Beijing Forest Studio

北京理工大学信息系统及安全对抗实验中心

小样本命名实体识别

林朝坤　硕士研究生

2021年05月30日

内容提要

- 背景简介
- 基本概念
- 算法原理
- 应用总结

预期收获

- 预期收获
  - 1. 了解命名实体识别的研究历史及现状
  - 2. 了解远程监督、小样本学习的基本概念
  - 3. 了解一些少量标注命名实体识别的解决方案

背景简介

背景简介

- 命名实体识别（ Named entity recognition，简称NER ）
  - 又称作“专名识别”，是指识别文本中具有特定意义的实体，主要包括人名、地名、机构名、专有名词等

小明	在	北京大学	的	燕园	看了	中国男篮	的一场比赛。
PER		ORG		LOC		ORG	

背景简介

- 命名实体识别
  - 命名实体共性：
    - 数量无穷
    - 构词灵活
    - 类别模糊
  - 命名实体识别难点：
    - 实体的无穷
    - 歧义多
    - 边界界定困难
      - “李鹏飞起来了”
    - 标注数据缺失

人名

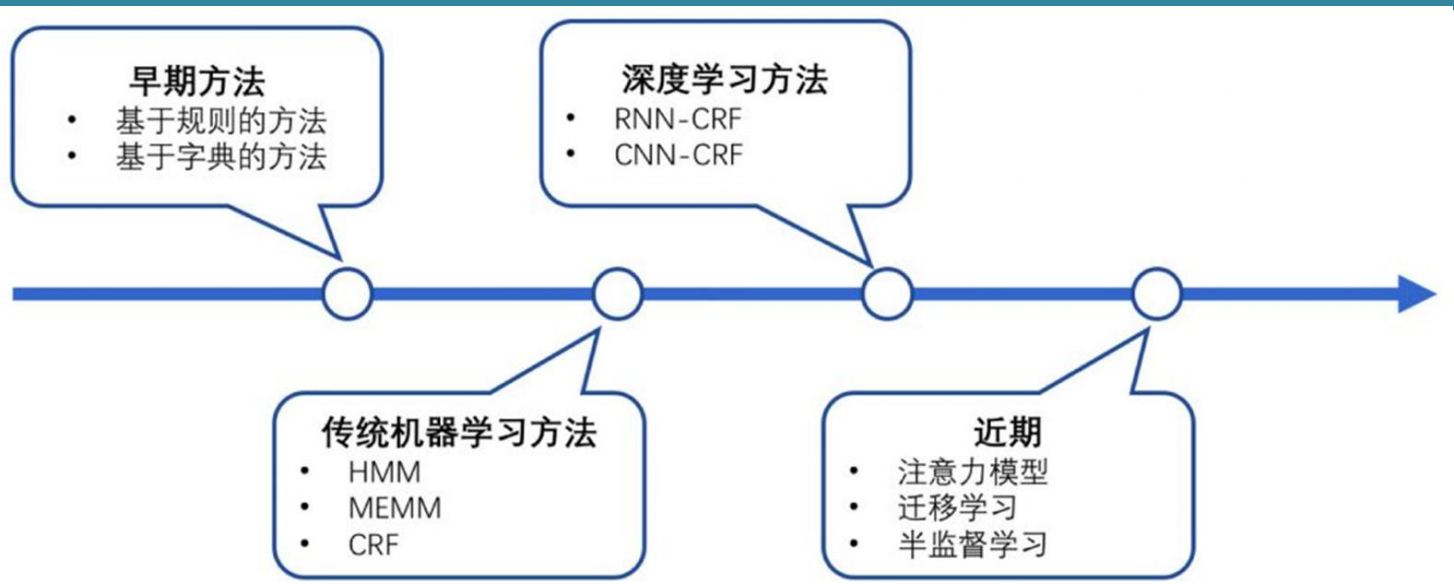
地名

可以用的马桶

马桶里的菊纹检测

背景简介

- NER发展趋势



- 特定领域NER：收集大量的有标签的数据是非常昂贵、困难、甚至不可能。

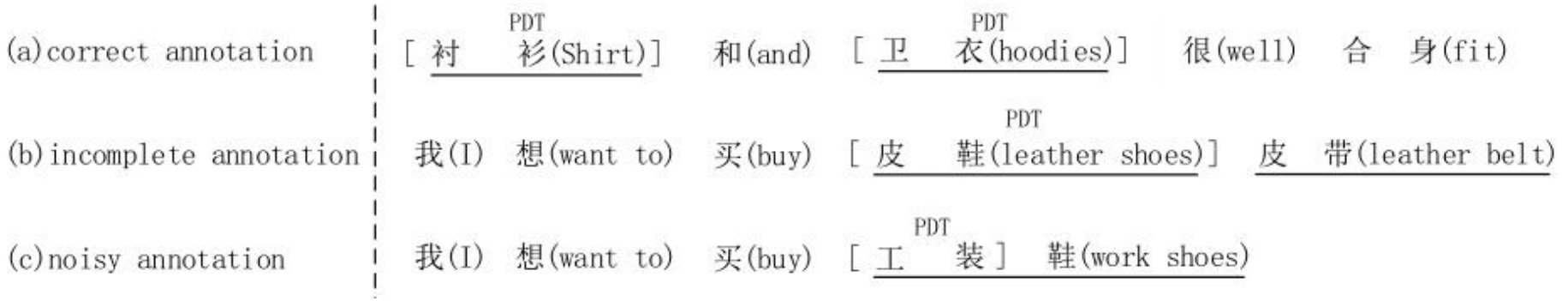
如何进行小样本NER?

半监督学习、迁移学习

基本概念

基本概念

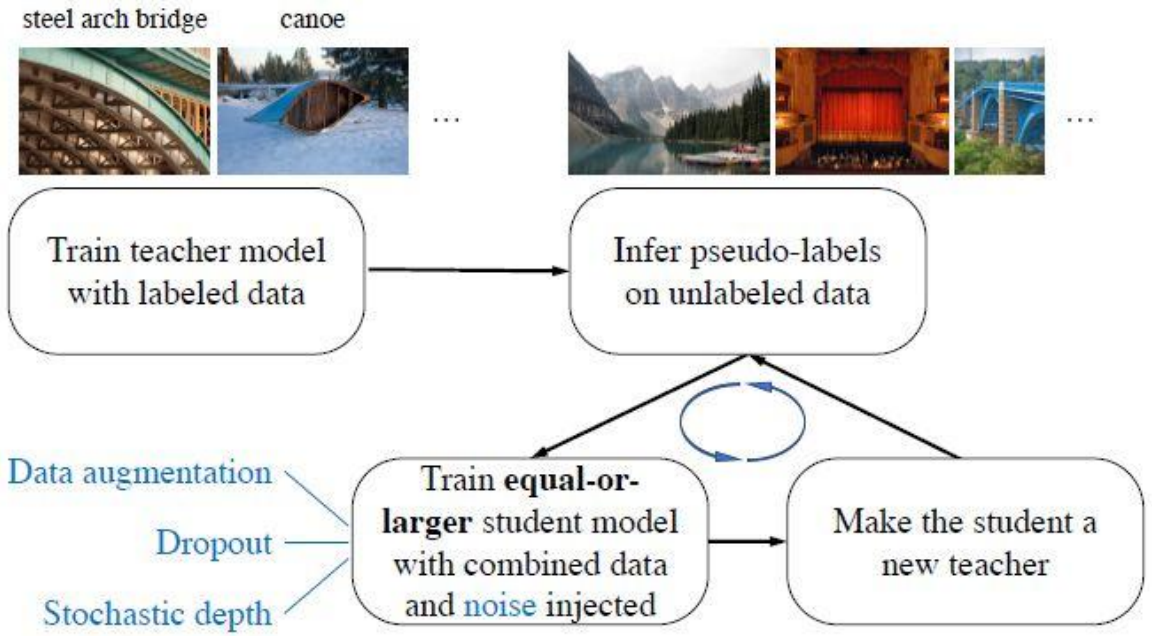
- 远程监督（Distant Supervision）
  - 目的：借助外部知识库为数据提供标签，从而省去人工标注的麻烦。
  - DSNER假设：如果文本中的字符串包含在预定义的实体字典中，则该字符串可能是一个实体。
  - 存在问题
    - 不完整标注
    - 噪声标注



