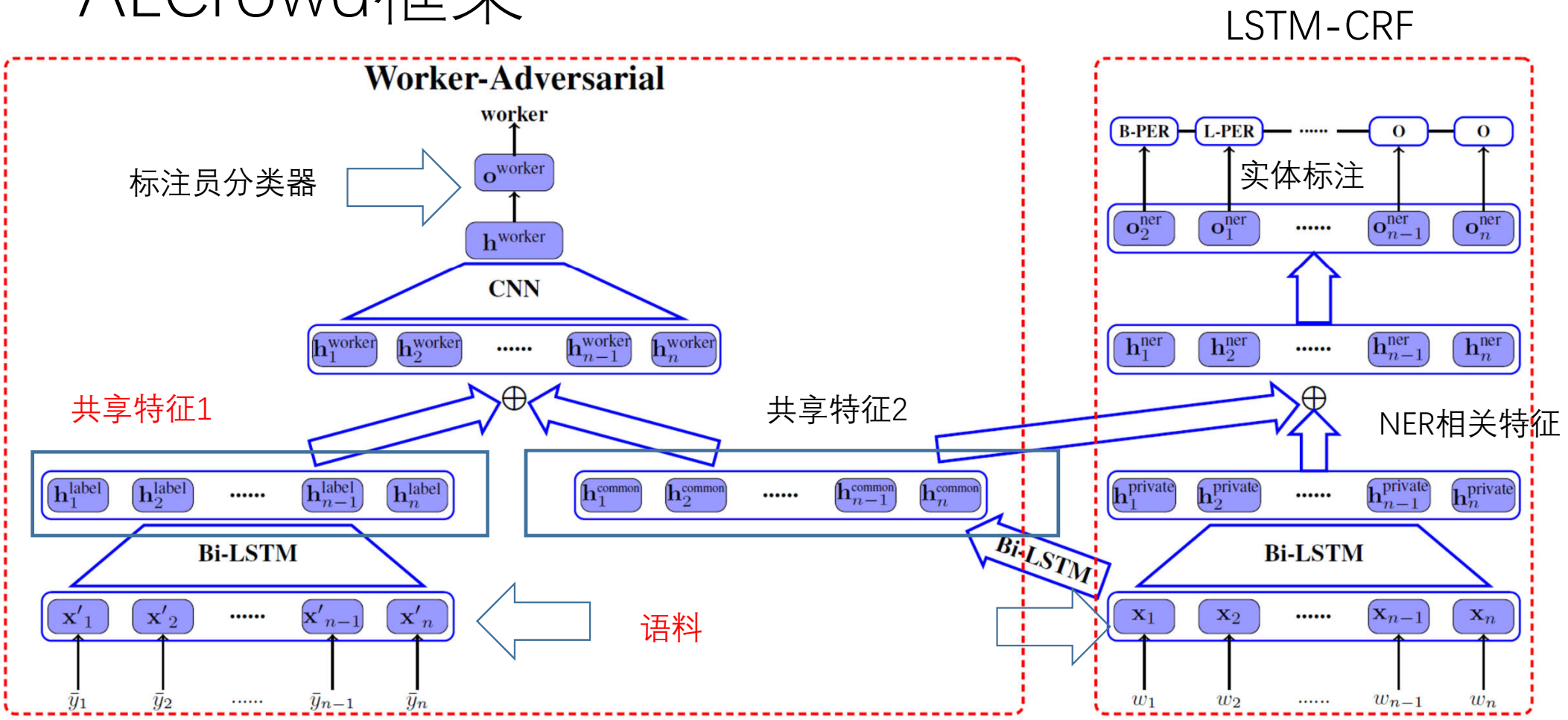
跨领域


Softmax Output
Domain Discriminator
Gradient Reversal
CNN/MaxPool Encoder
Bidirectional Language Model
Common BLSTM
Private BLSTM
Word (private) + Char (common) Embeddings
分类器分不出输入领域
两种领域的共性表示
两种领域的语料
任务相关模块

众包数据学习

- 对抗学习：学习共性
  - 步骤1：输入各个标注员标注的语料
  - 步骤2：学习出来的标注员的“共性”
  - 步骤3：分类器分不清是谁标注的
- 困难
  - 不同领域、不同语言的特征明显
  - 如何区别标注员呢？

ALCrowd框架


LSTM-CRF
Worker-Adversarial
worker
标注员分类器
CNN
共享特征1
共享特征2
Bi-LSTM
语料
Bi-LSTM
B-PER
L-PER
O
O
实体标注
NER相关特征
Bi-LSTM

