一种多源领域自适应命名实体识别方法

李佳芮，刘健 ，陈钰枫*，徐金安，张玉洁

（北京交通大学计算机与信息技术学院，北京 100044）

摘要：领域自适应是解决低资源问题的一种通用方式，可应用于各种自然语言处理的任务中。当前针对命名实体识别任务的领域自适应研究通常从单一的源领域迁移到目标领域，在目标领域和源领域相近的情况下，这种方式能够取得一定的效果，但是在目标领域与源领域相关度不高的情况下，单一领域迁移方式存在很大的局限性。针对这一问题，我们提出一种融合多源领域贡献度加权的命名实体识别自适应模型，通过多个领域的知识迁移来提升目标领域的实体识别性能，并针对不同领域及其领域内样本对目标领域的重要性程度进行建模，加权到命名实体识别模型中，以此来实现更好的模型领域适应性。最终在多个目标领域上进行了实验，性能皆优于当前性能最好的方法，验证了模型的有效性。

关键词：命名实体识别；领域自适应；贡献度加权；多源

中图分类号：TP391　　**文献标志码：** A

命名实体识别（Named entity recognition，NER）旨在识别文本中特定类型的实体（如人名、地名、组织机构名、专有名词等），是自然语言处理中一项重要的基础任务，可广泛应用于信息抽取、问答系统、机器翻译等上游任务中[1-3]。随着深度学习技术的不断深入，基于神经网络的命名实体识别模型研究已经取得了不错的进展[4]，然而，这些方法通常需要依靠大规模的标注语料来提升性能，很多低资源领域难以获取到如此大规模的标注语料，限制了现有方法的性能。为解决这一问题，领域自适应这一研究课题随之诞生，它的研究目标是利用拥有丰富语料资源的源领域来提升低资源领域的模型性能[5]。

当前，Peng 和 Dredze 应用了多任务学习来进行领域自适应[6]。Jia 和 Zhang 针对神经网络结构进行改进，提出一种多神经元合成的神经网络模型[7]。Liu 等人提出基于自适应预训练进行领域增强，同时在下游任务中进行微调[8]。尽管现有的命名实体识别领域自适应方法已经取得了一定的研究进展，但仍存在以下问题：1）大多数自适应方法从单一源领域迁移到目标领域（源领域通常为新闻领域）。当目标领域和新闻领域差异较大时，模型效果并不理想。2）当前基于预训练的自适应方法通常需要大规模的目标域（相关）无标注语料对语言模型进行继续预训练，然而并非所有目标域都能满足这一条件。

考虑以上问题，本文提出了一种融合多源领域贡献度加权的自适应命名实体识别模型（Multi-Domain Adaptation based on Importance Weighting，MDAIW）。该模型基于多个源领域的知识迁移提升目标领域的模型效果，针对多源领域数据混合可能存在的分布不一致问题，设计了一种融合领域层级和样本层级贡献度的自适应方法，同时采用对抗训练策略联合训练命名实体识别任务和领域分类任务。通过贡献度加权，模型能够可以更好地迁移到目标领域，另一方面，为了降低对大规模目标领域数据的依赖，我们提出了一种两阶段领域自适应预训练的方法，它能够通过小规模的数据实现领域自适应预训练。我们在多个目标领域上进行了实验，结果表明本文提出的模型优于当前的性能最佳方法（SOTA）和其他基线模型，验证了方法的有效性。

本文的组织结构如下：第 1 节介绍本文提出的 MDAIW 模型架构和原理；第 2 节通过实验验证模型的有效性，并与现有研究方法进行比较；第 3 节介绍领域自适应的相关研究工作；最后对全文进行总结，并展望未来研究方向。

1 模型架构

基金项目：国家自然科学基金“面上”项目(61976016、61976015、61876198)，国家重点研发计划（2019YFB1405200）。

* **通信作者：**chenyf@bjtu.edu.cn

我们用S和T分别表示源领域和目标领域。源领域可通过公式表示为$S_i = \sum_1^k \{x_j^{S_i}, y_j^{S_i}\}_{j=1}^{m_i}$，其中$i$表示不同的源领域，每个源领域有$m_i$个样本，共有$k$个领域。同样，目标领域可表示为$T = \{x_j^T, y_j^T\}_{j=1}^n$。融合多源领域的自适应命名实体识别的目标是使用多个源领域的知识来提升目标领域的实体识别性能。

模型的整体架构如图 1 所示，它可以分为两部分，第一部分是两阶段领域自适应预训练，考虑到目标领域语料匮乏的情况，我们设计了一种两阶段自适应预训练的方式，目的是使经过预训练语言模型获取的文本向量融合源领域和目标领域信息。第二部分是融合多源领域贡献度加权的迁移模型，通过计算领域层级和样本层级的贡献度参数，并加权到命名实体识别任务的损失函数中来提升模型性能。下接下来对各部分进行详细阐述。

