

Beijing Forest Studio  
北京理工大学信息系统及安全对抗实验中心



# 跨语言命名实体识别

硕士研究生 刘小丫

2024年05月12日

- **总结反思**

- 背景知识，基本概念部分铺垫不足
- 讲述太快，缺乏停顿，缺少互动
- 没有结合例子，理论比较抽象

- **相关内容**

- 2023.12.27 马西洋：《平面多标签文本分类方法》
- 2023.08.27 刘小丫：《小样本命名实体识别》
- 2023.05.04 门元昊：《开放式信息抽取技术》

- 预期收获
- 题目内涵解析
- 研究背景与意义
- 研究历史与现状
- 知识基础
- 算法原理
  - ContProto
  - CoLaDa
- 特点总结与工作展望
- 参考文献

- 预期收获
  - 1. 了解跨语言命名实体识别的基本概念
  - 2. 理解跨语言命名实体识别的研究方法
  - 3. 了解跨语言命名实体识别的应用场景和发展方向

- 研究目标
  - 以源语言、目标语言为研究对象，面向命名实体识别任务
  - 结合知识蒸馏、对比学习、原型学习等技术
  - 利用源语言中的标记数据训练的NER模型，迁移到目标语言上的进行NER任务
- 内涵解析
  - 命名实体识别（NER）：定位给定文本中的实体跨度，分类为预定义的实体类型
  - 跨语言命名实体识别（CLNER）：用源语言的标记数据训练NER模型，并在目标语言的测试数据上对其进行评估
    - 零样本跨语言迁移

*Original English:* There have been many tenants, including [The Sunday Times\[ORG\]](#) " and [The Daily News\[ORG\]](#) ".  
有许多租户，包括《[星期日时报\[ORG\]](#)》和《[每日新闻\[ORG\]](#)》。

- 研究背景

- 命名实体识别（NER）是信息提取的一个基本子任务。它被广泛用于许多下游应用，如**关系提取和问答**。
- 目前全球使用的语言总数约为7000多种，收集**足够高质量的标记数据**的成本很高，尤其是对于**低资源的语言**。

- 研究意义

- 利用具有丰富实体标签的高资源（源）语言（如英语）的**知识来克服低（零）资源（目标）语言的数据稀缺问题**。
- 弥合源语言和目标语言的语言结构、特征差异，打破语言障碍，实现信息的跨语言共享。



早期的一些方法，通过元学习、对抗性学习，学习与语言无关的特征，以便在源语言上训练的模型可以直接适应目标语言

Jain提出了一种利用谷歌翻译系统的实体投影算法

Chen等人提出AdvPicker模型，训练语言鉴别器来选择较少依赖语言的未标记数据进行知识蒸馏

Li等人提出MTMT模型，以实体相似性为辅助任务，改进了基于单一任务的知识蒸馏

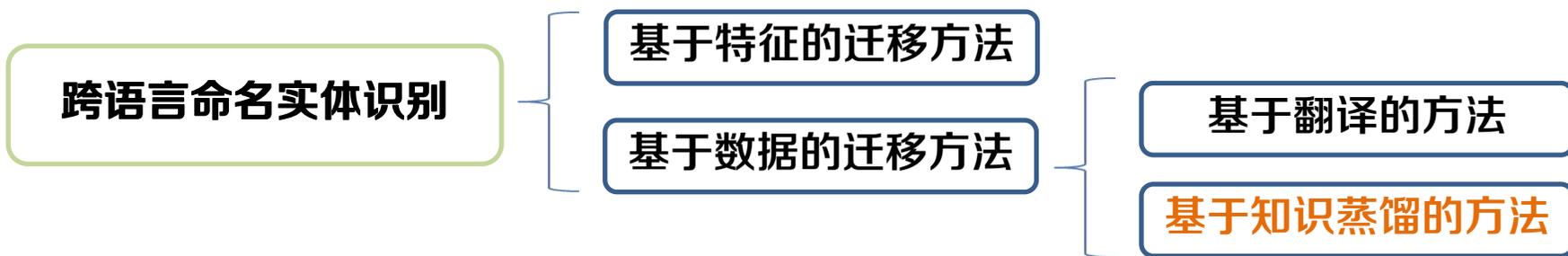


Xie等人提出将源语言数据翻译成目标语言，然后执行从源侧到目标侧的标签投影，使用逐字翻译、短语翻译等

Liu等人提出用预定义的标记翻译句子以进行标签投影

Liang等人提出RIKD模型，设计了一种强化学习算法，根据模型置信度等特征训练实例选择器，迭代选择可靠的伪标签

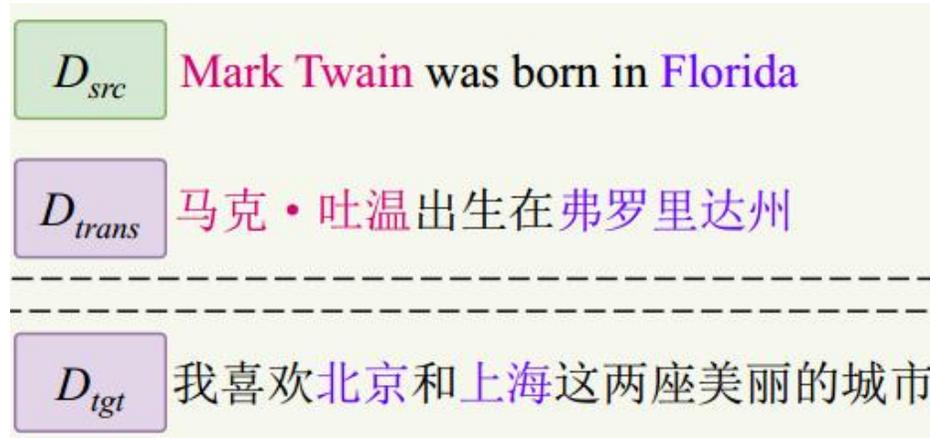
Ge等人提出ProKD改进自训练，采用了传统的序列标记公式，并且只在特定类别的原型上应用对比学习，而不是实际跨度



## 跨语言命名实体识别

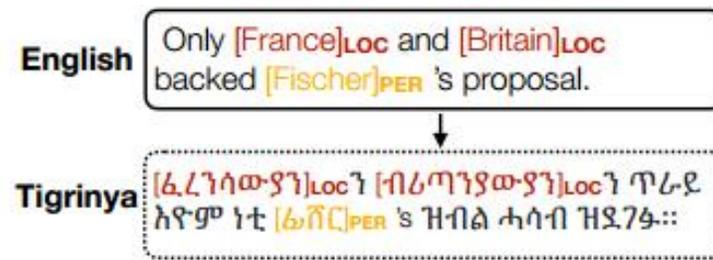
- 问题定义

- 输入文本序列  $x = \{x_1, x_2, \dots, x_n\}$
- NER模型为每个token  $x_i$  分配标签  $y_i$
- 源语言标记训练数据  $D_{src} = \{(x^s, y^s)\}$
- 目标语言未标记数据  $D_{tgt} = \{x^u\}$
- 通过数据投影从  $D_{src}$  获得的翻译数据  $D_{trans}$



- 目标

- 训练一个NER模型  $M$ ，利用这些资源推广到目标语言
- 得到目标语言数据的实体标签  $y^u$



## 知识蒸馏

- 基本概念
  - 使知识从教师网络转移到学生网络
    - 学生网络通过拟合教师网络生成的软标签来优化
- 基本原理
  - 源语言标注数据上训练教师模型

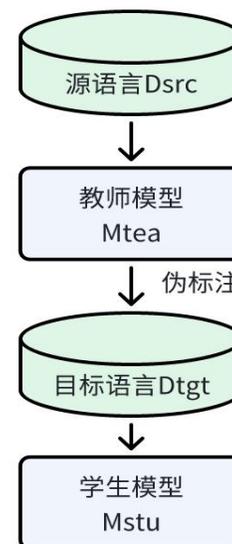
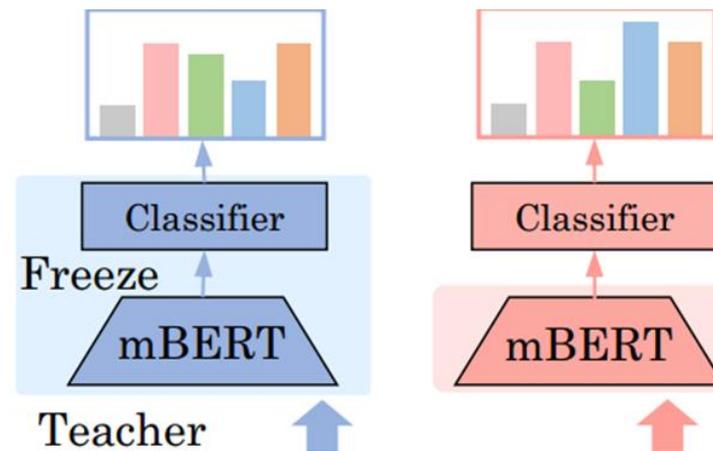
$$L_{src} = -\frac{1}{N} \sum_{X \in D_l^{src}} \frac{1}{|S(X)|} \sum_{s_{jk} \in S(X)} \sum_{c \in \mathcal{C}} y_{jk}^c \log P_{\theta_t}^c(s_{jk})$$

- 教师模型分配软伪标签

$$\hat{y}_{jk} = P_{\theta_t}(s_{jk}) \in \mathbb{R}^{|\mathcal{C}|}$$

- 在伪标签的目标语言数据上训练学生模型

$$L_{tgt} = -\frac{1}{N} \sum_{X \in D_{ul}^{tgt}} \frac{1}{|S(X)|} \sum_{s_{jk} \in S(X)} \sum_{c \in \mathcal{C}} \hat{y}_{jk}^c \log P_{\theta_s}^c(s_{jk})$$





