计算机科学
COMPUTER SCIENCE

基于半监督学习的域适应实体解析算法

戴超凡, 丁华华

引用本文

戴超凡, 丁华华. 基于半监督学习的域适应实体解析算法[J]. 计算机科学, 2024, 51(9): 214-222.

DAI Chaofan, DING Huahua. Domain-adaptive Entity Resolution Algorithm Based on Semi-supervised Learning [J]. Computer Science, 2024, 51(9): 214-222.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230800102

基于半监督学习的域适应实体解析算法

戴超凡　丁华华
国防科技大学信息系统工程全国重点实验室　长沙 410073
(cfdai@nudt.edu.cn)

摘　要　实体解析旨在查找两个数据实体是否引用同一实体,是许多自然语言处理任务中的一项基本任务。现有的基于深度学习的实体解析解决方案通常需要大量的标注数据,即使利用预训练的语言模型进行训练,仍然需要数千个标签才能达到令人满意的准确性。现实场景中,这些标注数据并不容易获得。针对上述问题,提出了一个基于半监督学习的域适应实体解析模型。首先,在源域上训练一个分类器,然后利用域适应减小源域和目标域的分布差异,同时用数据增强后的目标域软伪标签加入源域迭代训练,从而实现从源域到目标域的知识迁移。在13个来自相同或不同领域的数据集上对所提模型进行了对比实验和消融实验,实验结果表明,与无监督基线模型相比,所提模型在多个数据集上的F1值平均提升了2.84%,9.16%和7.1%;与有监督基线模型相比,所提模型只需要20%～40%的标签就可以达到与有监督学习相当的性能。消融实验进一步证明了所提模型的有效性,其总体上可以获得更好的实体解析结果(相关代码已开源[1])。

关键词:实体解析;域适应;伪标签;预训练语言模型;数据增强
中图分类号　TP391

Domain-adaptive Entity Resolution Algorithm Based on Semi-supervised Learning

DAI Chaofan and DING Huahua
National Key Laboratory of Information Systems Engineering, National University of Defense Technology, Changsha 410073, China

Abstract　Entity resolution is a fundamental task in many natural language processing tasks, which aims to find out whether two data entities refer to the same entity. Existing deep learning-based solutions for entity resolution typically require a large amount of annotated data, even when pre-trained language models are used for training. Obtaining such annotated data is challenging in real-world scenarios. To address this issue, a domain-adaptive entity resolution model based on semi-supervised learning is proposed. First, a classifier is trained on the source domain, and then domain adaptation is used to reduce the distributional difference between the source and target domains. Soft pseudo-labels from the augmented target domain are then added to the source domain for iterative training, enabling knowledge transfer from the source to the target domain. Comparison and ablation experiments are performed on 13 datasets from various domains. The results show that, compared to unsupervised baseline models, the proposed model achieves an average F1 score improvement of 2.84%, 9.16%, and 7.1% across multiple datasets. Compared to supervised baseline models, it achieves comparable performance with only 20% to 40% of the labels required. Ablation experiments further demonstrate the effectiveness of the proposed model, and better entity resolution results can be obtained in general (The relevant code is available[1]).

Keywords　Entity resolution, Domain adaptation, Pseudo-labels, Pre-trained language model, Data augmentation

1　引言

实体解析旨在查找两个数据实体是否引用同一实体。作为许多自然语言处理任务中的一项基本任务,实体解析并不容易,因为自然语言文本通常由于上下文信息的质量和主题连贯性而存在消歧困难。有许多文献利用模型中心来解决实体解析问题,包括基于规则的方法(如析取范式[1]和通用布尔公式[1])、基于机器学习的方法(如SVM[2]和随机森林[3]),以及基于深度学习的方法(如DeepMatcher[4],DeepER[5]和Ditto[6])。截至目前,基于深度学习的解决方案取得了最先进的结果。

然而,基于深度学习的实体解析方法通常需要大量标记的训练数据。例如,即使利用预训练的语言模型(如Ditto[6])进行训练,仍然需要数千个标签才能达到令人满意的准确性。实际上,基于深度学习的实体解析方法的主要痛点是需要大量标注工作来创建足够的训练数据。

但是在大数据时代,无论是公共基准数据集(例如WDC[7]和DBLP-Scholar[4]),还是企业内部,都提供了许多在

[1] https://github.com/cainiao12306/SSDAER

到稿日期:2023-08-16　返修日期:2023-11-27
通信作者:丁华华(dinghuahua@nudt.edu.cn)

相同或相关领域中可用的标记的实体识别数据集。因此,如果能再利用这些标记的源实体解析数据集,将其用于新的目标实体解析数据集,将能显著减少昂贵的人力标注工作量。

目前通过深度领域自适应进行知识转移[8]是一种有效的解决方法,它通过在学习过程中嵌入领域自适应来学习从源领域到目标领域的可迁移表示。图 1 为域适应框架图。图 1 显示了通过域适应对齐源域和目标域的数据分布之后,在源域上训练的模型能在目标域上也能起到很好的匹配效果,将匹配实体对(实心)和不匹配实体对(空心)区分开。但是,目前仅基于域适应的方法,不能很好地关注到目标域未标记数据中的潜在信息,在源域和目标域差异过大时,仅仅依靠域对齐并不能让模型很好地学习目标域特征。

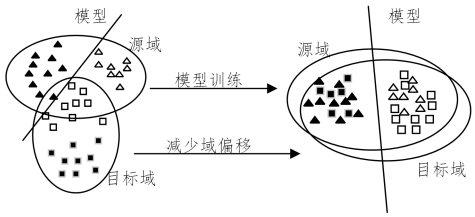

图 1 域适应框架图

Fig. 1 Framework of domain daptation

半监督学习[9]在利用目标域未标记数据进行学习方面具有巨大潜力,它关注的是具有稀疏标记数据和大量无标签数据的场景。半监督学习是一个跨不同领域的任务,包括图像[10]、音频[11]、时间序列[12]和文本[13]。图像分类方法主要集中在利用不同扰动下对相同样本的预测的一致性(一致性正则化)[14-15],而其他方法则直接为无标签数据生成标签以指导学习过程(伪标签)[16-17]。

针对以上问题,本文提出了一种基于半监督学习的域适应实体解析模型(Domain-Adaptive Entity Resolution Model Based on Semi-Supervised Learning,SSDAER)。该模型充分利用了源域带标签数据和目标域无标签数据进行学习,以解决传统深度实体解析模型需要大量标签的问题。SSDAER 分为两个阶段。第一阶段,在源域进行有监督学习训练分类器,同时使用最大均值差异来最小化源域和目标域之间的距离,利用域标签训练一个能提取域公共特征的特征提取器。第二阶段,利用第一阶段训练好的特征提取器和分类器,来给目标域打上软伪标签,将伪标签数据用于进一步训练分类器,并设计了新的损失函数,以更好地学习目标域特征。同时,利用 Mixup[18]特征级线性插值的数据增强方法来减小伪标签学习在源域上的确认偏差,增强模型的鲁棒性。本文的主要贡献在于:

1)据调查,本文是第一个将域适应和半监督学习相结合并应用于实体解析问题的,相比目前最优的无监督实体解析模型,本文模型在多个数据集上可以取得更优的效果;与有监督模型相比,本文模型可以以更少的标注数据取得与有监督学习相当的结果。

2)使用了 Mixup[18]数据增强方法,当源域和目标域来自不同领域时,可以显著增强模型的性能。

3)在 13 个来自不同领域的数据集上进行了广泛的实验,所提方法在各个数据集上的表现均优于仅基于域适应的实体解析模型,特别是当源域和目标域来自不同领域时,F1 值平均提升了 8.73%。

2 相关工作

2.1 实体解析

早期,用于实体解析的方法主要有基于规则的方法[19]、基于匹配函数的方法[3,20]、基于聚类的方法[21-24]和传统的机器学习模型[25-27]。

近几年,基于深度学习的方法在实体识别中被广泛使用,并取得了最先进的结果。DeepER[5]设计了两个深度神经网络来提取实体对的特征,并将实体解析建模为二分类任务。DeepMatcher[4]系统地定义了深度学习解决方案在实体解析中的架构和设计空间。Ditto[6]首次将预训练的语言模型应用于实体解析,可以减少所需的训练数据数量。HierGAT[28]首次提出了一种基于分层图注意力变换网络的新的实体解析方法,其在部分数据集上取得了最先进的结果。即使如此,基于深度学习的方法仍然需要大量标记的训练数据。

2.2 域适应

域适应[29-33]是迁移学习[34]的一种情况,是一种有效的利用标记的源数据来适应不同的目标数据的方法。现有的域适应解决方案可以广泛分为实例级别、特征级别和其他级别。

1)实例级别:这些方法旨在通过使其适应目标分布来重用源数据实例。为此,传统研究对源数据实例进行重新加权,以强调与目标相似的实例。最近的 DL 方法提出了学习映射函数,将源数据实例调整为与目标类似的方法[35],或者为目标数据生成伪标签[36]。

2)特征级别:这些方法专注于学习领域不变且具有区分性的特征。现有的特征级别方法可以分为 3 类。(1)基于差异的方法利用各种度量方法,例如最大均值差异[37](Maximum Mean Discrepancy,MMD)、二阶统计量[38]和高阶矩[39],来计算和减小源域和目标域之间的分布差异。(2)对抗性方法利用对抗性学习框架,例如梯度反转[40]和基于 GAN 的极小极大训练[41],来学习领域不变的特征。(3)基于重构的方法将数据重构作为辅助任务引入,以提升特征学习过程[42]。

3)其他级别:还有许多其他领域自适应研究,例如处理多个领域自适应的元学习等。

2.3 基于域适应的实体解析

由于域适应在计算机视觉和自然语言处理领域都取得了不错的效果,一些学者也开始尝试将域适应用于实体解析工作。Thirumuruganathan 等介绍了一种实例级方法来重新加权源数据实例并使它们适应目标[43]。Kasai 等将具有梯度反转的域适应用于实体解析[44]。Tu 等设计了一个用于深度实体解析的域适应的框架[45]。Trabelsi 等提出了新的基于域适应的方法,将知识从多个源域转移到目标域[46]。但是目前还没有将半监督学习用于深度实体解析域适应问题的研究,本文实验填补了这一空缺并证明了半监督与基于差异的域适应相结合的有效性。

3 问题描述

3.1 实体解析

实体解析旨在查找两个数据实体是否引用同一实体。令

A 和 B 是两个具有多个属性的关系表，每个元组 $a \in A$（或 $b \in B$）表示由一组属性值对 $\{(attr_i, val_i)\}(1 \leqslant i \leqslant k)$ 组成的实体，其中 $attr_i$ 和 val_i 分别表示第 i 个属性名称和值。实体解析就是找到所有的指代现实世界中同一实体的实体对 $(a, b) \in A \times B$，如果实体对引用相同的现实世界对象，则称它们是匹配的，反之就是不匹配。

图 2(a)给出了一个源域的带标记数据集，每个实体对由来自不同表的两条实体记录组成，如 $(a_{s1}, b_{s1}, 1)$，并用 0/1 来表示两条记录是否匹配。图 2(b)给出了目标域的未标记数据集。

id	title	category	price		id	title	category	price
a_{s1}	balt wheasel …	stationery …	239.88	$(a_{s1}, b_{s1}, 1)$	b_{s1}	balt inc. …	laminating …	134.45
a_{s2}	kodak esp …	printers	68	$(a_{s2}, b_{s2}, 0)$	b_{s2}	kodak esp 7 …	kodak	149.29
a_{s3}	hp q3675a …	printers	194.84	$(a_{s3}, b_{s3}, 1)$	b_{s3}	hewlett …	cleaning repair	NULL

(a)带标记的源域

id	title	description	price		id	title	description	price
a_{T1}	samsung 52 ' series 7 black flat …	samsung 52 ' series 7 black flat panel lcd …	NULL	$(a_{T1}, b_{T1}, \text{nan})$	b_{T1}	samsung ln52a750 …	dynamic contrast ratio 120hz 6ms respons …	2148.99
a_{T2}	sony 46 ' bravia …	bravia z series …	NULL	$(a_{T2}, b_{T2}, \text{nan})$	b_{T2}	sony bravia …	ntsc 16:9 1366 x 768 …	597.72
a_{T3}	linksys wirelessn …	security router …	NULL	$(a_{T3}, b_{T1}, \text{nan})$	b_{T3}	ntsc 16:9 1366 x 768 …	54mbps	NULL

(b)不带标记的目标域

图 2 域适应的实体解析示例

Fig. 2 Example of domain adaptation for entity resolution

3.2 深度实体解析

现有的深度实体解析解决方案[4-6]，通常由特征提取和特征匹配两部分组成。如图 3 所示，具体来说，给定一个实体对 (a, b) 和特征提取函数 $F(a, b): A \times B \rightarrow R^d$，将这个实体对转化为 d 维向量（又称为特征），用 x 表示，即 $x = F(a, b)$。然后，特征 x 将作为特征匹配函数 M 的输入，这是一个基于深度学习的二元分类模型，然后输出预测实体对匹配的概率 $\hat{y}$，即：

$$\hat{y} = M(x) = M(F(a, b)) \tag{1}$$

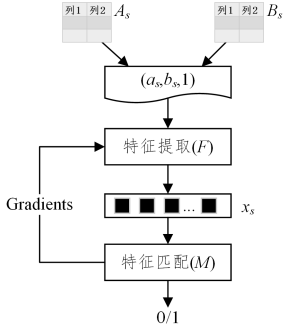

图 3 深度实体解析框架

Fig. 3 Framework of deep entity resolution

随后，通过迭代小批量随机梯度下降来优化 F 和 M 的参数，从而让模型更好地区分匹配实体对和非匹配实体对。

4 基于半监督域适应的实体解析模型

针对在深度实体解析中往往需要大量标记数据的问题，可以通过域适应利用已经标记好的数据集作为源数据集来训练深度实体解析模型，用于未标记目标数据集的实体对匹配工作。同时，为了让模型更好地在目标数据集上学习，提出了一种基于差异的域适应和半监督学习的实体解析模型（SSDAER），充分利用源域带标记数据和目标域未标记数据来进行学习。