- PU Learning（Positive-unlabeled learning）：在只有正类和无标记数据的情况下，训练二分类器。

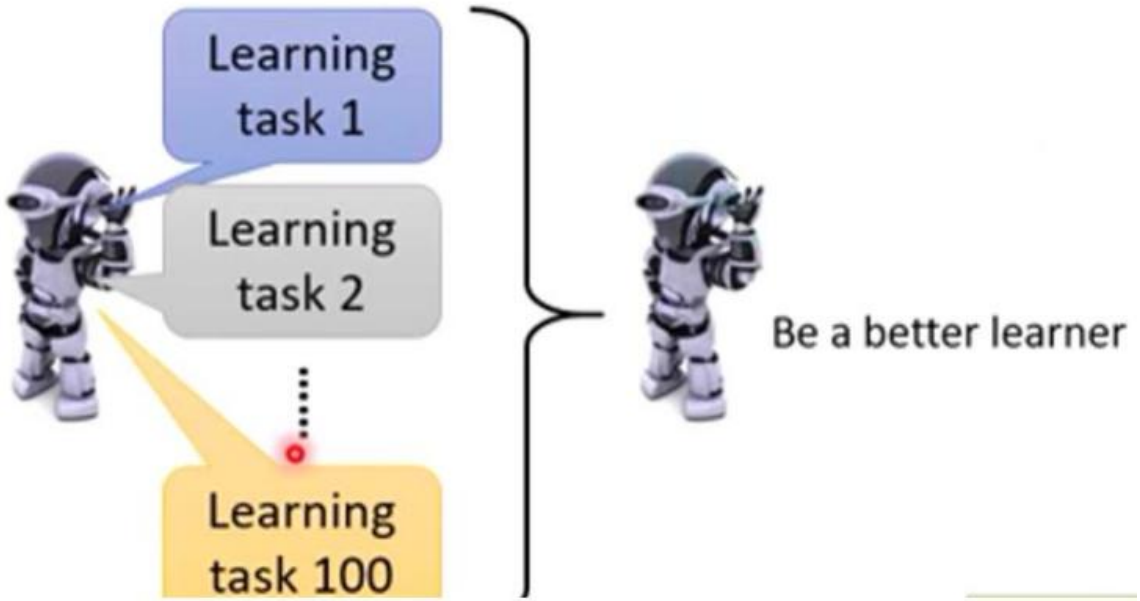
基本概念

- 自训练原理
  - ①首先使用部分标签样本训练一个教师网络；
  - ②使用教师网络生成无标签样本的伪标签；
  - ③使用①中的标签样本和伪标签样本训练学生网络（过程中会加入“噪声”）；
  - ④将学生网络作为新的教师网络，重复步骤②-④。