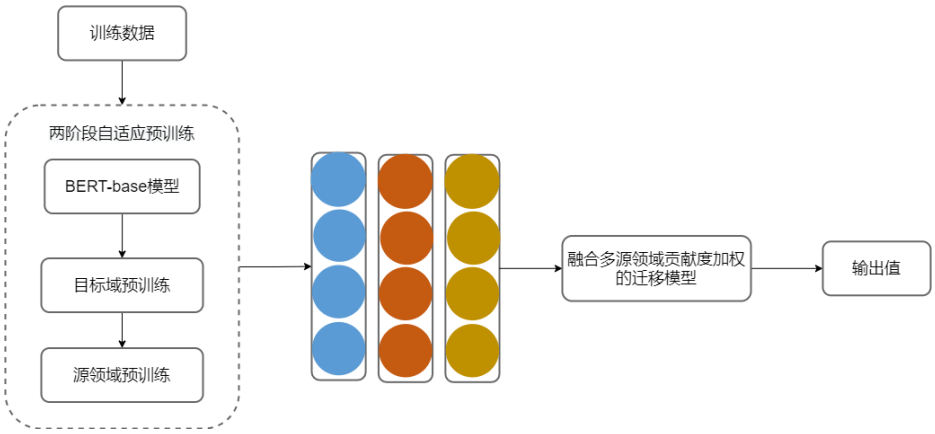

图 1 模型整体架构

Fig. 1 Overall framework of the model

1. 1 两阶段领域自适应预训练

由于直接使用预训练语言模型获取到的文本向量特征可能无法有效地捕捉到源领域和目标领域之间的差异，我们提出了一种两阶段自适应预训练方法，它的具体步骤如图 1 的左半部分所示。自适应预训练的训练过程利用了 BERT[9]中的掩码语言模型（masked language model），它通过上下文来预测被掩码的字段，从而获得深层的双向表示。目标领域和源领域的自适应细节如下。

1.1.1 目标领域自适应预训练

针对目标领域语料资源匮乏的情况，在目标领域预训练阶段，我们不需要像当前方法中那样额外搜集大量的目标领域文本，而是采用所有可用的目标领域命名实体识别语料进行训练。训练任务和参数都遵循 BERT-base 模型中的设置。 此外，我们添加了一个超参数来表示目标领域样本的数据重复次数，这一参数可以弥补数据不足的问题。具体来说，首先根据重复次数这一超参数对目标领域语料进行重复迭代，随后使用得到的数据对 BERT 进行继续预训练，得到经过目标领域自适应的语言模型。

1.1.2 源领域自适应预训练

我们希望在目标领域自适应预训练的基础上使文本向量表示进一步融合多个源领域的信息，因此设计了源领域自适应预训练。在这一阶段，我们使用多个源领域混合的未标注文本作为训练数据，对前一步中得到的目标领域自适应语言模型进行二阶段的继续预训练。同时，为了保持语言模型对源领域和目标领域的适应性，通过设置数据重复次数使源领域自适应预训练的训练数据规模小于目标领域自适应预训练中的训练数据规模。

1. 2 融合多源领域贡献度加权的迁移模型

考虑到不同的源领域以及领域内样本对目标领域的重要程度可能存在差别，我们设计了一种融合贡献度加权的迁移方法，以此来缓解源领域和目标领域间的领域偏移问题，同时，应用了对抗训