图 4 给出了 SSDAER 的模型框架。该模型主要包括两个阶段：第一阶段，通过在源域进行有监督学习来训练分类器，同时使用统计指标 MMD 来最小化源域和目标域之间的距离，利用域标签训练一个能提取域公共特征的编码器；第二阶段，利用第一阶段训练好的分类器，来给目标域打上软伪标签。将伪标签数据加入源域，用于进一步训练分类器，并重新设计了损失函数，以更好地学习目标域特征。

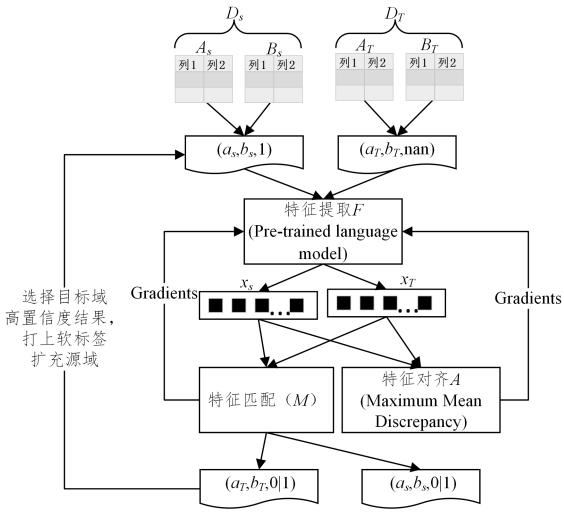

图 4 SSDAER 模型的框架

Fig. 4 Framework of SSDAER

4.1 第一阶段——域适应

特征提取（Feature Extractor，F），即表示学习，将一个实体对 $(a, b) \in A \times B$ 转换为基于向量的表示 $\boldsymbol{x}$。现有的深度实体解析研究通常利用循环神经网络[4-5]（Recurrent Neural Network，RNN）和预训练语言模型[6,47]（Pre-trained Language Models，LMs）来进行特征提取。Tu 等已经通过实验证明在基于域适应的实体解析问题中，LMs 的表现优于 RNN[8]。因此，我们选取了 LMs 来进行特征提取。代表性的 LMs 包括 Bert[48]，RoBerta[49] 和 DistilBert[50] 等。下面以 Bert 为例，来说明它如何将实体对 (a, b) 编码为基于向量的表示。

以图 2(a)中的数据集为例，给定实体 a，首先应用式(2)所示的函数将 a 的所有属性值对 $\{(attr_i, val_i)\}(1 \leqslant i \leqslant k)$ 转化为标记序列（即文本）。

$$S(a) = [ALL]attr_1[VAL]val_1 \cdots [ALL]attr_k[VAL]val_k \tag{2}$$

其中，$[ALL]$ 和 $[VAL]$ 是分别用于起始属性和值的两个特殊

标记。例如，将实体a_{S1}序列化为一个标记序列，即$S(a_{S1})=[ALL]title[VAL]balt\cdots[ALL]price[VAL]239.88$。然后将$(a_s,b_s)$转换为标记序列$S(a_S,b_S)=[CLS]S(a_S)[SEP]S(b_S)[SEP]$，其中$[SEP]$是一个特殊标记，用于分隔两个实体，$[CLS]$是 Bert 中用于编码整个序列的特殊标记。最后，使用全连接层来实现特征匹配(Feature Matcher，M)，这是基于深度实体解析中最常见的选择(如 DeepMatcher[4]，DeepER[5]，Ditto[6])，然后从$\boldsymbol{x}$产生匹配概率$\hat{y}$。

对于域适应，本文选择的是基于差异的域适应，即通过统计指标 MMD 来最小化源域和目标域之间的分布差异，从而实现特征对齐(Feature Aligner，A)。在训练过程中，D_S和D_T作为F的输入，输入特征$\boldsymbol{x}_S$和$\boldsymbol{x}_T$，它们对应的特征空间分别是p_S和p_T。然后，用 MMD 计算出p_S和p_T之间的分布差异，记为L_{MMD}。

$$L_{\mathrm{MMD}}=\sup_{\|\phi\|_H\leqslant 1}\|E_{x_S\sim p_S}[\phi(x_S)]-E_{x_T\sim p_T}[\phi(x_T)]\|_H^2 \tag{3}$$

其中，ϕ表示将$\boldsymbol{x}_S$和$\boldsymbol{x}_T$映射到再生核希尔伯特空间(RKHS)的核函数，$\|\phi\|_H\leqslant 1$定义了$RKHS(H)$的单位球中的一组函数。当p_S和p_T的分布相同时，L_{MMD}为 0。

同时，M给出源域的标签预测，并基于交叉熵损失来计算源域上的匹配损失，记为L_{M1}。

$$L_{\mathrm{M1}}=E_{(x_S,y_S)\sim(D_S,y_S)}[L_{\mathrm{CE}}(M(F(x_S)),y_S] \tag{4}$$

其中，L_{CE}表示交叉熵损失函数。总的优化目标是同时减少L_{MMD}和L_{M1}，以学习域不变和匹配特征为目标。因此第一阶段总的损失函数为：

$$L_{\mathrm{T1}}=L_{\mathrm{M1}}+\beta L_{\mathrm{MMD}} \tag{5}$$

其中，β是权衡匹配损失和域特征对齐损失的超参数。

4.2 第二阶段——半监督域适应

在第一阶段训练完成之后，将目标域的预测结果作为软伪标签加入源域再次进行训练。这个过程中，选择高质量的伪标签影响是提高第二阶段训练性能的重要因素之一。因此，我们的目标是减少所选样本中存在的噪声，以提高整体性能。选择伪标签的一种直接方法是选择具有高置信度预测的样本。然而，不正确的预测在校准不佳的网络中可能具有很高的置信度分数[51]。此外，如果第一阶段已经预测了一些高置信度样本，那么这些样本对第二阶段训练来说几乎没有好处[52]。基于可以利用预测不确定性来抵消不良校准影响的观察结果[51]，在用置信度筛选标签的同时，还采用了不确定性感知伪标签选择策略。形式上，不确定性可以分为认知不确定性和任意不确定性。前者来自模型参数的不确定性，后者是数据固有的不确定性(例如，不同类别的两个样本相似)。我们只关注认知不确定性，使用 MC-Dropout[53]，通过计算随机前向传播的固定数量(本文实验中取 10 次)的标准差来获得不确定性的度量。

综合来看，将经过 10 次 MC-Dropout[53]之后的输出进行 softmax 层处理后的概率分布的平均值作为置信度，将标准差作为不确定性的度量，以此综合筛选出高质量的伪标签，减少伪标签中的噪声。