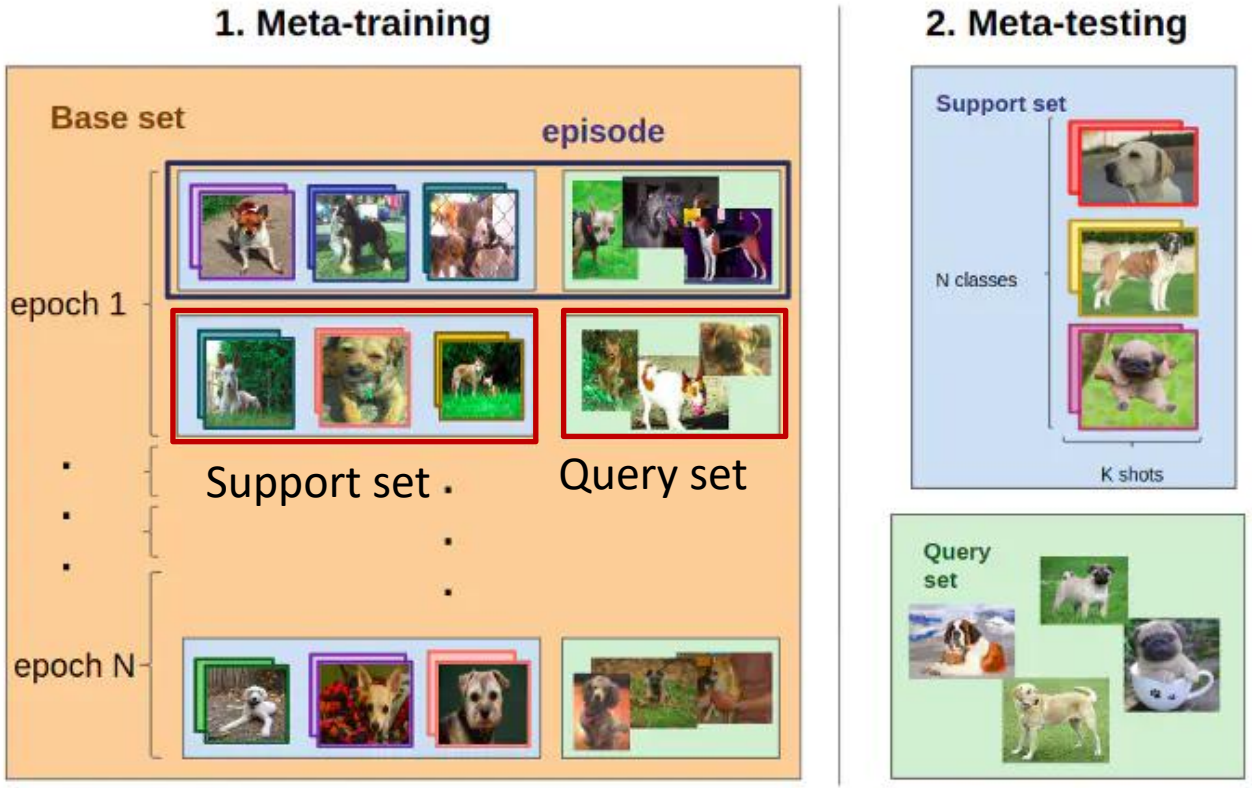
基本概念

- 元学习（ meta-larning）
  - 利用以往的知识经验来指导新任务的学习，具有学会学习的能力，也被称为 learn to learn。
  - 例如：
    - 让Alphago迅速学会下象棋；
    - 让一个猫咪图片分类器，迅速具有分类其他物体的能力；



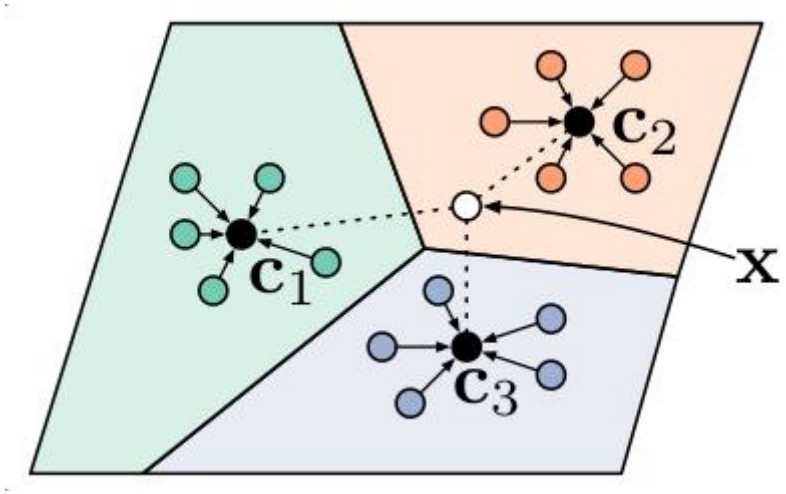
基本概念

- 元学习训练设置
  - 训练单位分层：
    - 第一层训练单位是任务，目的是从每个任务中快速获取知识。
    - 第二层训练单位是每个任务对应的数据，从而缓慢将信息从所有任务中取出并消化。
  - 基本概念
    - episode
    - support set
    - N-way K-shot
    - query set
    - meta training set
    - meta test set