**ContProto**

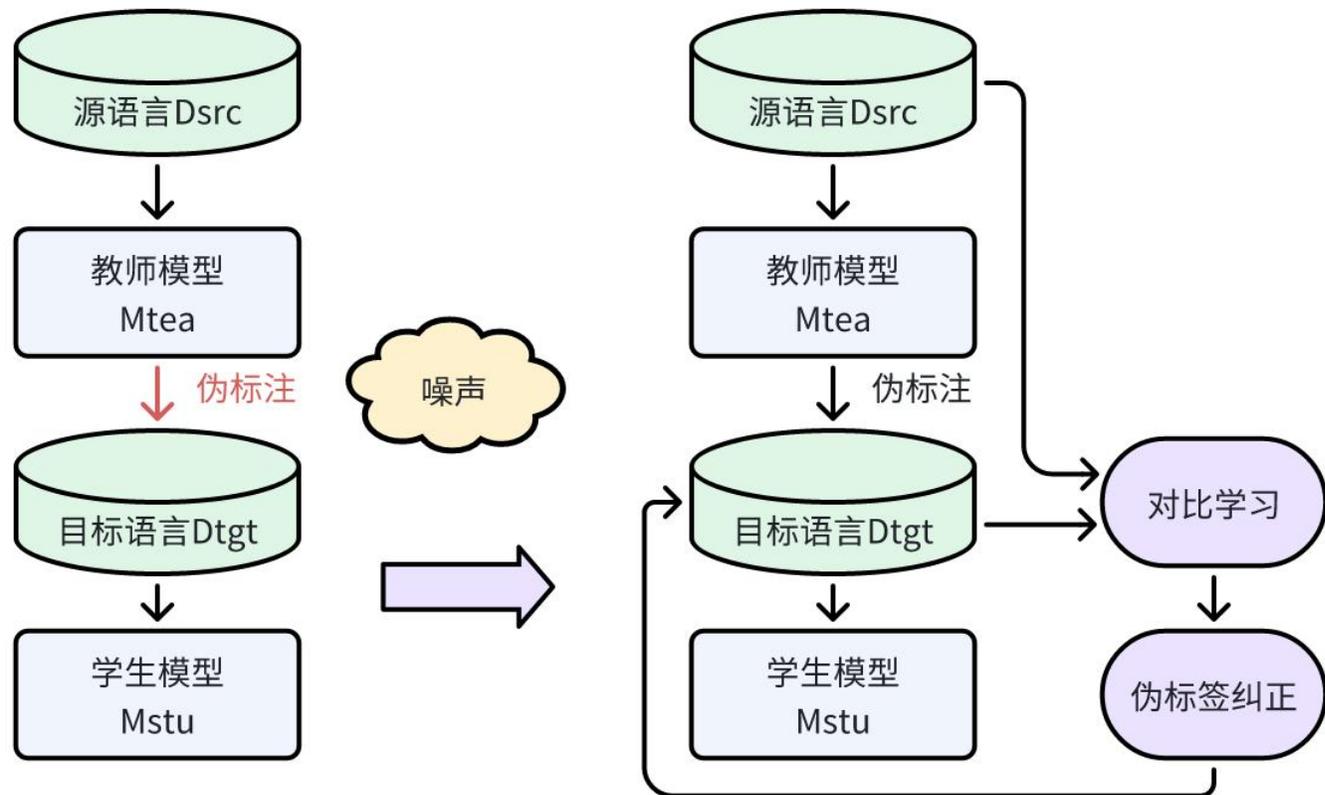
## TIPO

T	目标	用源语言标注数据训练NER模型，在目标语言测试数据上识别命名实体
I	输入	源语言标注数据，目标语言未标注数据
P	处理	1.对比自训练，使用对比学习改进跨度表示 2.基于原型的伪标记，通过原型学习逐渐提高伪标记质量
O	输出	目标语言上的命名实体识别结果

P	问题	自训练通常用于通过对伪标记的目标语言数据进行训练来弥合语言差距。然而，由于目标语言的次优性能，伪标签往往是有噪声的，并限制了整体性能
C	条件	源语言标注数据，目标语言未标注数据
D	难点	增强跨语言可迁移性，有效提高伪标签准确性
L	水平	ACL 2023 CCF-A

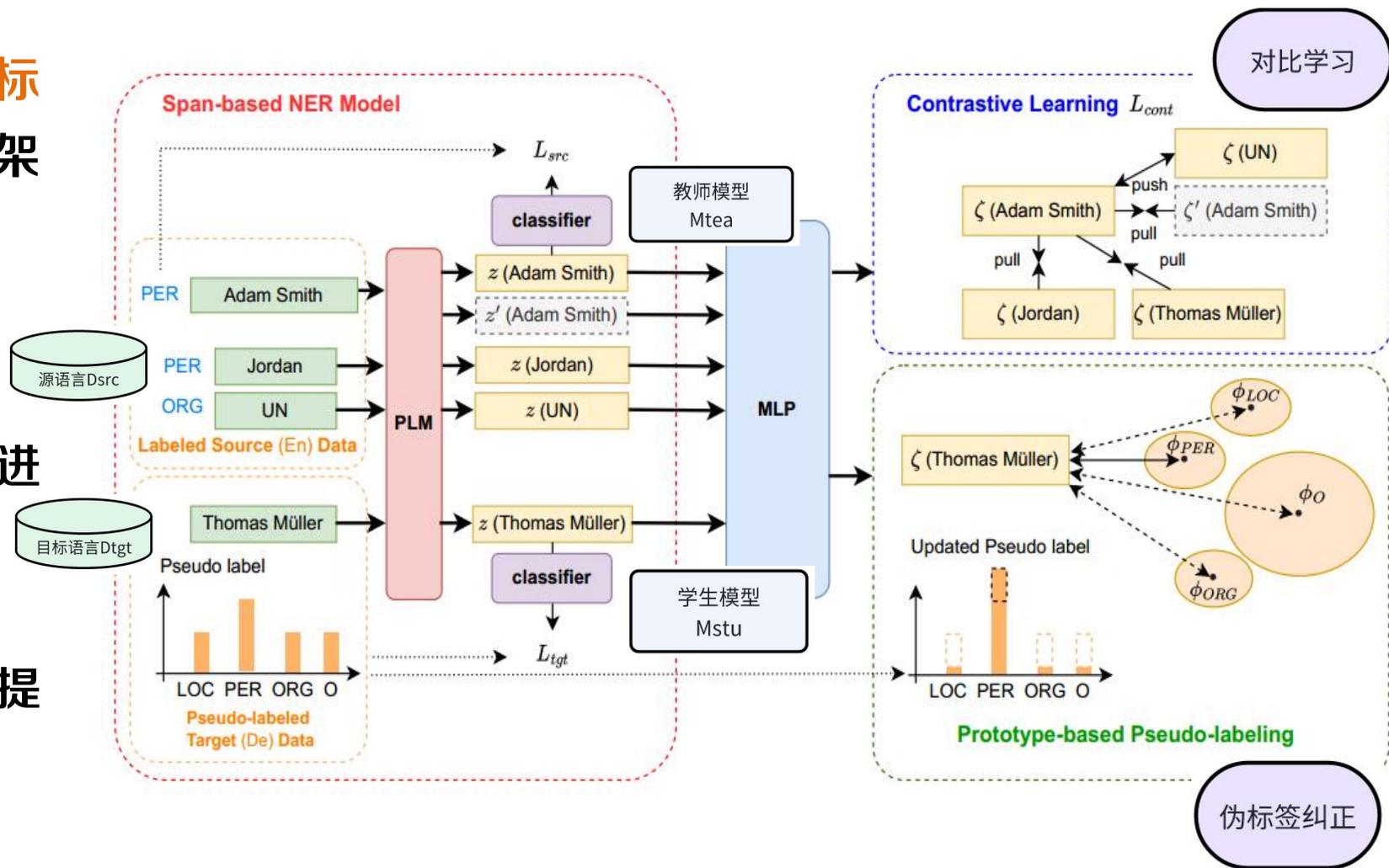
## 算法原理图

- 核心思想
  - 通过将表示学习和伪标签细化结合在一个框架中，提高伪标签质量
- 算法步骤
  - 对比自训练
    - 使用对比学习来改进跨度表示
  - 基于原型的伪标注
    - 通过原型学习逐渐提高伪标注质量



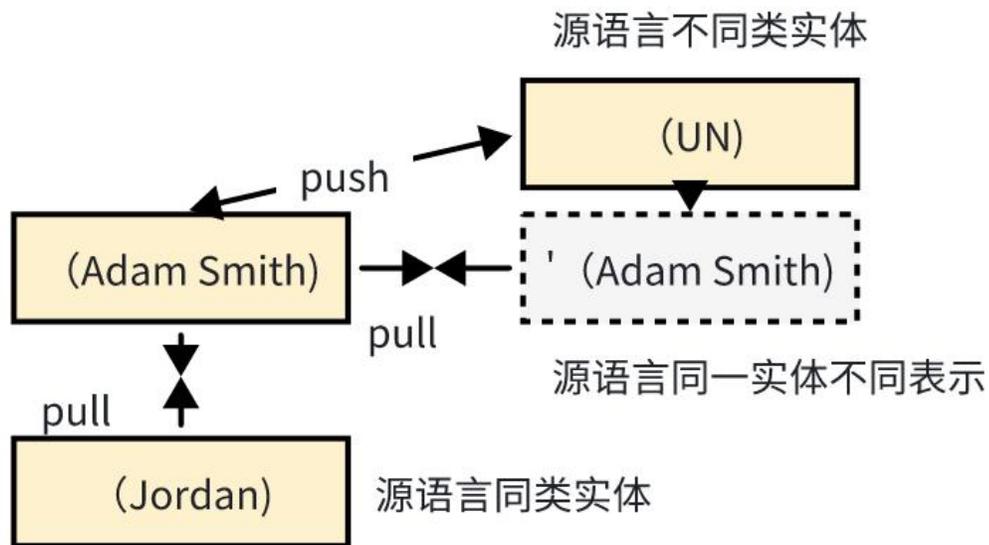
## 算法原理图

- 核心思想
  - 通过将表示学习和伪标签细化结合在一个框架中，提高伪标签质量
- 算法步骤
  - 对比自训练
    - 使用对比学习来改进跨度表示
  - 基于原型的伪标注
    - 通过原型学习逐渐提高伪标注质量