设$g_C^i=[g_0^i,g_1^i]\subseteq\{0,1\}$表示样本$i$中类别$C$的伪标签选择情况的二值向量。当类别$C$的标签$g_C^i$被选择时，$g_C^i=1$；当类别$C$的标签$g_C^i$未被选择时，$g_C^i=0$。该向量由式(6)得到：

$$g_C^i=\mathbb{I}[p_c^i\geqslant\tau_p]\,\mathbb{I}[u(p_c^i)\leqslant k_p] \tag{6}$$

其中，p_c^i表示样本i属于预测为类别C的概率，$u(p_c^i)$表示预测p_c^i的不确定性，τ_p和k_p表示标签预测的置信阈值和预测的不确定性阈值。

将筛选出来的预测结果作为目标域上未标记样本的软标签。我们认为对伪标签的初步筛选可以在一定程度上减少标签噪声，同时用概率分布作为软标签代替传统的硬标签，这比直接用网络输出类或使用最近邻图上的标签传播估计的类来作为硬伪标签性能更好[16]。同时，匹配损失L_{M}随之发生变化，分为两部分，一部分是源域的硬标签损失L_{M1}，另一部分是目标域的软标签损失L_{M2}。

$$L_{\mathrm{M2}}=E_{(x_T,y_T)\sim(D_T,y_T)}[L_{\mathrm{CE}}(M(F(x_T)),\hat{y}_T] \tag{7}$$

使用两个正则化[16]来提高收敛性。当网络预测大多不正确且模型倾向于预测同一类以最小化损失时，第一个正则化用于处理在早期训练阶段收敛困难的问题。通过添加以下内容，不鼓励将所有样本分配给一个类。

$$R_A=\sum_{c=1}^{C}p_c\log\left(\frac{p_c}{\bar{h}}\right) \tag{8}$$

其中，p_c是c类的先验概率分布，$\bar{h}$表示数据集中所有被预测为c类的平均 softmax 概率。假设先验概率的分布为均匀分布，即$p_c=\frac{1}{C}(C=2)$。R_A代表对正类和负类分别进行正则化。第二个正则化是为了将每个软伪标签的概率分布集中在一个单独的类别上，从而避免网络因为弱指导而陷入局部最优解的情况。

$$R_H=-\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C}h_\theta^c(x_i)\log(h_\theta^c(x_i)) \tag{9}$$

其中，$h_\theta^c(x)$是模型输出的c类的 softmax 概率，N是样本总数(但这里用 mini-batch 来作近似替换)。第二个正则化是每个样本的平均熵(R_H代表熵正则化)，因此带正则化的软标签损失L_{M3}和匹配损失L_M表示为：

$$L_{M3}=L_{M2}+\lambda_A R_A+\lambda_H R_H \tag{10}$$

$$L_M=L_{M1}+\theta L_{M3} \tag{11}$$

其中，λ_A和λ_H用于权衡两个正则化的影响，θ用于权衡硬标签损失和软标签损失。

但是网络预测有时是不正确的。当将不正确的预测用作未标记样本的标签时，这种情况会得到加强，噪声无法通过伪标签筛选彻底排除。如果能通过数据增强，降低网络对其预测的信心，则可能会缓解这个问题并提高泛化能力。受 Mixup[18]启发，对目标域中提取的特征进行线性插值，同时引入正则化技术[16]，将数据增强与标签平滑相结合，这可能有助于提升模型在新的数据上的泛化性能。Mixup[18]在从目标域中随机选取的特征对$(x_{T,p},x_{T,q})$和相应的伪标签$(\hat{y}_{T,p},\hat{y}_{T,q})$的凸组合上进行训练。

$$x_i=\mu x_{T,p}+(1-\mu)x_{T,q} \tag{12}$$

$$y_i=\mu\hat{y}_{T,p}+(1-\mu)\hat{y}_{T,q} \tag{13}$$

其中，$p\neq q$；$\mu\in\{0,1\}$是从 Beta 分布$Be(\alpha,\alpha)$中随机抽取的，以支持训练样本之间的线性行为，减少远离它们的区域的振荡。此外，式(6)中的损失L_{M2}可以重新定义为：

$$L_{M2}=-\sum_{i=1}^{N}\mu[\hat{y}_{i,p}\log(M(F(x_i)))]+ \tag{14}$$

$(1-\mu)[\hat{y}_{i,q}\log(M(F(x_i)))]$

故第二阶段总的损失函数为：

$$L_{T2}=L_M+\beta L_{MMD} \tag{15}$$

总的优化目标是在减少源域的硬标签损失L_{M1}和目标域的软标签损失L_{M3}的同时，最小化源域和目标域之间的分布差异L_{MMD}。

本文提出的基于伪标签的域适应实体解析算法(SSDAER)的流程如算法1所示。

算法1 SSDAER算法

输入：标记源数据集D_S，未标记目标数据集D_T，对齐损失权值β，伪标签损失权值θ，置信阈值和预测的不确定性阈值τ_p和k_p，迭代次数N_1和N_2

输出：F，M

1. 初始化 F，M
2. for t←1 to N_1 do
3. for(x_S，x_T)in (D_S，D_T)
4. $\theta_{M1}=\underset{\theta_M}{\operatorname{argmin}}L_{M1}$
5. $\theta_{F1}=\underset{\theta_F}{\operatorname{argmin}}(L_{M1}+\beta L_{MMD})$
6. end for
7. end for
8. 用 M 预测目标域得到软伪标签$\hat{y}_T$
9. 用式(6)筛选标签得到$D_T'=(x_T,\hat{y}_T)$
10. $D_S\leftarrow D_S\cup D_T'$
11. 用θ_{M1}和θ_{F1}初始化 F 和 M
12. for t←1 to N_2 do
13. for(x_S，x_T)in (D_S，D_T)
14. $\theta_{M2}=\underset{\theta_M}{\operatorname{argmin}}L_M$
15. $\theta_{F2}=\underset{\theta_F}{\operatorname{argmin}}(L_M+\beta L_{MMD})$
16. end for
17. end for
18. 根据θ_{M2}和θ_{F2}更新 F 和 M，用于预测目标域

5 实验结果以及分析

5.1 数据集

实验使用了来自 DeepMatcher[4] 和 Magellan[3] 的基准数据集，这些数据集涵盖了各种领域，如商品、引文、餐厅等。每个数据集都包含来自两个关系表的实体，每个实体具有多个属性，并且包含一组标记的匹配/不匹配实体对。以表1中的 DBLP-Scholar(D_S)数据集为例，从 DBLP 和 Scholar 分别提取了两个表，每个表都有4个对齐的属性(标题、作者、会议地点、年份)。该数据集包含28707个实体对，其中5347个实体对被标记为匹配，其余的实体对是非匹配的。此外，还考虑了4个 WDC 产品数据集[7]，这些数据集也在 Ditto[6] 中使用。WDC 数据集从电子商务网站收集，其中包含计算机、相机、手表和鞋子4个类别，其中每个类别有1100个标记实体对。由于 Zomato-Yelp 数据集的原始版本较简单，所有方法在该数据集上表现良好，我们按照 DeepMatcher[4] 的做法，使用了一个包含错误的 Zomato-Yelp 数据集进行评估。有关数据集的更多详细信息如表1所列。为了方便表达，使用$D_S\rightarrow D_T$表示源域和目标域。例如，$WA\rightarrow AB$ 表示 Walmart-Amazon(WA)是源数据集，Abt-Buy(AB)是目标数据集。