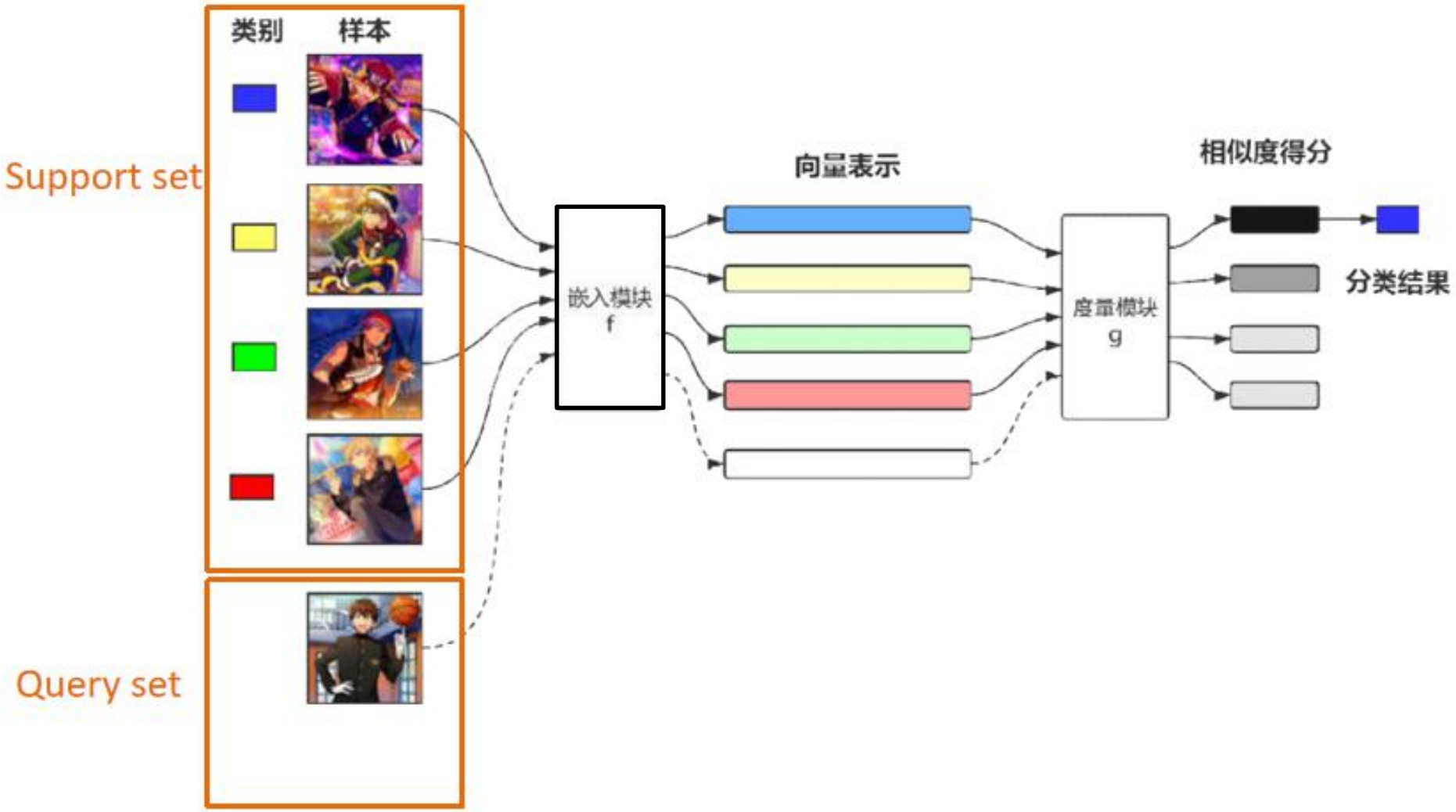
基本概念

- 原型网络的假设: 每个类别都能在空间中对应到一个类比为“原型”的点上，该类的其他数据的表示以这个点为中心分布。
- 原型网络组成是:
  - 嵌入模块：CNN、BiLSTM等；
  - 度量模块：余弦距离、欧式距离等。



基本概念

- 原型网络的原理图



算法原理

算法原理

T	仅使用未标记数据和领域内相关实体字典进行命名实体识别
I	实体字典和未标注语料
P	1、利用字典标注语料； 2、词嵌入； 3、词嵌入输入到BiLSTM中，最后经过sigmoid函数进行分类； 4、损失计算，更新神经网络参数； 5、字典更新；
O	PU分类器

P	远程监督存在噪声标注和不完整标注
C	存在领域实体字典
D	如何减小远程监督语料中不完整标注和噪声标注对NER模型的影响
L	ACL 2019

算法原理

- 步骤：
  - 1. 数据标注：对于实体采用词典进行数据标注，包含在实体字典里的单词标为正例，其余的为未标注数据。
  - 2. 词嵌入：字符级别向量、 GloVe词向量、人工特征向量。

$$\boldsymbol{e}(w) = [\boldsymbol{e}_c(w) \oplus \boldsymbol{e}_w(w) \oplus \boldsymbol{e}_h(w)]$$