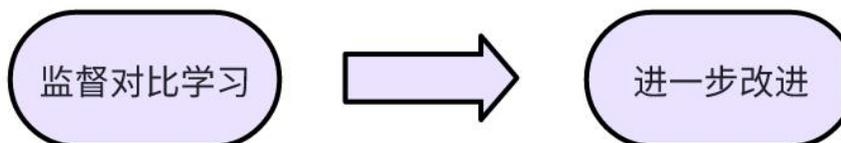
## • 监督对比学习

- 利用标签信息从**同类的源语言样本**中构建**正对**，和其他类样本进行对比
- 生成同一源语言句子的**多个视图**：
  - 每个句子 $X$ 经过NER模型两次（不同随机dropout）
  - 每个跨度 $S_{jk}$ ，产生两个表示 $\zeta_{jk}, \zeta'_{jk}$ ，得到 $\{s_i, y_i, \zeta_i\}_{i=1}^{2m}$ 作为多视图跨度集



将**同类**的源语言实体拉到一起，将**不同类**聚类分开，有利于分类

目标语言?



如何促进跨语言对齐?

源语言 ← ..... → 目标语言

$$L_{cont} = -\frac{1}{2m} \sum_{i=1}^{2m} \frac{1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(\zeta_i \cdot \zeta_p / \tau)}{\sum_{a \in A(i)} \exp(\zeta_i \cdot \zeta_a / \tau)}$$

## 对比自训练

- 伪正对

- 构建**目标语言跨度**的伪正对

- 添加**未标记的目标语言跨度**扩展多视图集
    - 当前训练步骤的**预测**视为其**标签的近似值**

$$y_i = \begin{cases} y_i^{gold} & \text{if } s_i \in D_l^{src} \\ \operatorname{argmax} P_\theta(s_i) & \text{if } s_i \in D_{ul}^{tgt} \end{cases}$$

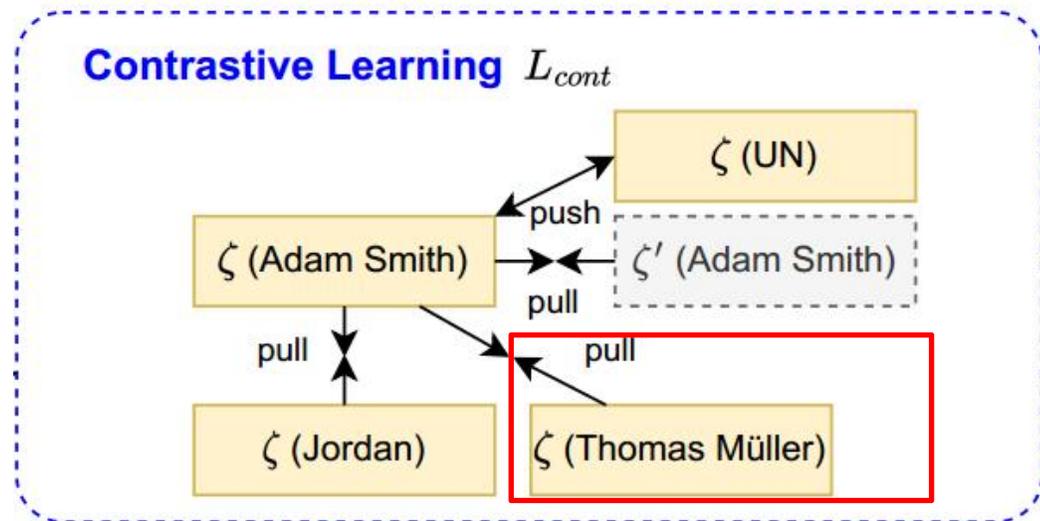
- 例如，PER(person)类的正对组成

- 两个**源语言**PER名称

- 一个**源语言**PER名称和预测为PER的一个**目标语言跨度**

- 预测为PER的两个**目标语言跨度**

- 除了分离不同集群，加强了**语言之间的一致性**，有利于**跨语言迁移**

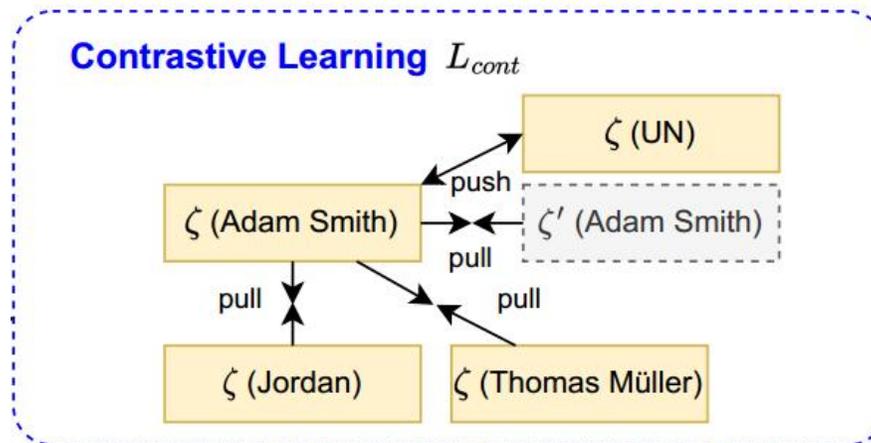
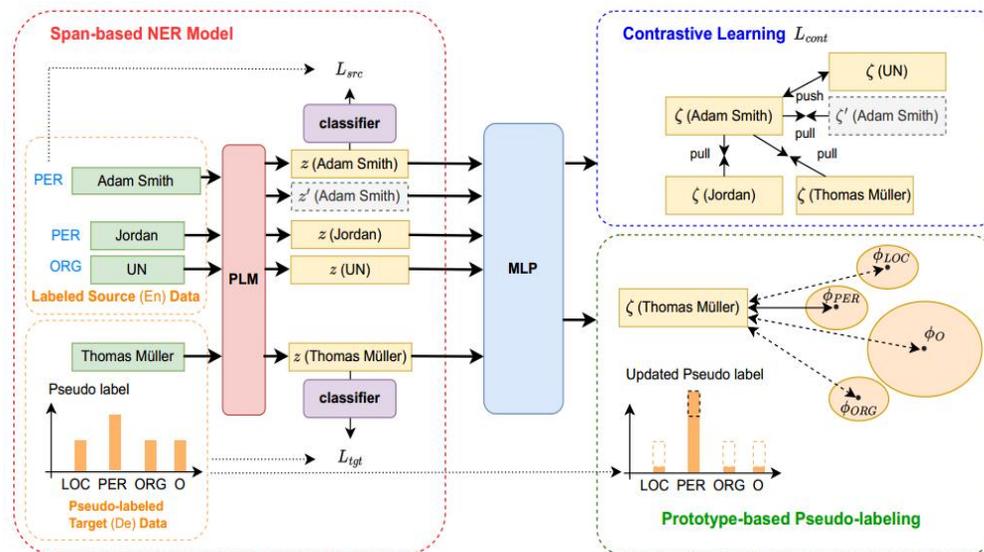
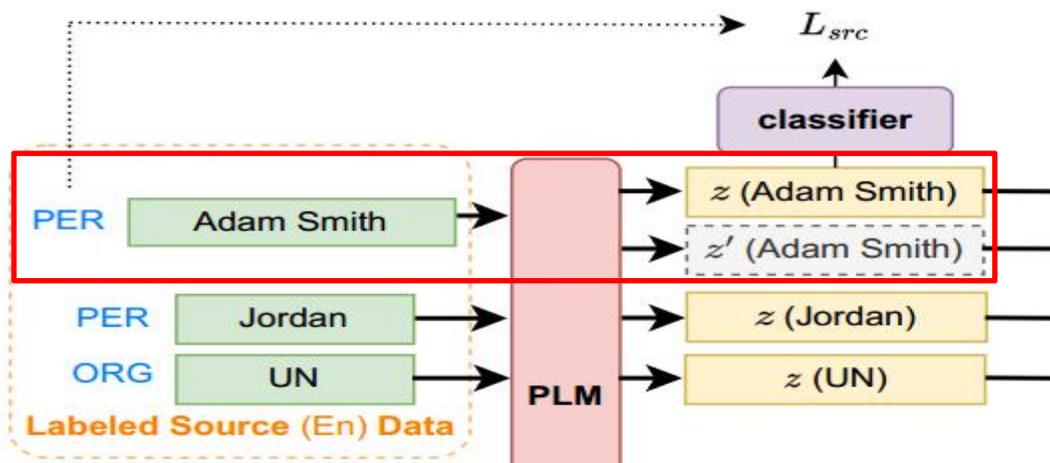


- 一致性正则化
  - 进一步**增强模型鲁棒性**
  - 最小化KL偏差强制模型输出**一致的预测**

$$L_{reg} = -\frac{1}{m} \sum_{i=1}^m \text{KL}(P_{\theta}(s_i) || P'_{\theta}(s_i))$$

- 模型总损失

$$L = L_{src} + L_{tgt} + L_{cont} + L_{reg}$$



## 基于原型的伪标注

- 类特定原型

- 每个epoch结束，对同一类混合的源语言和目标语言跨度的表示取平均值得到聚类中心表示向量

- 不理想，因为跨度表示和簇在剧烈变化

- 移动平均更新原型

$$\phi_c = \text{Normalize} (\alpha \phi_c + (1 - \alpha) \zeta_i), \\ \forall i \in \{i \mid y_i = c\}$$