表1 实体解析数据集

Table 1 Datasets of entity resolution

数据集	领域	实体对	实体对(匹配)	属性
Walmart-Amazon(WA)	商品	10242	962	5
Abt-Buy(AB)	商品	9575	1028	3
DBLP-Scholar(DS)	引文	28707	5347	4
DBLP-ACM(DA)	引文	12363	2220	4
Fodors-Zagats(FZ)	餐馆	946	110	6
Zomato-Yelp(ZY)	餐馆	894	214	3
iTunes-Amazon(IA)	音乐	532	132	8
RottenTomatoes-IMDB(RI)	电影	600	190	3
Books2(B2)	书本	394	92	9
WDC-Computers(CO)	商品	1100	300	2
WDC-Cameras(CA)	商品	1100	300	2
WDC-Watches(WT)	商品	1100	300	2
WDC-Shoes(SH)	商品	1100	300	2

5.2 评估指标

遵循大多数实体解析研究[4-6]的做法，使用 $F1$ 分数来评估上述的预测结果。F1 分数是匹配实体对的精确率和召回率的调和平均值。具体来说，设 TP(True Positive)表示真正例，即模型预测为匹配的匹配实体对；设 FP(False Positive)表示假正例，即模型预测为匹配的非匹配实体对；设 FN(False Negative)表示假负例，即模型预测为非匹配的匹配实体对。基于此，可以计算精确率和召回率分别为 $P=TP/(TP+FP)$和$R=TP/(TP+FN)$，根据精确率 P 和召回率R可以计算 $F1$ 分数为 $F1=2\times P\times R/(P+R)$。

5.3 实验设置

如图4所示，为了评估 SSDAER 在实体解析上的性能，便于和 DADER[8] 进行比较，在数据集划分上保持和它一样的选择。选择一个数据集作为标记源域(D_S，y_S)，另一个数据集作为未标记目标域D_T。将目标数据集D_T按1:9分成验证集D_T^{val}和测试集D_T^{test}，请注意，目标域测试集标签没有用于 SSDAER。使用目标域验证集标签来选择超参数，如在候选集{0.001，0.01，0.1，1.5}中选择超参数β，通过对每个值进行实验，选择在验证集上 $F1$ 分数最高的值，由于$\mu\in\{0,1\}$是从 Beta 分布 $Be(\alpha,\alpha)$中随机抽取的，因此 SSDAER 所有实验重复了3次并报告平均值。按此思路，α设置为1，β设置为0.1，λ_A和λ_H均设置为1，θ根据源域和目标域的分布差异设置为0.1或0.0003。

第一阶段，使用20个 epoch 进行训练，并在验证集D_T^{val}上评估每个周期的模型的性能，选择 F1 分数最高的模型并保存相应的 F 和 M 的参数。第二阶段，用第一阶段保存的参数初始化 F 和 M，使用50个 epoch 进行训练，在验证集D_T^{val}上评估每个周期的模型的性能，选择 $F1$ 分数最高的模型并保存相应的 F 和 M，然后用保存的 F 和 M 进行预测。

5.4 基线模型

将本文模型 SSDAER[1)] 与4种不同的实体解析模型进行实验对比。

DADER-MMD：DADER[8] 是目前效果最优的实体解析

1) https://github.com/cainiao12306/SSDAER

域适应框架，探索了实体解析各种域适应方法的好处和局限性，其中基于 MMD 的域适应方法在众多数据集上优于框架中的其他方法。

DAME：提出了一种新的基于域适应的实体匹配方法，将任务知识从多个源域转移到目标域[46]。

ZeroER：目前效果最优的无监督实体匹配解决方案[54]。ZeroER 通过考虑几个特定于实体匹配的属性来为实体匹配定制高斯混合模型。

Ditto：目前效果最优的有监督深度实体解析方法[6]。Ditto 提供了一些优化，然而这些优化是基于领域知识的，不能推广到其他数据集。此外，它们并不总是优于基本版的 Ditto，因此只比较 Ditto 的基本版本。

5.5 实验结果

5.5.1 对比实验

模型 SSDAER 和 DADER-MMD 的对比结果如表 2 所列，其中同一对源数据集和目标数据集中最好的结果加粗标注，$\Delta F1$ 表示 SSDAER 与 DADER-MMD 的 F1 的差值。SSDAER 与 DAEM 和 ZeroER 的对比结果如图 5 所示。DAEM，ZeroER 和 DADER-MMD 的指标来自于原论文中的实验结果。

表 2 SSDAER 与 DADER-MMD 的结果对比

Table 2 Results comparison between SSDAER and DADER-MMD

Datasets	Similar domains						Different domain					
Source	WA	AB	DS	DA	ZY	FZ	RI	RI	IA	IA	B2	B2
Target	AB	WA	DA	DS	FZ	ZY	AB	WA	DA	DS	FZ	ZY
SSDAER	**76.6**	66.3	**97.2**	**91.8**	**94.1**	**79.1**	40.2	**45.6**	**94.5**	**90.8**	**81.3**	**96.8**
DADER-MMD	72.6	**71.10**	**97.2**	91.5	92.2	64.5	**43.6**	41.5	**94.5**	86.9	73.0	91.5
$\Delta F1$	4.0	−4.8	0	0.3	1.9	14.6	−3.5	4.1	0	3.9	8.3	5.3

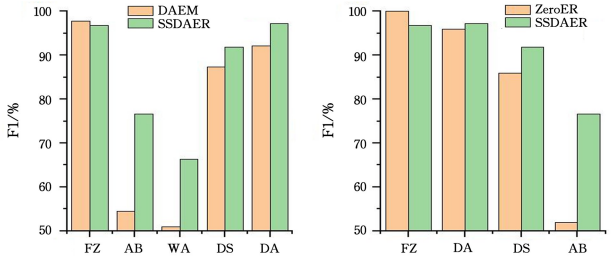

图 5 SSDAER 与 DAEM 和 ZeroER 的效果对比

Fig. 5 Performance comparison between SSDAER, DAEM and ZeroER

由结果可以看出，当源数据集和目标数据集来自相似的领域（如 WA→AB 都描述商品）时，除了 AB→WA，在其他数据集对上 SSDAER 的结果均优于 DADER-MMD；当源数据集和目标数据集来自不同的领域（如 RI→AB 中源域 RI 描述电影，目标域 AB 描述商品）时，除了 RI→AB，在其他的数据集对上 SSDAER 的结果均优于 DADER-MMD。综合来看，SSDAER 在大多数据集上普遍优于 DADER-MMD，使所有数据集上的 F1 值平均提升 2.84%。

将 SSDAER 中同一目标数据集（ZY→FZ 和 B2→FZ）中的最优值与 DAEM（zero-shot）和 ZeroER 进行比较。从图 5 中可以发现，除了在目标域 Fodors-Zagats（FZ）数据集与 DAEM（zero-shot）和 ZeroER 性能相当外，对于其他数据集，同样都是零样本情况下，SSDAER 在相应目标域数据集上的效果均优于 DAEM（zero-shot）和 ZeroER，其在所有数据集上的 F1 值平均提升了 9.16%和 7.1%。