  - 3. BiLSTM词编码：$\boldsymbol{e}(w_t|s) = [\overrightarrow{\mathbf{h}}_t \oplus \overleftarrow{\mathbf{h}}_t]$

  - 4. sigmoid函数分类：$f(w|s) = \sigma(\boldsymbol{w}_p^T \boldsymbol{e}(w|s) + b)$

算法原理

- 5. 损失计算：
  - 词w被预测为正类的概率 $f(w|s) = \sigma(\boldsymbol{w}_p^T \boldsymbol{e}(w|s) + b)$
  - 预测风险 $\ell(f(w|s), y) = |y - f(w|s)|.$
  - 经验训练损失

$$\hat{R}_\ell(f) = \pi_p \hat{R}_p^+(f) + \max\left\{0, \hat{R}_u^-(f) - \pi_p \hat{R}_p^-(f)\right\}$$

$$\hat{R}_p^+(f) = \frac{1}{|\mathcal{D}^+|} \sum_{w|s \in \mathcal{D}^+} \ell(f(w|s), 1),$$

$$\hat{R}_p^-(f) = 1 - \hat{R}_p^+(f),$$

$$\hat{R}_u^-(f) = \frac{1}{|\mathcal{D}^u|} \sum_{w|s \in \mathcal{D}^u} \ell(f(w|s), 0),$$

- 6. 利用AdaSampling扩充字典

算法原理

- 实验结果：

Dataset	Type	MEMM	CRF	BiLSTM	BiLSTM+CRF	Matching	bnPU	AdaPU
CoNLL (en)	PER	91.61	93.12	94.21	95.71	6.70	87.21	90.17
	LOC	89.72	91.15	91.76	93.02	67.16	83.37	85.62
	ORG	80.60	81.91	83.21	88.45	46.65	75.29	76.03
	MISC	77.45	79.35	76.00	79.86	53.98	66.88	69.30
	Overall	86.13	87.94	88.30	90.01	44.90	80.74	82.94
CoNLL (sp)	PER	86.18	86.77	88.93	90.41	32.40	84.30	85.10
	LOC	78.48	80.30	75.43	80.55	28.53	73.68	75.23
	ORG	79.23	80.83	79.27	83.26	55.76	69.82	72.28
	Overall	81.14	82.63	80.28	84.74	42.23	74.43	75.85
MUC	PER	86.32	87.50	85.71	84.55	27.84	84.21	85.26
	LOC	81.70	83.83	79.48	83.43	62.82	75.61	77.35
	ORG	68.48	72.33	66.17	67.66	51.60	58.75	60.15
	Overall	74.66	76.47	73.12	75.08	50.12	70.06	71.60
Twitter	PER	73.85	80.86	80.61	80.77	41.33	72.68	74.66
	LOC	69.35	75.39	73.52	72.56	49.74	63.44	65.18
	ORG	41.81	47.77	41.39	41.33	32.38	35.77	36.62
	Overall	61.48	67.15	65.60	65.32	37.90	57.54	59.36