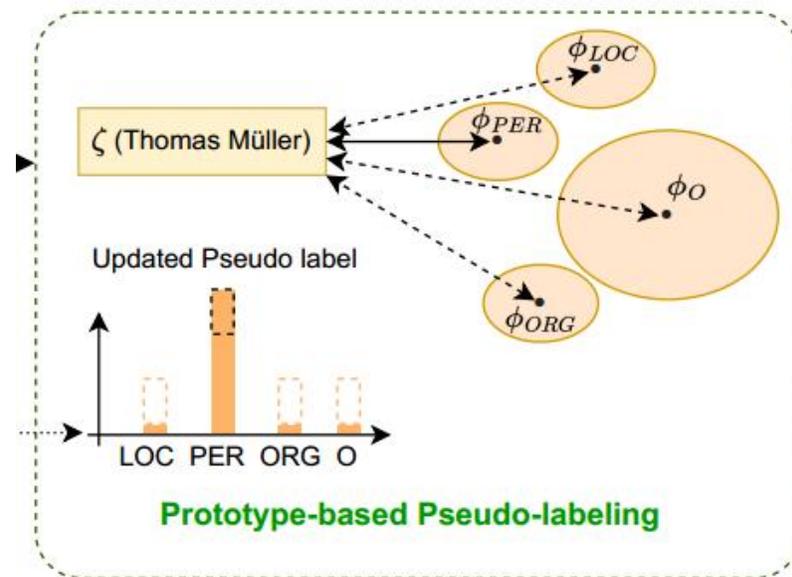
- 伪标签细化

- 参考原型细化目标语言跨度的伪标签

- 分配硬标签失去在自训练中使用软标签的优势

- 原型扰动比较大，训练不稳定

- 移动平均更新伪标签



$$\hat{y}_t^c = \begin{cases} \beta \hat{y}_{t-1}^c + (1 - \beta) & \text{if } c = \arg \max_{\gamma \in C} (\phi_\gamma \cdot \zeta) \\ \beta \hat{y}_{t-1}^c & \text{otherwise} \end{cases}$$

## 伪标签细化

### – 参考原型细化目标语言跨度的伪标签

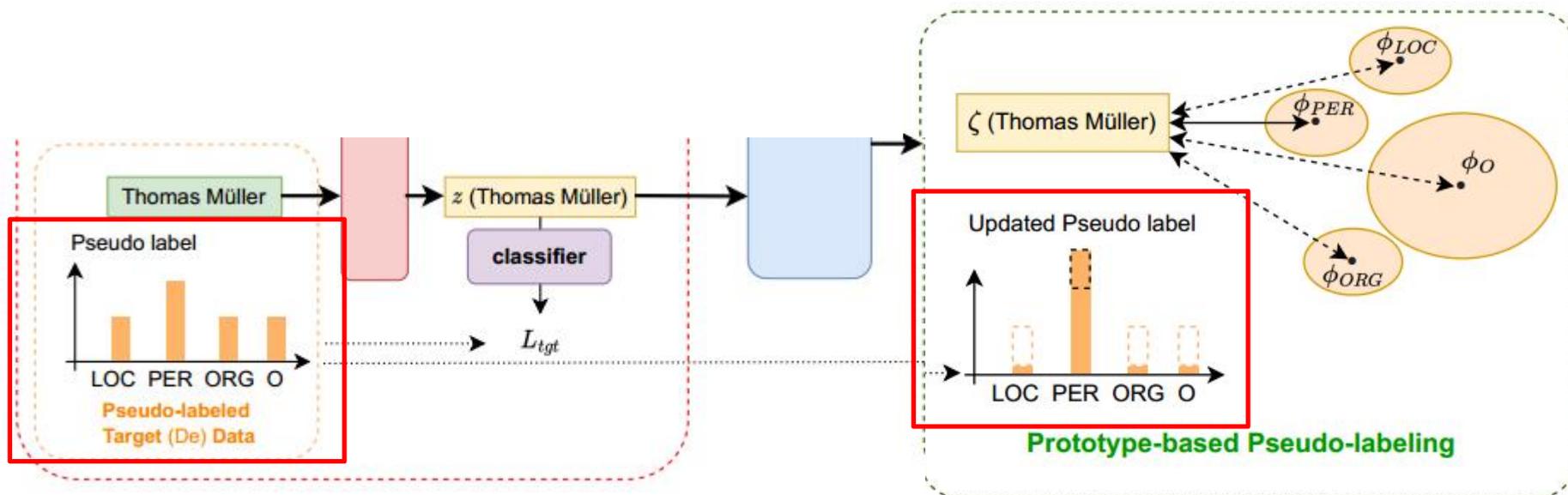
- 分配硬标签失去在自训练中使用软标签的优势

- 原型扰动比较大，训练不稳定

### – 移动平均更新伪标签

- 校正最可能不正确的伪标签，增强正确伪标签的置信度

$$\hat{y}_t^c = \begin{cases} \beta \hat{y}_{t-1}^c + (1 - \beta) & \text{if } c = \arg \max_{\gamma \in \mathbb{C}} (\phi_\gamma \cdot \zeta) \\ \beta \hat{y}_{t-1}^c & \text{otherwise} \end{cases}$$

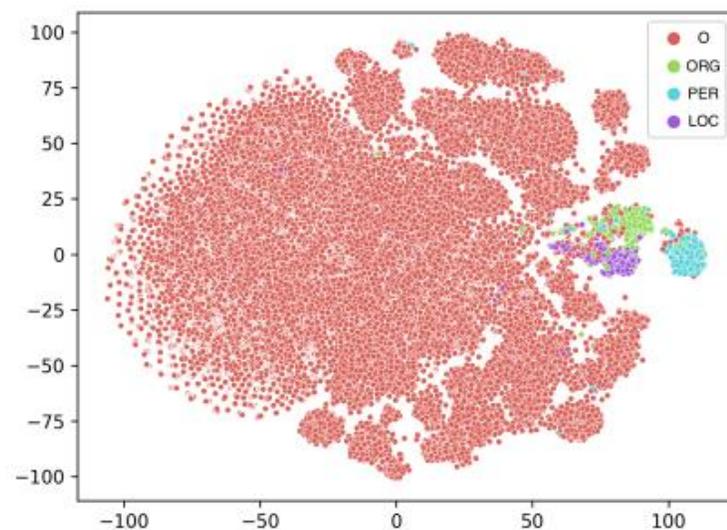
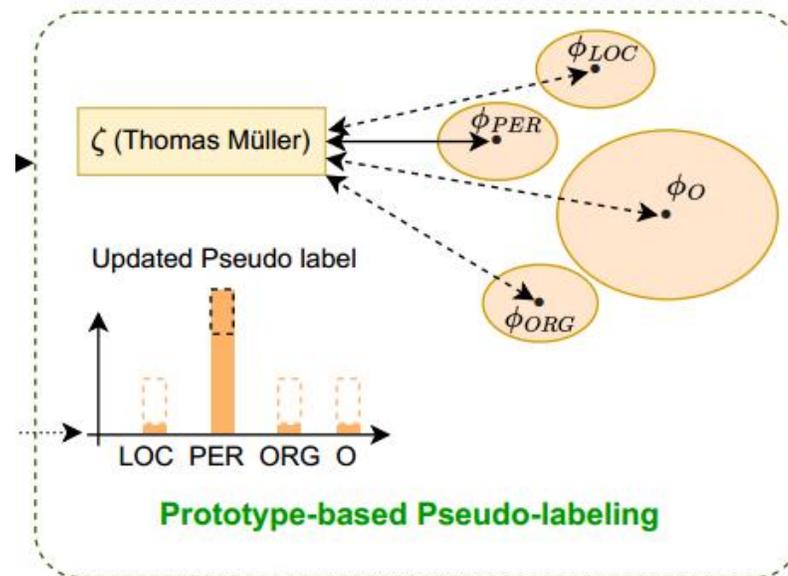
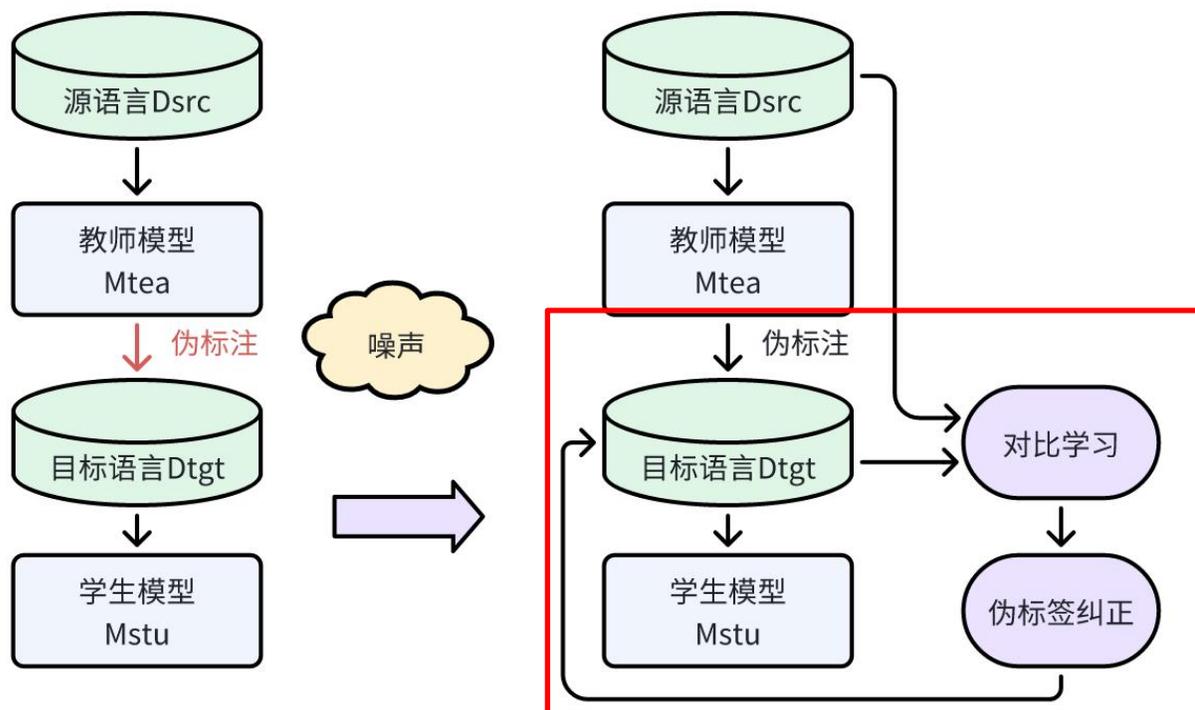