在两个较大的目标域数据集上分别使用 SSDAER 与 Ditto 算法进行微调，遵循 Ditto 中的实验设置，目标数据集按 3∶1∶1 分为训练集、测试集和验证集。逐步加入训练集数据对 SSDAER 进行微调，并在测试集上记录 $F1$ 分数。如图 6 所示，当标签数量按比例逐渐增加时，SSDAER 的性能也随之上升。结果表明，为了达到良好的性能，Ditto 需要大量的标记数据，而 SSDAER 可以在少量甚至不需要标记数据的情况下达到与 Ditto 相当的性能。

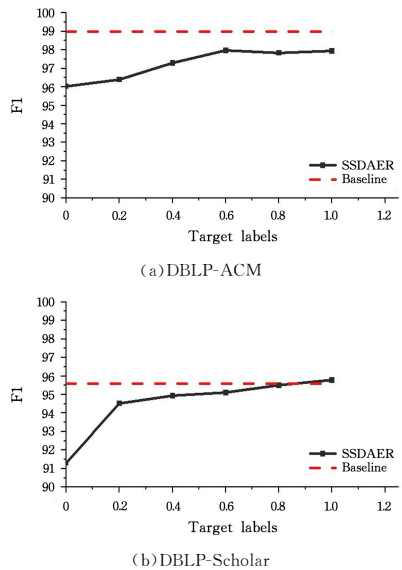

图 6 SSDAER 与 Ditto 的效果对比

Fig. 6 Performance comparison between SSDAER and Ditto

5.5.2 消融实验

为了进一步验证模型中不同模块的有效性，通过消融实验进行对比分析。为了分别验证域适应、域适应加半监督学习以及 Mixup 数据增强对解决实体解析任务性能的提升，我们将 SSDAER 与如下变体模型的结果进行比较。

NoDA：没有用域适应的实体解析模型，直接将在源域训练的模型用于目标域。

DAER：仅进行了第一节阶段域适应训练的实体解析模型。

NoMix：在第二阶段训练过程中没有用 Mixup 数据增强的实体解析模型，而是直接将在目标域上预测的软伪标签直接加入源域进行第二次训练。

如表 3、表 4 所列，$\Delta_1 F1$ 表示 DAER 相比 NoDA 方法 $F1$ 值的提升；$\Delta_2 F1$ 表示 SSDAER 相比 DAER 方法 F1 值的提升；$\Delta_3 F1$ 表示 SSDAER 相比 NoMix 方法 F1 值的提升。

同一对源数据集和目标数据集中最好的结果加粗标注。

相同领域下模型各模块效果的比较如下：如表 3 所列，当源域和目标域来自同一领域时，从$\Delta_1 F1$的值可以看出，除了 FZ→ZY，在其他数据集上，域适应能明显提高实体解析模型的性能，相比无域适应直接训练，F1 平均提高了 4.03%；从$\Delta_2 F1$的值可以看出，SSDAER 相对于 DAER 的 F1 值提升并不明显且不形成一般性结论；从$\Delta_3 F1$的值可以看出，Mixup 数据增强对 SSDAER 性能的提升情况因数据集而异，在部分数据集上有一定的提升效果，但是不具有普适性。但是综合来看，F1 最高值都出现在 SSDAER 或 SSDAER(NoMix)，说明在相同领域，本文模型对基于域适应的实体解析性能提升存在一定的潜力。

表 3 模型各模块效果的比较

Table 3 Performance comparison of each module

Datasets	Similar domains						Different domain					
Source	WA	AB	DS	DA	ZY	FZ	RI	RI	IA	IA	B2	B2
Target	AB	WA	DA	DS	FZ	ZY	AB	WA	DA	DS	FZ	ZY
NoDA	67.4	53.1	96.0	86.4	91.2	72.7	12.4	9.2	92.4	88.2	39.2	60.4
DAER	77.0	64.7	97.1	92.2	93.7	66.0	33.8	39.5	92.5	89.4	50.6	91.2
NoMix	**76.9**	66.2	97.0	**92.7**	93.6	77.8	38.0	45.2	94.1	89.9	49.4	96.0
SSDAER	76.6	**66.3**	**97.2**	91.8	**94.1**	**79.1**	**40.2**	**45.6**	**94.5**	**90.8**	**81.3**	**96.8**
$\Delta_1 F1$	9.7	11.7	1.2	5.8	2.5	−6.7	21.5	30.3	0.1	1.2	11.4	30.9
$\Delta_2 F1$	−0.4	1.6	0.0	−0.4	0.3	13.1	6.4	6.2	2.1	1.4	30.8	5.5
$\Delta_3 F1$	−0.3	0.1	0.2	−0.9	0.4	1.3	2.2	0.5	0.4	0.9	31.9	0.8

注：$\Delta_1 F1$表示 DAER 相比 NoDA 方法 F1 值的提升；$\Delta_2 F1$表示 SSDAER 相比 DAER 方法 F1 值的提升；$\Delta_3 F1$表示 SSDAER 相比 NoMix 方法 F1 值的提升。

相同领域不同类别下模型各模块效果的比较如下：如表 4 所列，为了进一步讨论本文模型在相同领域对实体解析问题的性能，对表 1 中的 WDC 数据集进行了实验。这 4 个 WDC 数据集不仅来自同一领域，而且具有相同的文本属性。从$\Delta_1 F1$可以看出，域适应对模型的提升并不明显，在不同数据集对上表现有所不同；但是从$\Delta_2 F1$可以看出，通过在源域加入目标域软伪标签进行第二阶段的训练，SSDAER 相对于 DAER 的 F1 值平均提升 1.82%，验证了本文模型对基于域适应的实体解析性能提升的有效性；从$\Delta_3 F1$的值可以看出，Mixup 数据增强对 SSDAER 性能提升并不明显。综合来看，当源域和目标域来自相同领域的不同类别时，伪标签学习能在一定程度上改进基于域适应的实体解析模型。

表 4 相同领域不同类别下模型各模块效果的比较

Table 4 Performance comparison of each module in different categories of domains

Datasets	Source	CO	WT	CA	WT	SH	WT	CO	SH	CA	SH	CO	CA
	Target	WT	CO	WT	CA	WT	SH	SH	CO	CA	WT	CA	CO
NoDA		84.1	80.3	81.3	87.2	**85.8**	73.9	68.7	83.7	**69.5**	81.7	84.9	81.3
DAER		83.2	86.6	86.7	87.0	80.0	76.9	76.7	83.2	63.6	86.9	85.4	86.9
NoMix		79.3	86.7	**87.4**	86.7	84.4	**77.2**	77.2	**86.7**	69.1	**87.9**	86.7	87.4
SSDAER		**87.1**	**86.9**	87.3	**88.1**	83.6	77.0	**77.7**	86.4	68.2	87.3	**87.2**	**88.1**
$\Delta_1 F1$		−0.9	6.3	5.5	−0.1	−5.8	3.0	8.1	−0.6	−6.0	5.3	0.5	5.6
$\Delta_2 F1$		4.0	0.2	0.5	1.0	3.6	0.1	1.0	3.3	4.6	0.4	1.8	1.3
$\Delta_3 F1$		7.9	0.2	−0.2	1.4	−0.8	−0.2	0.5	−0.3	−0.9	−0.6	0.5	0.7