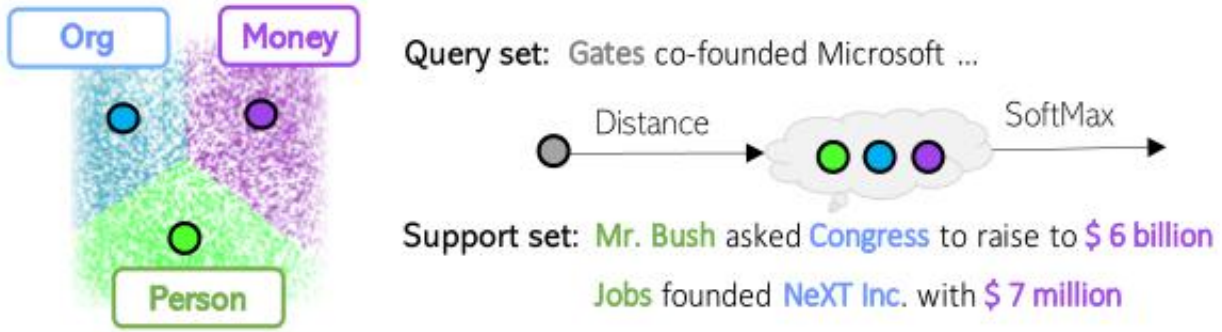
算法原理

T	仅使用少量标注数据进行命名实体识别
I	少量的标注数据
P	原型方法、噪声监督预训练方法、自训练方法
O	命名实体识别模型

P	在少标注情况下如何进行NER?
C	存在少料标注语料
D	如何在仅使用少料标注数据的情况下提升NER模型的性能
L	CoRR 2021

算法原理

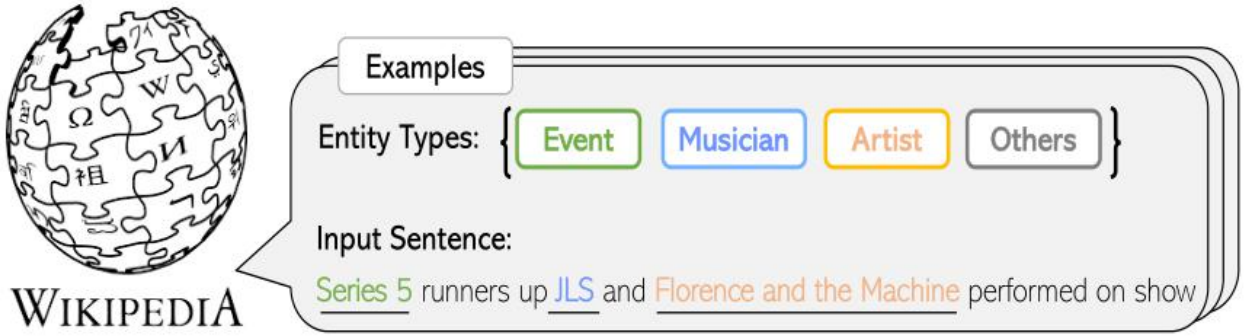
- 原型网络
  - Training阶段：
    - 首先基于Support Set对每一个实体类型构建原型表示(提取所有构成该实体类型的token表示并进行平均加权)
    - 然后对Query Set中的每个token与实体原型表示进行距离度量并分类、计算loss并更新参数。
  - Testing阶段：对每一个新token与实体原型表示进行距离度量，选取最近邻的实体原型标签。



(b) Prototype-based method

算法原理

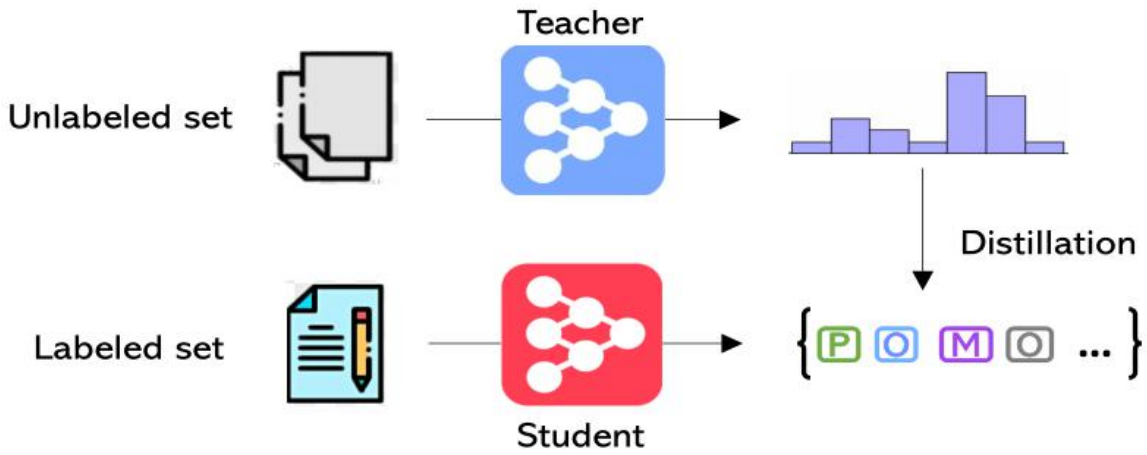
- 噪声监督预训练
  - 利用大规模、带噪的NER标注数据进行“有监督预训练”，提升下游NER任务的鲁棒性和泛化性，是持续预训练的一种改进方式。