- **Margin-based标准**

- 和每个原型相似性都不超过一个范围，不更新

$$\beta = \begin{cases} \beta & \text{if } \arg \max_{\gamma \in \mathcal{C}} (\phi_{\gamma} \cdot \zeta_i) = O \\ \beta & \text{if } \max_{\gamma \in \mathcal{C} \setminus \{O\}} (\phi_{\gamma} \cdot \zeta_i) > r \\ 1 & \text{otherwise} \end{cases}$$

- **对比学习和基于原型伪标记两个组件协同工作**



(a) Vanilla self-training

## 数据集

### – CoNLL数据集

- 4种语言：英语(En)、德语(De)、西班牙语(Es)和荷兰语(Nl)

### – WikiAnn数据集

- 4种语言：英语(En)、阿拉伯语(Ar)、印地语(Hi)和汉语(Zh)

## 评价指标：F1值

## 实验结果

- 优于未结合的，结合目标语言未标注数据有利于跨语言NER
- 优于多教师和多轮自训练的，说明产生了更准确的伪标签

Method	De	Es	Nl	Avg
<i>w/o unlabeled data</i>				
Wiki	48.12	60.55	61.56	56.74
WS	58.50	65.10	65.40	63.00
TMP	61.50	73.50	69.90	68.30
BERT-f	69.56	74.96	77.57	74.03
AdvCE	71.90	74.30	77.60	74.60
<i>self-training</i>				
TSL	75.33	78.00	81.33	78.22
Unitrans	74.82	79.31	82.90	79.01
RIKD	<b>78.40</b>	79.46	81.40	79.75
AdvPicker				
- seq-tagging	75.01	79.00	82.90	78.97
- span-based <sup>†</sup>	73.93	84.70	81.01	79.88
MTMT	76.80	81.82	83.41	80.68
<b>ContProto (Ours)</b>	<b>76.41</b>	<b>85.02</b>	<b>83.69</b>	<b>81.71</b>

Method	Ar	Hi	Zh	Avg
<i>w/o unlabeled data</i>				
BERT-f	42.30	67.60	52.90	54.27
XLm-R <sub>Large</sub>	53.00	73.00	33.10	53.03
mT5 <sub>XXL</sub>	66.20	77.80	56.80	66.93
<i>self-training</i>				
TSL	50.91	72.48	31.14	51.51
RIKD	54.46	74.42	37.48	55.45
AdvPicker				
- seq-tagging <sup>†</sup>	53.76	73.69	41.24	56.23
- span-based <sup>‡</sup>	70.70	80.37	56.57	69.21
MTMT	52.77	70.76	52.26	58.60
<b>ContProto (Ours)</b>	<b>72.20</b>	<b>83.45</b>	<b>61.47</b>	<b>72.37</b>

## 消融实验

- 消融实验

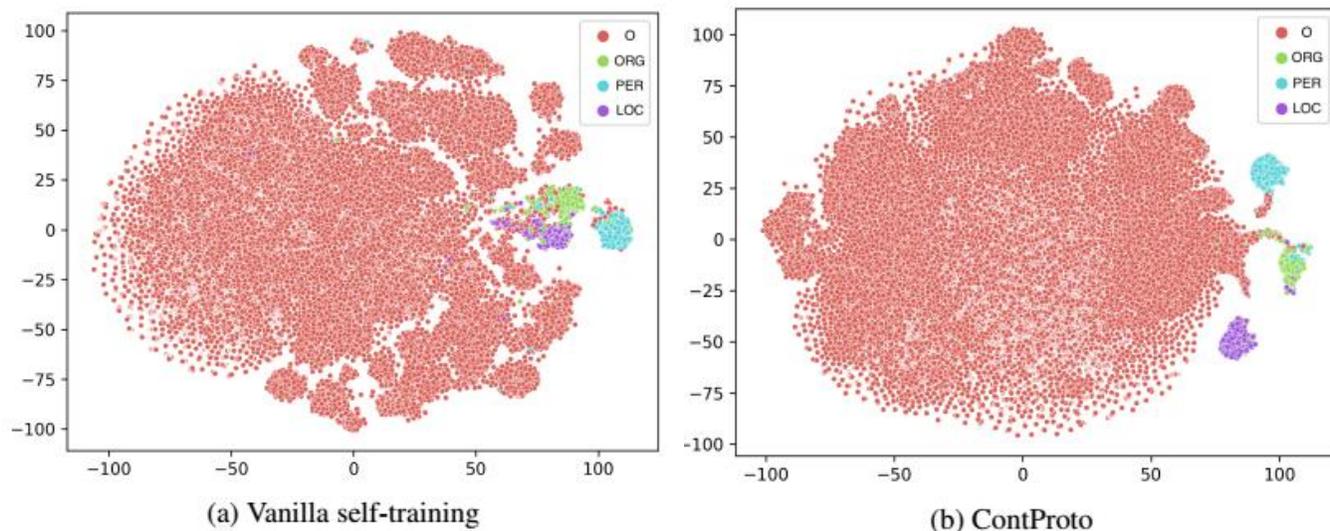
- proto: 基于原型的伪标注
- cl: 对比自训练
- reg: 一致性正则化
- fixed margin: 固定范围

Method	De	Es	Nl	Ar	Hi	Zh
<b>ContProto</b>	<b>76.41</b>	<b>85.02</b>	<b>83.69</b>	<b>72.20</b>	<b>83.45</b>	<b>61.47</b>
- w/o proto	74.87 (-1.54)	84.08 (-0.94)	81.44 (-2.25)	71.49 (-0.71)	83.10 (-0.35)	59.57 (-1.90)
- w/o proto & cl	74.17 (-2.24)	84.47 (-0.54)	81.03 (-2.66)	70.40 (-1.80)	81.00 (-2.45)	56.30 (-5.16)
- w/o reg	76.23 (-0.18)	84.96 (-0.06)	83.56 (-0.13)	72.15 (-0.05)	83.21 (-0.24)	61.31 (-0.16)
- fixed margin	74.65 (-1.76)	84.49 (-0.52)	83.09 (-0.60)	69.19 (-3.01)	83.07 (-0.38)	60.61 (-0.86)
- proto w/o cl	72.59 (-3.82)	81.18 (-3.84)	80.76 (-2.93)	69.72 (-2.48)	58.38 (-25.07)	53.52 (-7.95)

- 结论

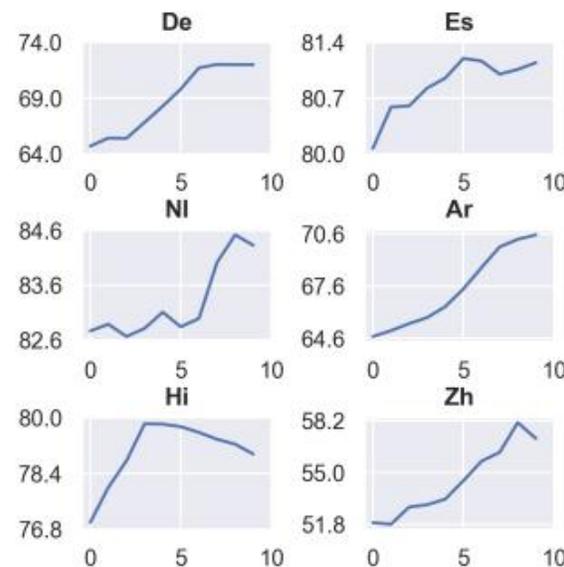
- w/o-proto在所有目标语言上都表现出降低的性能，验证了基于原型的伪标记在**提高伪标记质量**方面的能力
- proto w/o-cl导致性能急剧下降。如果没有对比的目的，不同类别的集群就会相互重叠，导致伪标签质量恶化。对比学习的聚类效应对于准确的基于原型的伪标记至关重要

## • 可视化跨度分布



- 不同类别**重叠**->更可区分的表示, **对比目标**的有效性
- 非实体集群明显大于其他实体类别, margin-based标准的必要性

## • 伪标签质量



- 德语、阿拉伯语、汉语F1显著增加
- 印地语epoch3之后, F1下降, 很多假阴性。印地语**实体分布方差较大**。



**CoLaDa**

## TIPO

T	目标	用源语言标注数据训练NER模型，在目标语言测试数据上识别命名实体
I	输入	源语言标注数据，目标语言未标注数据
P	处理	1.实例协作去噪 2.模型协作去噪
O	输出	目标语言上的命名实体识别结果

P	问题	通过翻译源语言标注数据或对目标语言未标注数据进行知识蒸馏来缓解数据稀缺问题，受到标注噪声的影响
C	条件	源语言标注数据，目标语言未标注数据
D	难点	结合翻译数据和目标语言数据的互补特性来进行去噪
L	水平	ACL 2023 CCF-A