优化目标

$$\begin{aligned}\mathrm{R}(\Theta, \Theta', \mathbf{X}, \bar{\mathbf{y}}, \bar{z}) &= \mathrm{loss}(\Theta, \mathbf{X}, \bar{\mathbf{y}}) - \mathrm{loss}(\Theta, \Theta', \mathbf{X}) \\ &= -\log p(\bar{\mathbf{y}}|\mathbf{X}) + \log p(\bar{z}|\mathbf{X}, \bar{\mathbf{y}}),\end{aligned}$$

数据

- 数据1：DL-PS
  - 狗尾草公司对话数据
  - 16948句子
  - 标注类别：人名和歌名
  - 43名标注员，每个句子3名标注员
- 数据2：EC-MT/UQ
  - 阿里电商Title和Query
  - 2337句Title和2300句Query
  - 类别：品牌、产品、型号、材料、规格
  - 5名标注员，每个句子2名标注员
- 无标注数据：5M互联网用户生成数据

	#Sent	AvgLen	Kappa
DL-PS	16,948	9.21	0.6033
UC-MT	2,337	34.97	0.7437
UC-UQ	2,300	7.69	0.7529

实验结果：DL-PS

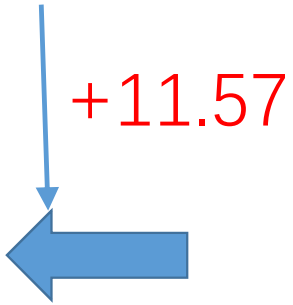
Model	P	R	F1
CRF	89.48	70.38	78.79
CRF-VT	85.16	65.07	73.77
CRF-MA	72.83	**90.79**	80.82
LSTM-CRF	**90.50**	79.97	84.91
LSTM-CRF-VT	88.68	75.51	81.57
LSTM-Crowd	86.40	83.43	84.89
ALCrowd	89.56	82.70	**85.99**

- 众包数据直接用也是可以的
- ALCrowd效果明显(+1.1)

实验结果：EC

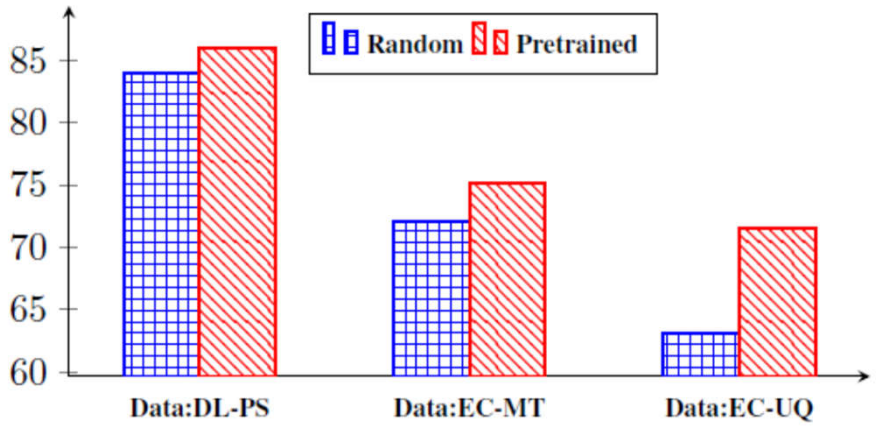
Model	Data: EC-MT			Data: EC-UQ		
	P	R	F1	P	R	F1
CRF	75.12	66.67	70.64	65.45	55.33	59.96
LSTM-CRF	75.02	72.84	73.91	71.96	66.55	69.15
LSTM-Crowd	73.81	**75.18**	74.49	67.51	**71.10**	69.26
ALCrowd	**76.33**	74.00	**75.15**	**74.72**	68.60	**71.53**

+4.51



- 众包数据直接用也是可以的
- ALCrowd效果明显(+1.2 ~ + 2.4)

预先训练的Embeddings的作用



- Pre-trained Embeddings还是很有用的

分析

- 封闭测试（train-train）

Annotations	你说谢谢的诗意哥哥吗？ 你说**[谢谢]**的诗意哥哥吗？ 你说谢谢的**[诗意哥哥]**吗？
Majority-Voting	你说谢谢的诗意哥哥吗？
LSTM-CRF	你说谢谢的**[诗意哥哥]**吗？
ALCrowd	你说**[谢谢]**的**[诗意哥哥]**吗？

- ALCrowd可以较好综合普通标注员的标注结果

结束语

- 专家标注数据质量高，但是不好搞[少慢好贵]
- 普通标注员可以快速得到大规模标注数据[多快糙省]
- 在众包数据上可以构建较好的NER系统
  - 如何充分利用众包数据还有很长的路要走

- 谢谢