不同领域下模型各模块效果的比较如下：如表 3 所列，当源域和目标域来自不同领域时，从$\Delta_1 F1$的值可以看出，域适应能通过域对齐显著提高实体解析模型的性能，相比无域适应直接训练，F1 平均提高了 15.9%；从$\Delta_2 F1$的值可以看出，通过在源域加入目标域软伪标签进行第二阶段的训练，SSDAER 相对于 DAER 的 F1 值平均提升了 8.73%；从$\Delta_3 F1$的值可以看出，Mixup 数据增强对 SSDAER 性能在各个数据集对上均有明显的提升，F1 值平均提升了 6.12%。综合来看，F1 的最高值都出现 SSDAER，验证了在不同领域本文模型对基于域适应的实体解析性能提升的有效性。

为了进一步展示所提方法的工作原理，帮助我们理解域适应如何减少域转移，以 WA→AB 为例，使用 t-SNE[55] 将源数据集和目标数据集的特征映射到二维空间。其分布如图 7 和图 8 所示，其中红色和灰色点分别表示源实体对和目标实体对。

对比图 7 和图 8 可以发现，当应用域对齐之后，源实体对和目标实体对明显更加混合。结果表明，域对齐可以引导特征提取器 F 让源域和目标域的特征分布更加接近，从而帮助匹配器 M 对目标域进行正确的分类。

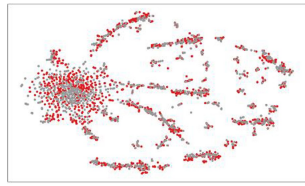

图 7 SSDAER：源域和目标域分布(电子版为彩图)

Fig. 7 SSDAER: distribution of source domain and target domain

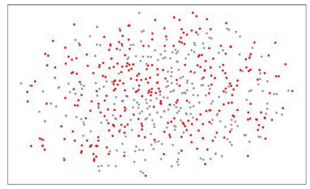

图 8 NoDA：源域和目标域分布(电子版为彩图)

Fig. 8 NoDA: distribution of source domain and target domain

本文提出的 SSDAER 模型，通过域对齐来减小源域和目标域的分布差异，同时经过目标域软伪标签迭代训练之后，能进一步学习目标域中的特征，大大提升了源域训练的模型在

目标域上的性能。尤其是当源域和目标域来自不同领域时,SSDAER 对模型在目标域的表现的提升尤为明显,域适应和伪软标签都能大大提升 F1 指标。同时,通过消融实验,验证了 Mixup[18] 数据增强方法在源域和目标域差距较大时对模型的增益。但是当源域和目标域来自不同领域,只有源域和目标域相似度足够高时,软伪标签训练的效果才比较明显,其他情况对模型性能的提升不显著,而且二次训练会增加训练开销,这也是本文模型的不足之处。在之后的研究中可以探究软伪标签和其他域适应方法相结合的方法在实体解析相同领域上的性能。

结束语 针对现有的深度实体解析模型需要大量标注数据的问题,本文提出了一种基于半监督学习的域适应实体解析算法。该模型充分利用了源域带标签数据和目标域无标签数据进行学习,以此解决传统深度实体解析模型需要大量标签的问题。第一阶段,在源域进行有监督学习训练分类器,同时使用统计指标(最大均值差异)来最小化源域和目标域之间的距离,利用域标签训练一个能提取域公共特征的特征提取器;第二阶段,利用第一阶段训练好的特征提取器和分类器,来给目标域打上软伪标签。将伪标签数据用于进一步训练分类器,并设计了新的损失函数,以更好地学习目标域特征。本文模型的 F1 指标均高于目前最先进的无监督实体解析模型和域适应实体解析模型,获得了更好的匹配效果;相比目前最优的有监督实体解析模型,SSDAER 可以用更少的标签达到最优的性能。但是当源域和目标域来自相同领域时,半监督域对齐对模型的增益并没有很明显且不稳定,在之后的研究中会探索更加有效的域适应实体解析模型,将源域的知识更好地迁移到目标域中,以进一步提高模型在目标域中的表现。

参 考 文 献

[1] SINGH R,MEDURI V V,ELMAGARMID A,et al. Synthesizing entity matching rules by examples[J]. Proceedings of the VLDB Endowment,2017,11(2):189-202.

[2] BILENKO M,MOONEY R J. Adaptive duplicate detection using learnable string similarity measures[C]// Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2003:39-48.

[3] DOAN A H,KONDA P,SUGANTHAN GC P,et al. Magellan:toward building ecosystems of entity matching solutions[J]. Communications of the ACM,2020,63(8):83-91.

[4] MUDGAL S,LI H,REKATSINAS T,et al. Deep learning for entity matching:A design space exploration[C]// Proceedings of the 2018 International Conference on Management of Data. 2018:19-34.

[5] EBRAHEEM M,THIRUMURUGANATHAN S,JOTY S,et al. Distributed representations of tuples for entity resolution[J]. Proceedings of the VLDB Endowment,2018,11(11):1454-1467.

[6] LI Y,LI J,SUHARA Y,et al. Deep entity matching with pre-trained language models[J]. arXiv:2004.00584,2020.

[7] PRIMPELI A,PEETERS R,BIZER C. The WDC training dataset and gold standard for large-scale product matching[C]// Companion Proceedings of The 2019 World Wide Web Conference. 2019:381-386.

[8] TU J,FAN J,TANG N,et al. Domain adaptation for deep entity resolution[C]// Proceedings of the 2022 International Conference on Management of Data. 2022:443-457.

[9] ARAZO E,ORTEGO D,ALBERT P,et al. Unsupervised label noise modeling and loss correction[C]// International Conference on Machine Learning. PMLR,2019:312-321.

[10] OLIVER A,ODENA A,RAFFEL C A,et al. Realistic evaluation of deep semi-supervised learning algorithms[C]// Proceedings of the Advances in Neural Information Processing Systems. 2018:3239-3250.

[11] ZHANG Z,RINGEVAL F,DONG B,et al. Enhanced semi-supervised learning for multimodal emotion recognition[C]// 2016 IEEEInternational Conference on Acoustics,Speech and Signal Processing(ICASSP). IEEE,2016:5185-5189.

[12] GONZÁLEZ M,BERGMEIR C,TRIGUERO I,et al. Self-labeling techniques for semi-supervised time series classification:an empirical study[J]. Knowledge and Information Systems,2018,55:493-528.

[13] MIYATO T,DAI A M,GOODFELLOW I. Adversarial training methods for semi-supervised text classification[J]. arXiv:1605.07725,2016.

[14] LI Y,LIU L,TAN R T. Certainty-driven consistency loss for semi-supervised learning[J]. arXiv. 1901.05657,2019.