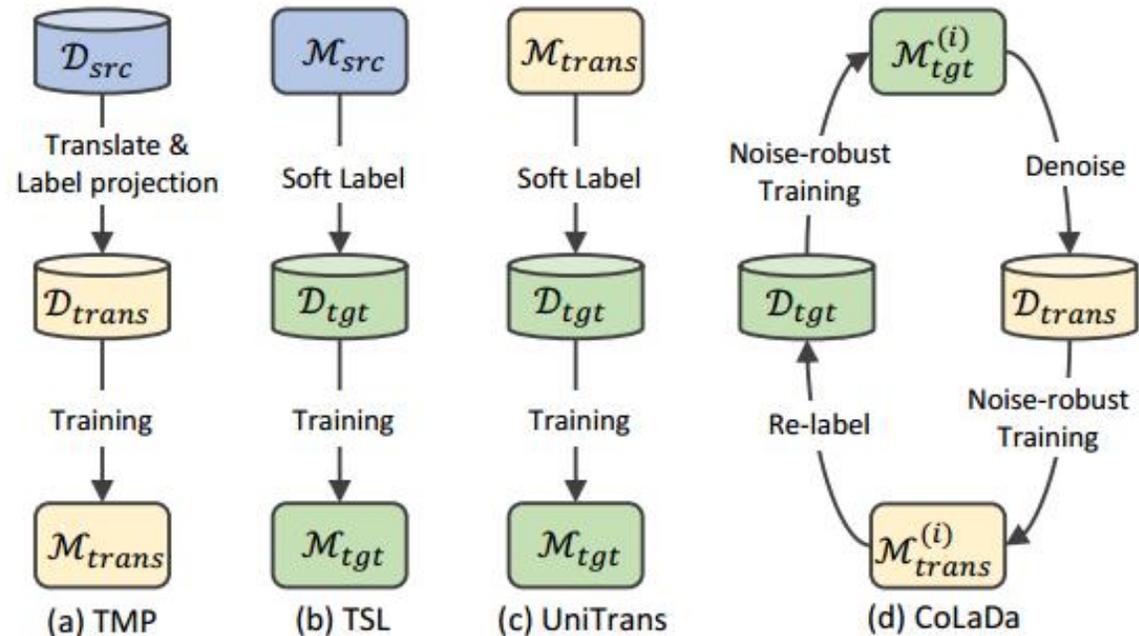
## 整体框架

- 核心思想

- 基于模型协作去噪，结合了在两个数据源上训练的模型，以**迭代的方式**对两个数据来源的伪标签进行**相互去噪**
- 基于实例去噪，利用表示空间中每个标记**邻域的标签一致性**来重新加权知识蒸馏中的软标记示例

- 算法流程

- 1.对**翻译数据**的噪声鲁棒学习
- 2.对**未标注的目标语言数据**的噪声稳健学习
- 3.将**实例协作**集成到基于**模型协作**的去噪过程



- 核心思想：给定token的标签与其许多邻居不一致，更可能是有噪声的
- 方法：通过测量给定token的邻域的标签一致性评估其标签可靠性，然后使用可靠性得分对教师模型中的噪声标签加权

- 创建内存库  $B_D$

- 获得和存储所有句子token的表示

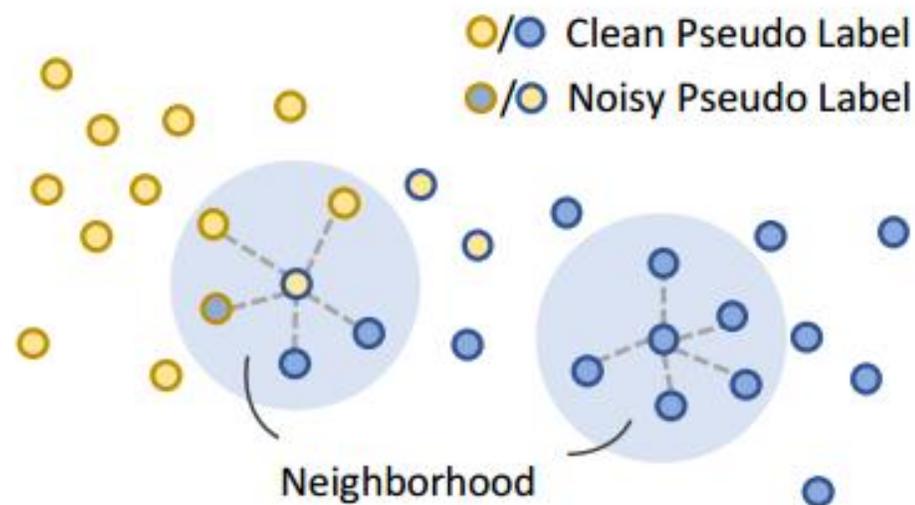
- 计算标签一致性

- 计算  $x_i$  的邻居被分配与  $x_i$  相同标签的分数

$$\lambda(p_i; x_i) = \frac{1}{K} \sum_{x_j \in \mathcal{N}_k(x_i)} I(\hat{y}_j = \hat{y}_i)$$

- 得到可靠性分数

- 加权的类自适应可靠性得分
- 模型训练期间用作软标签概率的权重



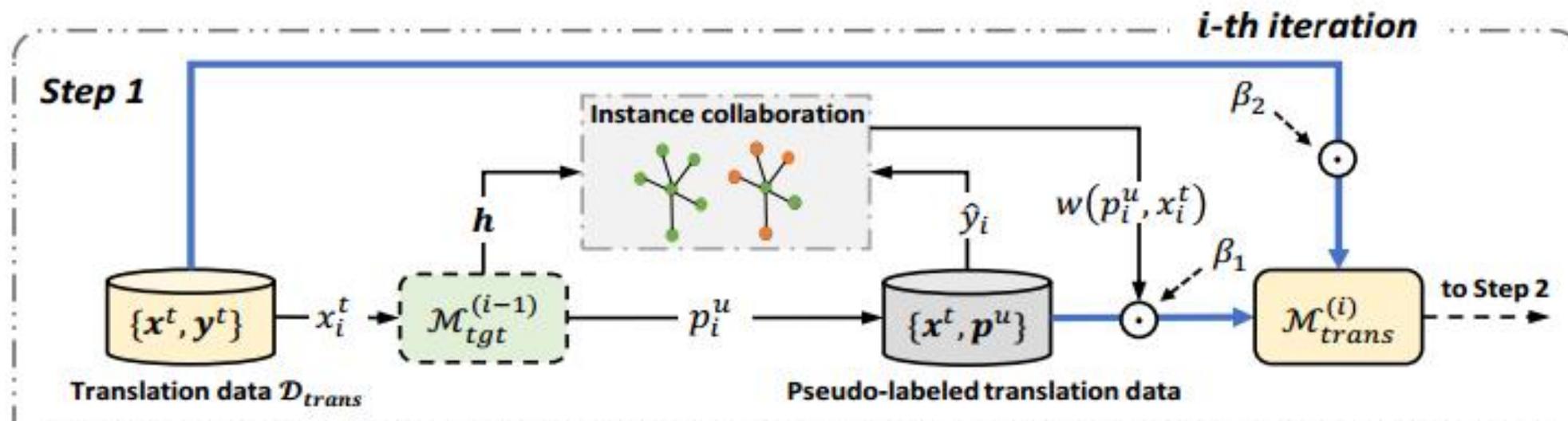
$$w(p_i; x_i) = \text{Sigmoid}(\alpha(\lambda(p_i; x_i) - \mu(\hat{y}_i)))$$

- 在翻译数据上的噪声鲁棒训练
  - 从协作模型  $M_{tgt}$  获得软标签  $p_i^u$
  - 原来标签  $y_i^t$  和软标签  $p_i^u$  作为监督训练模型  $M_{trans}$
  - 加入实例协作去噪
  - 向协作者  $M_{trans}$  提供可靠性评估参数  $\beta_1$

利用  $M_{tgt}$  来减少翻译数据标签噪声

$$\mathcal{L}^{x^t} = \frac{1}{L} \sum_{i=1}^L (\beta_1 \text{CE}(\hat{p}_i^t, p_i^u) + \beta_2 \text{CE}(\hat{p}_i^t, y_i^t))$$

$$\beta_1(x_i^t) \leftarrow \beta_1 * w(p_i^u, x_i^t)$$

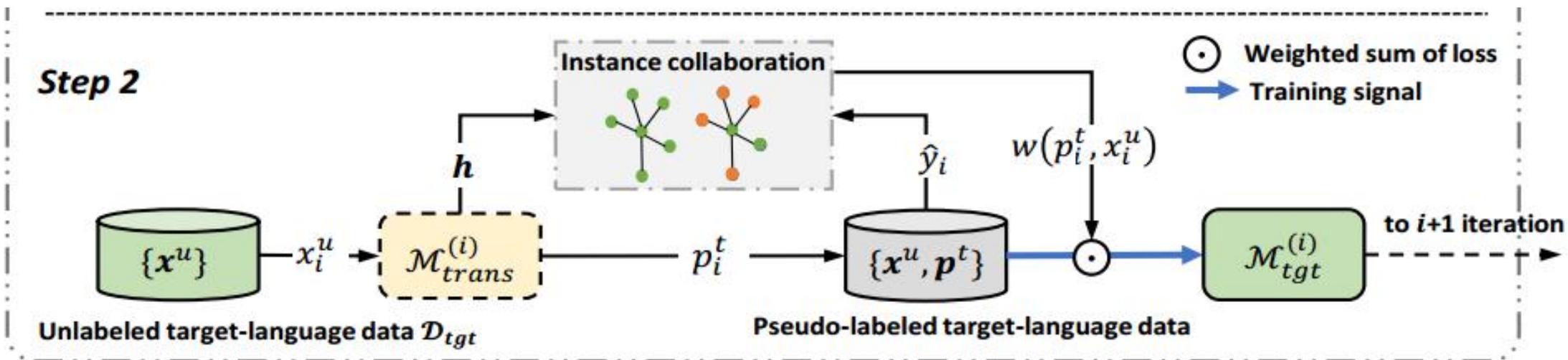
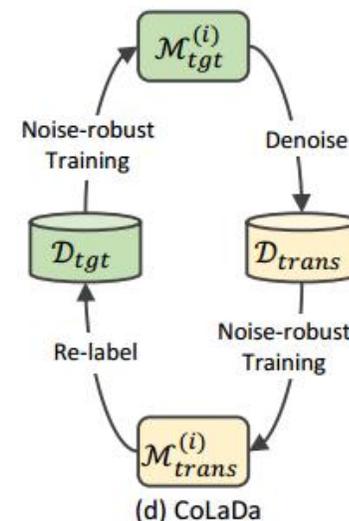