(c) Noisy supervised pre-training

算法原理

- 自训练
  - 步骤1: 基于标注数据训练一个NER教师模型
  - 步骤2: 通过NER教师模型对未标注数据预测soft标签
  - 步骤3: 基于“原始标注+soft标注”数据训练NER学生模型
  - 步骤4: 重复步骤2~3多次



(d) Self-training

算法原理-算法执行结果

Datasets	Settings	① LC	② LC + NSP	③ P	④ P + NSP	⑤ LC + ST	⑥ LC + NSP + ST
CoNLL	5-shot	0.535	0.614	0.584	0.609	0.567	**0.654**
	10%	0.855	0.891	0.878	0.888	0.878	**0.895**
	100%	0.919	**0.920**	0.911	0.915	-	-
Onto	5-shot	0.577	0.688	0.533	0.570	0.605	**0.711**
	10%	0.861	**0.869**	0.854	0.846	0.867	0.867
	100%	0.892	**0.899**	0.886	0.883	-	-
WikiGold	5-shot	0.470	0.640	0.511	0.604	0.481	**0.684**
	10%	0.665	0.747	0.692	0.701	0.695	**0.759**
	100%	0.807	**0.839**	0.801	0.827	-	-
WNUT17	5-shot	0.257	0.342	0.295	0.359	0.300	**0.376**
	10%	0.483	0.492	0.485	0.478	0.490	**0.505**
	100%	0.489	0.520	0.552	**0.560**	-	-
MIT Movie	5-shot	0.513	0.531	0.380	0.438	0.541	**0.559**
	10%	0.651	0.657	0.563	0.583	0.659	**0.666**
	100%	**0.693**	0.692	0.632	0.641	-	-
MIT Restaurant	5-shot	0.487	0.491	0.441	0.484	0.503	**0.513**
	10%	0.745	0.734	0.713	0.721	**0.750**	0.741
	100%	0.790	**0.793**	0.787	0.791	-	-
SNIPS	5-shot	0.792	0.824	0.750	0.773	0.796	**0.830**
	10%	0.945	**0.950**	0.879	0.896	0.946	0.942
	100%	0.970	**0.972**	0.923	0.956	-	-
ATIS	5-shot	**0.908**	**0.908**	0.842	0.896	0.904	0.905
	10%	0.883	0.898	0.785	0.896	0.898	**0.903**
	100%	0.953	**0.956**	0.929	0.943	-	-
Multiwoz	5-shot	0.123	0.198	0.219	**0.451**	0.200	0.225
	10%	0.826	0.830	0.787	0.805	0.835	**0.841**
	100%	0.880	**0.885**	0.837	0.845	-	-
I2B2	5-shot	0.360	0.385	0.320	0.366	0.365	**0.393**
	10%	0.855	0.869	0.703	0.762	0.865	**0.871**
	100%	0.932	**0.935**	0.895	0.906	-	-
Average	5-shot	0.502	0.562	0.488	0.555	0.526	**0.585**
	10%	0.777	0.794	0.734	0.758	0.788	**0.799**
	100%	0.833	**0.841**	0.815	0.827	-	-

- 结论
  - 原型方法在NER标注数据量极少的情况下，表现良好；
  - 带噪有监督预训练在众多“少样本NER”设置下表现最佳，可见持续预训练的有效性、同时可提升NER的鲁棒性和泛化性。
  - “带噪有监督预训练+自训练”结合指标可进一步提升。

总结

- 总结
  - 在存在领域内实体字典的时候，可以通过远程监督的方式进行NER；
  - 持续预训练确实对下游任务很有效，针对NER提出的“带噪有监督预训练”确实是一个不错的选择
  - “持续预训练+自训练”是未来解决小样本NER问题的一个有效方案。
- 应用
  - 对话系统
  - 机器翻译
  - 构建知识图谱
  - ……

谢谢！

大成若缺，其用不弊。大盈若冲，其用不穷。大直若屈。大巧若拙。大辩若讷。静胜躁，寒胜热。清静为天下正。