[15] SAJJADI M,JAVANMARDI M,TASDIZEN T. Regularization with stochastic transformations and perturbations for deep semi-supervised learning[C]// Proceedings of the 30th International Conference on Neural Information Processing Systems. 2016:1171-1179.

[16] ARAZO E,ORTEGO D,ALBERT P,et al. Pseudo-labeling and confirmation bias in deep semi-supervised learning[C]// 2020 International Joint Conference on Neural Networks(IJCNN). IEEE,2020:1-8.

[17] ISCEN A,TOLIAS G,AVRITHIS Y,et al. Label propagation for deep semi-supervised learning[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019:5070-5079.

[18] ZHANG H,CISSE M,DAUPHIN Y N,et al. mixup:Beyond empirical risk minimization[J]. arXiv:1710.09412,2017.

[19] SINGH R,MEDURI V,ELMAGARMID A,et al. Generating concise entity matching rules[C]// Proceedings of the 2017 ACM International Conference on Management of Data. 2017:1635-1638.

[20] SUN B,FENG J,SAENKO K. Return of frustratingly easy domain adaptation[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2016.

[21] CHAI C,LI G,LI J,et al. Cost-effective crowdsourced entity resolution:A partial-order approach[C]// Proceedings of the 2016 International Conference on Management of Data. 2016:969-984.

[22] CHAI C,LI G,LI J,et al. A partial-order-based framework for cost-effective crowdsourced entity resolution[J]. The VLDB Journal,2018,27:745-770.

[23] CUI L,CHEN J,HE W,et al. Achieving approximate global optimization of truth inference for crowdsourcing microtasks[J]. Data Science and Engineering,2021,6(3):294-309.

[24] LI G,CHAI C,FAN J,et al. CDB:A crowd-powered database system[J]. Proceedings of the VLDB Endowment,2018,11(12):1926-1929.

[25] AZZALINI F,JIN S,RENZI M,et al. Blocking techniques for entity linkage:A semantics-based approach[J]. Data Science and

Engineering, 2021, 6: 20-38.
[26] CHRISTEN P. Automatic record linkage using seeded nearest neighbour and support vector machine classification[C]// Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2008: 151-159.
[27] MCCALLUM A, WELLNER B. Conditional models of identity uncertainty with application to noun coreference[C]// Proceedings of the 17th International Conference on Neural Information Processing Systems. 2004: 905-912.
[28] YAO D, GU Y, CONG G, et al. Entity resolution with hierarchical graph attention networks[C]// Proceedings of the 2022 International Conference on Management of Data. 2022: 429-442.
[29] GANIN Y, USTINOVA E, AJAKAN H, et al. Domain-adversarial training of neural networks[J]. The Journal of Machine Learning Research, 2016, 17(1): 2096-2030.
[30] LIU T, FAN J, LUO Y, et al. Adaptive data augmentation for supervised learning over missing data[J]. Proceedings of the VLDB Endowment, 2021, 14(7): 1202-1214.
[31] LONG M, CAO Y, WANG J, et al. Learning transferable features with deep adaptation networks[C]// International Conference on Machine Learning. PMLR, 2015: 97-105.
[32] TANG H, JIA K. Discriminative adversarial domain adaptation [C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020: 5940-5947.
[33] TZENG E, HOFFMAN J, SAENKO K, et al. Adversarial discriminative domain adaptation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 7167-7176.
[34] KUMAGAI A, IWATA T, FUJIWARA Y. Transfer metric learning for unseen domains[J]. Data Science and Engineering, 2020, 5: 140-151.
[35] CHOI Y, CHOI M, KIM M, et al. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 8789-8797.
[36] GE Y, CHEN D, LI H. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification[J]. arXiv: 2001. 01526, 2020.
[37] LONG M, ZHU H, WANG J, et al. Deep transfer learning with joint adaptation networks[C]// International Conference on Machine Learning. PMLR, 2017: 2208-2217.
[38] SUN B, FENG J, SAENKO K. Return of frustratingly easy domain adaptation[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2016.
[39] ZELLINGER W, GRUBINGER T, LUGHOFER E, et al. Central moment discrepancy(cmd) for domain-invariant representation learning[J]. arXiv: 1702. 08811, 2017.
[40] GANIN Y, USTINOVA E, AJAKAN H, et al. Domain-adversarial training of neural networks[J]. The journal of machine learning research, 2016, 17(1): 2096-2030.
[41] TZENG E, HOFFMAN J, SAENKO K, et al. Adversarial discriminative domain adaptation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 7167-7176.
[42] GHIFARY M, KLEIJN W B, ZHANG M, et al. Deep reconstruction-classification networks for unsupervised domain adaptation[C]// Computer Vision-ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV 14. Springer International Publishing, 2016: 597-613.
[43] THIRUMURUGANATHAN S, PARAMBATH S A P, OUZZANI M, et al. Reuse and adaptation for entity resolution through transfer learning[J]. arXiv: 1809. 11084, 2018.
[44] KASAI J, QIAN K, GURAJADA S, et al. Low-resource deep entity resolution with transfer and active learning[J]. arXiv: 1906. 08042, 2019.
[45] TU J, FAN J, TANG N, et al. Domain adaptation for deep entity resolution[C]// Proceedings of the 2022 International Conference on Management of Data. 2022: 443-457.
[46] TRABELSI M, HEFLIN J, CAO J. DAME: Domain Adaptation for Matching Entities[C]// Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining. 2022: 1016-1024.
[47] TANG N, FAN J, LI F, et al. RPT: relational pre-trained transformer is almost all you need towards democratizing data preparation[J]. arXiv: 2012. 02469, 2020.
[48] LEE J, TOUTANOVA K. Pre-training of deep bidirectional transformers for language understanding [J]. arXiv: 1810. 04805, 2018.
[49] LIU Y, OTT M, GOYAL N, et al. Roberta: A robustly optimized bert pretraining approach[J]. arXiv: 1907. 11692, 2019.
[50] SANH V, DEBUT L, CHAUMOND J, et al. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter[J]. arXiv: 1910. 01108, 2019.
[51] RIZVE M N, DUARTE K, RAWAT Y S, et al. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning[J]. arXiv: 2101. 06329, 2021.
[52] MUKHERJEE S, AWADALLAH A. Uncertainty-aware self-training for few-shot text classification[J]. Advances in Neural Information Processing Systems, 2020, 33: 21199-21212.
[53] GAL Y, GHAHRAMANI Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning[C]// International Conference on Machine Learning. PMLR, 2016: 1050-1059.
[54] WU R, CHABA S, SAWLANI S, et al. Zeroer: Entity resolution using zero labeled examples[C]// Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data. 2020: 1149-1164.
[55] VAN DER MAATEN L. Barnes-hut-sne[J]. arXiv: 1301. 3342, 2013.

DAI Chaofan, born in 1973, Ph. D, professor. His main research interests include big data analytics and data quality.

DING Huahua, born in 1999, postgraduate. His main research interests include entity resolution and data integration.

(责任编辑:柯颖)