## 在目标语言未标注数据上的噪声鲁棒训练

- 利用  $M_{trans}$  预测软标签  $p_i^t$
- 加入实例去噪得到可靠性分数

利用  $M_{trans}$  为目标语言未标注数据提供高质量监督

$$\mathcal{L}^{x^u} = \frac{1}{L} \sum_{i=1}^L w(p_i^t, x_i^u) \cdot \text{CE}(\hat{p}_i^u, p_i^t)$$



## 数据集

### – CoNLL数据集

- 4种语言：英语(En)、德语(De)、西班牙语(Es)和荷兰语(Nl)

### – WikiAnn数据集

- 4种语言：英语(En)、阿拉伯语(Ar)、印地语(Hi)和汉语(Zh)

## 评价指标：F1值

## 实验结果

- WikiAnn上效果显著，说明在这些和源语言差距更大的语言中，标签噪声的问题甚至更为关键

Method	de	es	nl	avg
<i>mBERT based methods:</i>				
mBERT (Wu and Dredze, 2019)	69.56	74.96	77.57	73.57
AdvCE (Keung et al., 2019)	71.90	74.3	77.60	74.60
TSL (Wu et al., 2020c)	73.16	76.75	80.44	76.78
UniTrans (Wu et al., 2020b)	74.82	79.31	82.90	79.01
TOF (Zhang et al., 2021)	76.57	80.35	82.79	79.90
AdvPicker (Chen et al., 2021)	75.01	79.00	82.90	78.97
RIKD (Liang et al., 2021)	75.48	77.84	82.46	78.59
MTMT (Li et al., 2022)	76.80	<b>81.82</b>	83.41	80.67
<b>CoLaDa (ours)</b>	<b>77.30</b>	80.43	<b>85.09</b>	<b>80.94</b>
<i>XLM-R based methods:</i>				
MulDA (Liu et al., 2021)	74.55	78.14	80.22	77.64
xTune (Zheng et al., 2021)	74.78	80.03	81.76	78.85
ConNER (Zhou et al., 2022)	77.14	80.50	83.23	80.29
<b>CoLaDa (ours)</b>	<b>81.12</b>	<b>82.70</b>	<b>85.15</b>	<b>82.99</b>
Method	ar	hi	zh	avg
<i>mBERT based methods:</i>				
BERT-align (Wu and Dredze, 2020)	42.30	67.60	52.90	54.26
TSL (Wu et al., 2020c)	43.12	69.54	48.12	53.59
RIKD (Liang et al., 2021)	45.96	70.28	50.40	55.55
MTMT (Li et al., 2022)	52.77	70.76	52.26	58.59
UniTrans <sup>†</sup> (Wu et al., 2020b)	42.90	68.76	56.08	55.91
<b>CoLaDa (ours)</b>	<b>54.26</b>	<b>72.42</b>	<b>60.77</b>	<b>62.48</b>
<i>XLM-R based methods:</i>				
XLM-R (Conneau et al., 2020)	50.84	72.17	39.23	54.08
ConNER (Zhou et al., 2022)	59.62	74.49	39.17	57.76
<b>CoLaDa (ours)</b>	<b>66.94</b>	<b>76.69</b>	<b>60.08</b>	<b>67.90</b>

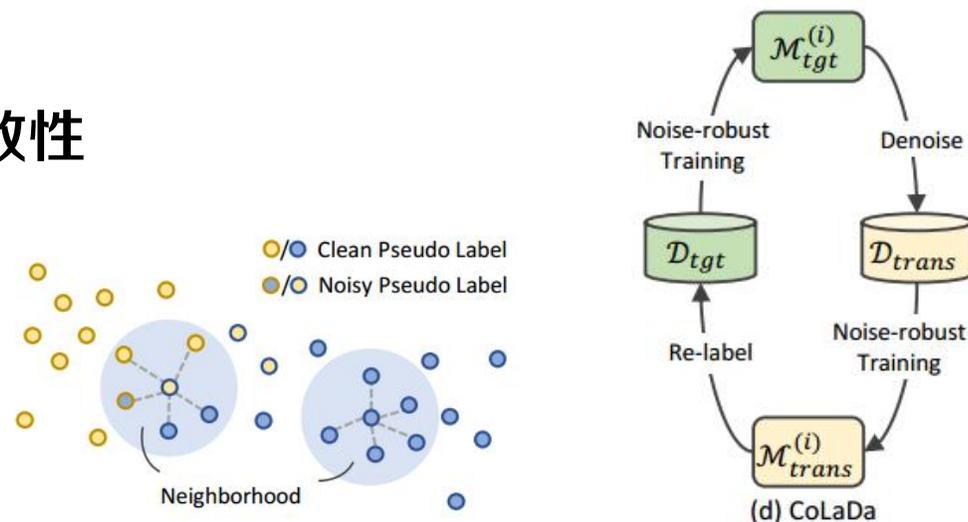
## 消融实验

- 去掉实例协作
- 去掉翻译数据去噪
- 去掉迭代去噪
- 去掉模型协作
- 去掉实例协作和模型协作

Method	de	es	nl	ar	hi	zh
CoLaDa	<b>77.30</b>	<b>80.43</b>	<b>85.09</b>	<b>54.26</b>	<b>72.42</b>	<b>60.77</b>
1) CoLaDa w/o instance collaboration	76.08	79.94	83.86	50.98	71.31	59.64
2) CoLaDa w/o translation data denoise	76.17	79.22	83.10	41.41	71.10	55.04
3) CoLaDa w/o iteratively denoise	75.77	79.64	83.50	47.82	71.31	57.64
4) CoLaDa w/o model collaboration	75.64	78.99	82.98	46.51	71.09	55.25
5) CoLaDa w/o instance & model collaboration	74.54	79.94	82.97	42.33	70.39	55.55

## 结论

- 利用邻域信息减少知识蒸馏中标签噪声的有效性
- 用协作模型去改进翻译数据的标签的重要性
- 通过两个协作模型的迭代提高标签质量
- 去噪框架的重要性



- 模型协作分析实验

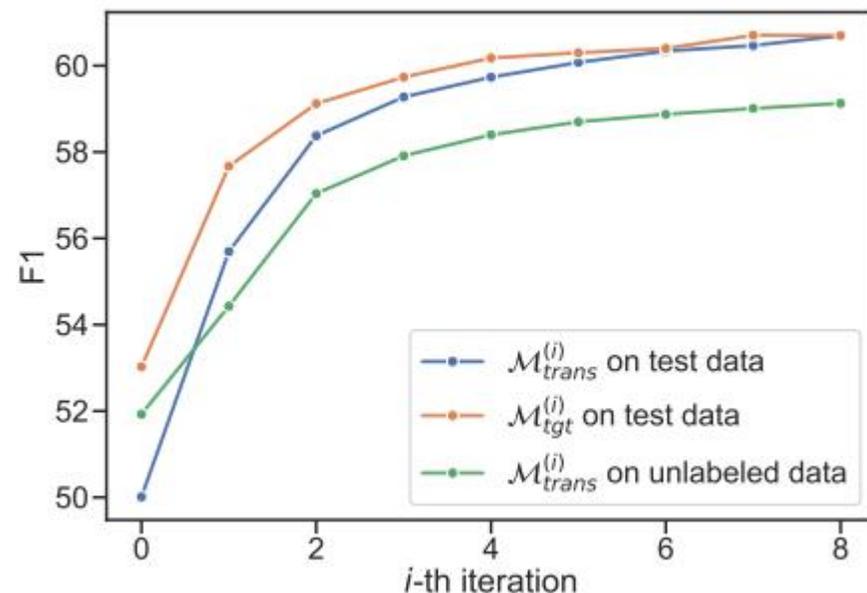
- 了解Mtrans和Mtgt两个模型如何协同改进

- 结论

- Mtrans和Mtgt的F1分数随着迭代的进行而不断提高，并最终在最后一次迭代时收敛。表明这两个模型都受益于所提出的模型协作方案

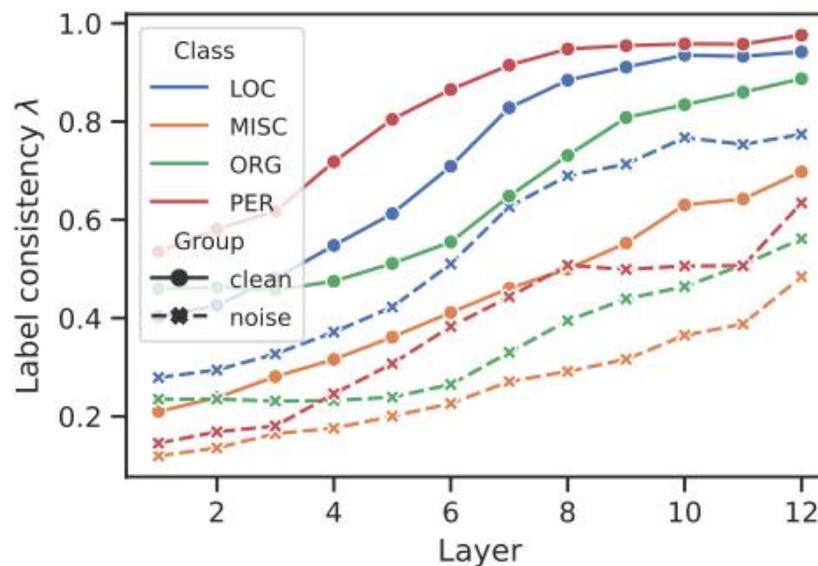
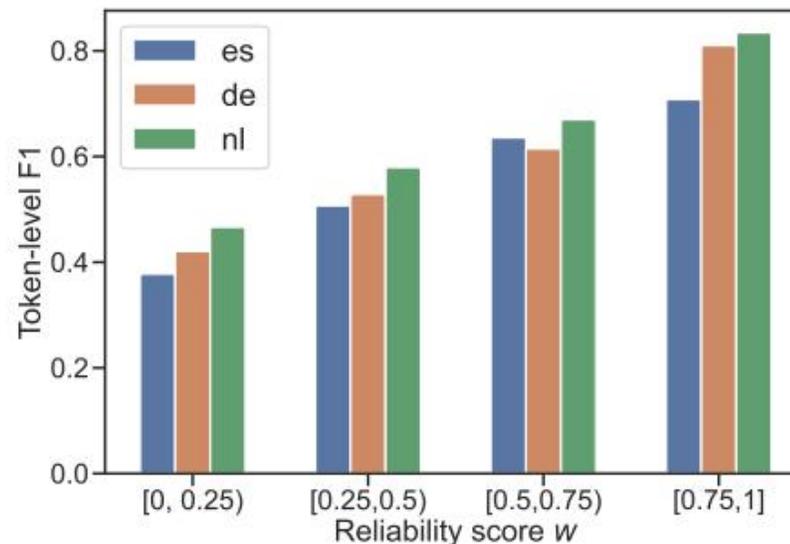
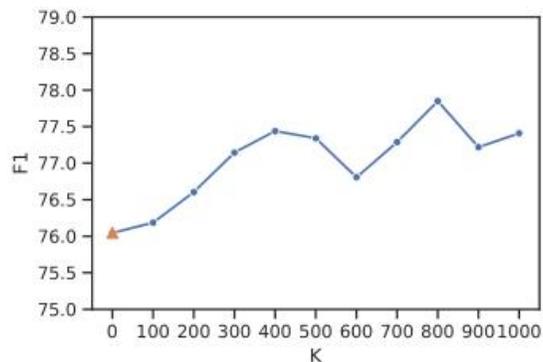
- 推测原因

- 改进的Mtgt可以在翻译数据上**提供更准确的标签**，这进一步有助于通过对这种翻译数据的噪声鲁棒学习来改进Mtrans
- 改进的Mtrans预测目标语言未标记数据上具有**更高质量的伪标签**，这进一步有利于Mtgt的学习



## 实例协作分析实验

- 研究基于实例协作的去噪的工作机制
- 可靠性分数和标签质量的关系
  - 标签质量与可靠性分数成正比
- 标签一致性分析
  - 干净标签比噪声标签有更大的一致性
  - 不同类型是不同的标签一致性分数
  - 上层大于底层，差距变得更大
- 邻域信息K的选择
  - 选择合适的K



## • 案例分析实验

### – 噪声类型：不准确的翻译、对齐错误、原始源语言数据中的注释错误

<p><i>Original English:</i> ** " Duchy of Aquitaine[LOC] " ' – William V[PER] ( 995–1030 )</p> <p><i>Original Translation:</i> ** " 爱因公爵[LOC] " – 威廉五世[PER] (995)</p> <p><i>Translate-train:</i> ** " 爱因公爵[LOC] " – 威廉五世[PER] (995)</p> <p><i>Ours:</i> ** " 爱因公爵[PER] " – 威廉五世[PER] (995)</p>	<p><b><i>Inaccurate translation</i></b> Translate “Duchy of Aquitaine” into “爱因公爵” (Duke of Ain) incorrectly.</p>
<p><i>Original English:</i> There have been many tenants, including The Sunday Times[ORG] " and The Daily News[ORG] ".</p> <p><i>Original Translation:</i> 有许多租户，包括 《星期日时报[ORG]》和 《每日新闻[ORG]》。</p> <p><i>Translate-train:</i> 有许多租户，包括 《星期日时报[ORG]》和 《每日新闻[ORG]》。</p> <p><i>Ours:</i> 有许多租户，包括 《星期日时报[ORG]》和 《每日新闻[ORG]》。</p>	<p><b><i>Inaccurate alignment boundary</i></b> The symbols 《 and 》 are corner brackets used to enclose the name of a newspaper, and other works.</p>
<p><i>Original English:</i> It is found in Democratic Republic of Congo[ORG] , Kenya[LOC] , Tanzania[LOC] ....</p> <p><i>Original Translation:</i> 它位于刚果民主共和国[ORG] , 肯尼亚[LOC] , 坦桑尼亚[LOC] ...</p> <p><i>Translate-train:</i> 它位于刚果民主共和国[ORG] , 肯尼亚[LOC] , 坦桑尼亚[LOC] ...</p> <p><i>Ours:</i> 它位于刚果民主共和国[LOC] , 肯尼亚[LOC] , 坦桑尼亚[LOC] ...</p>	<p><b><i>Label noise in English data</i></b> Wrong entity type in original English data for “Democratic Republic of China”.</p>

北京林业大学  
景观规划设计学院



## 特点总结与未来展望

- 算法优势

- ContProto

- 在源语言和目标语言之间产生紧密一致的表示来增强跨语言的可迁移性

- CoLaDa

- 结合翻译数据和目标语言数据的互补特性来去噪

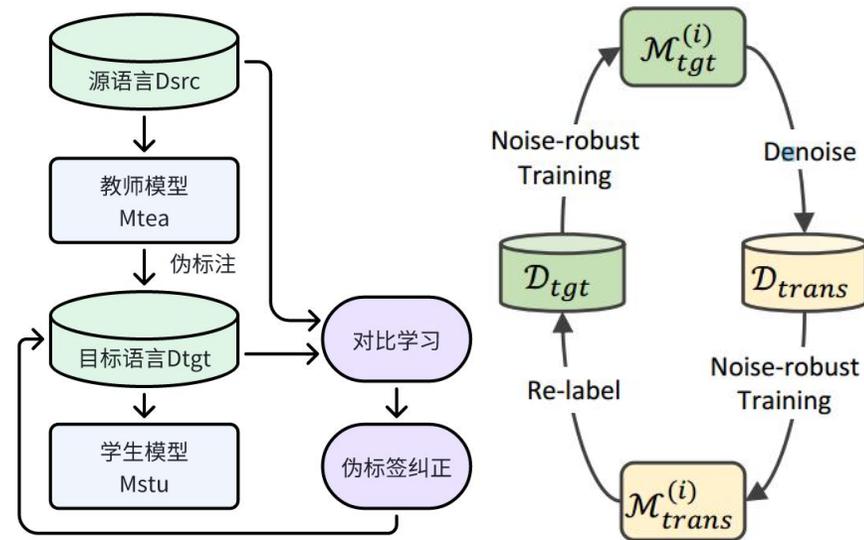
- 算法劣势

- ContProto

- 通过原型学习校正教师生成的伪标签，不可避免地会引入噪声
    - 只调整已经产生的伪标签，而忽略了提高教师网络产生高质量伪标签的能力

- CoLaDa

- 受限于两个协作模型能力的瓶颈



- 未来展望

- 结合数据增强技术

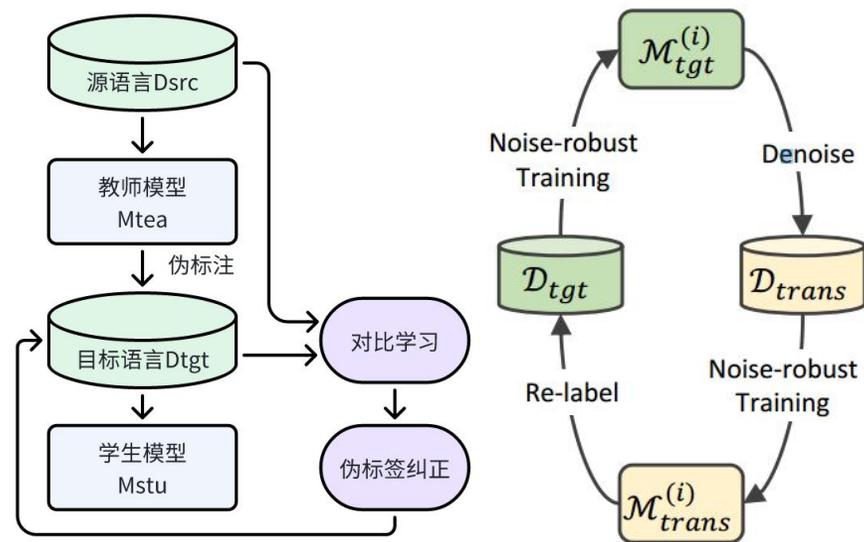
- 需要未标注的目标语言数据

- 结合更有效的表示学习技术

- 更多类型实体分布不平衡，长尾实体原型不准确

- 更有效的伪标签更新策略

- 某些语言在训练阶段伪标签质量略有下降



- [1] Ma T, Wu Q, Jiang H, et al. CoLaDa: A Collaborative Label Denoising Framework for Cross-lingual Named Entity Recognition[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2023: 5995-6009.
- [2] Zhou R, Li X, Bing L, et al. Improving Self-training for Cross-lingual Named Entity Recognition with Contrastive and Prototype Learning[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2023: 4018-4031.

知人者智，自知者明。胜人者有力，自胜者强。知足者富。强行者有志。不失其所者久。死而不亡者，寿。

## 谢谢！