练策略进一步增强模型性能，融合多源领域贡献度加权的迁移模型如图 2 所示。

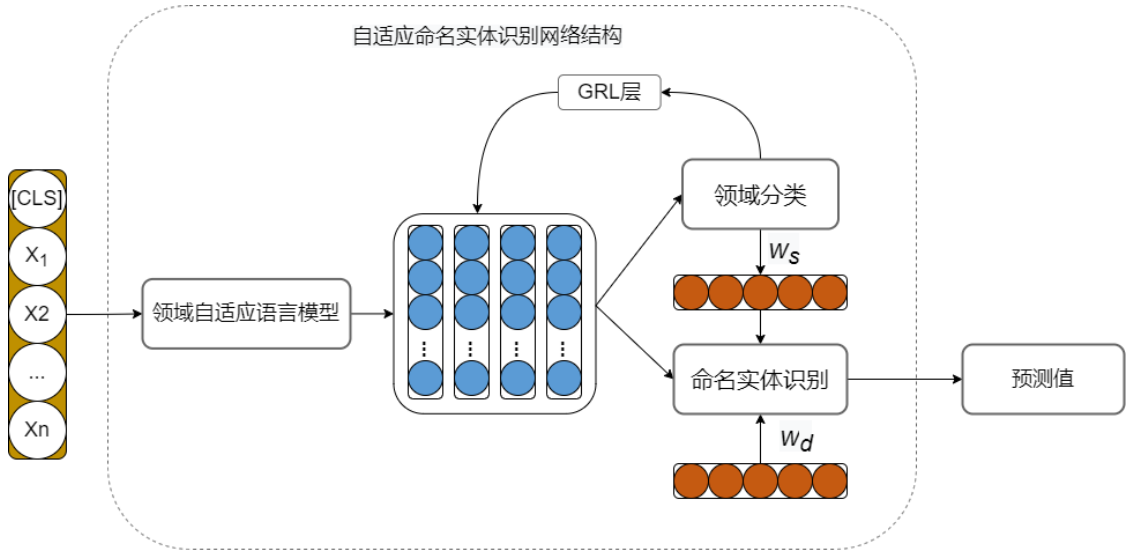

图 2 融合多源领域贡献度加权的自适应命名实体识别模型

Fig. 2 Modal of Multi-Domain Adaptation based on Importance Weighting

其中，输入向量由先前经过两阶段领域自适应预训练的语言模型获取，包含了源领域和目标领域的知识。命名实体识别的训练任务可通过公式表示为：

$$L_{NER} = -\frac{1}{m+n}\sum_{j=1}^{m+n} y_j \log \hat{y_j} \tag{1}$$

在 1.2.1 节中，详细介绍了领域层级和样本层级贡献度的建模计算方法，以及如何进行加权。在 1.2.2 节中介绍了对抗训练策略。

1.2.1 贡献度加权

贡献度加权的目标是量化计算不同领域和领域内样本对目标领域的重要程度，并将计算后得到的贡献度参数加权到命名实体识别模型中，以缓解在采用多源领域数据进行训练时，不同的领域分布导致的负迁移问题。我们分别设计了领域层级和样本层级的贡献度参数建模方法。

领域层级贡献度建模 领域层级贡献度的建模目标是计算不同源领域对目标领域的贡献程度。Aharoni and Goldberg 在研究中得出预训练语言模型获取的向量表示能够捕捉到领域信息这一结论[10]。在他们的研究基础上，我们提出通过两阶段自适应预训练后的语言模型获取向量表示，利用两个领域的向量表示来计算领域距离，并将其作为这两个领域相似性的量化表示。在具体实现过程中，首先通过 BERT-base 模型得到句子级别的向量表示，然后通过计算某一领域内句子向量的平均值作为该领域的向量表示，并计算两个领域向量的余弦相似度来表征领域距离。领域的向量表示和领域距离可通过公式表示为：

$$q_{s_i} = \frac{1}{m_i}\sum_{j=1}^{m_i} Bert(x_j) \tag{2}$$

$$d_{s_a,s_b} = cos\ similarity(q_{s_a}, q_{s_b}) \tag{3}$$

样本层级贡献度建模 样本层级贡献度的建模目标是计算领域内的不同样本对目标领域的贡献程度。我们设计了一个领域二分类器来实现这一目标，领域二分类器用于预测样本是否属于目标领域，二分类器将输出某一样本被预测为源领域样本和目标领域样本的概率值，如果某一源领域样本被预测为目标领域样本的概率值较低，则代表其对目标领域的重要程度比较低，相反，如果概率值较高，

则意味着样本对目标领域的重要程度较高，应赋予更高的权重。通过领域分类器即可得到源领域样本对于目标领域的贡献度权重。该过程可用公式（4）表示，其中，P_T和P_S分别表示被预测为目标领域样本和源领域样本的概率，P_T即为样本层级贡献度参数。

$$[P_T, P_S] = softmax(\text{classifier}(x_j)) \tag{4}$$

贡献度加权 在得到贡献度权重后，我们将其加权到命名实体识别模型的损失函数中。训练数据由少量的目标领域数据和多个源领域数据集构成。因此，如公式（5）所示，命名实体识别的损失函数可被分为两部分，L_S和L_T分别表示源领域和目标领域的损失函数，L的计算方法由公式（1）计算得到。

$$Loss = L_S + L_T \tag{5}$$

贡献度权重将加权到L_S中，每个领域内的样本共享该领域的贡献度权重，因此每个样本的权重由领域贡献度权重和样本贡献度权重点乘得出，可表示为公式（6）：

$$w = w_d \cdot w_s \tag{6}$$

贡献度加权的目标是体现源领域样本对于目标领域的重要程度，根据贡献度的高低来分配权重。因此，需要对权重进行归一化，归一化方法由公式（7）计算得出：

$$w = \frac{\exp w}{\sum_i \exp w_i} \tag{7}$$

对多个源领域内的样本加权后，融合多源领域贡献度加权的自适应命名实体识别模型的整体损失函数可表示为：

$$L = -\frac{1}{n}\sum_{j=1}^{n} y_j^T \log \widehat{y_j^T} - \frac{1}{m}\sum_{j=1}^{m} w_j y_j^S \log \widehat{y_j^S} \tag{8}$$

1.2.2 对抗训练

通过贡献度加权可以提高模型对不同领域和样本的鉴别能力，在此基础上，结合对抗训练可以进一步提高模型对多个不同的领域间通用知识的获取能力。受到梯度反转层（GRL）的启发[11]，我们在模型中加入 GRL 层来实现对抗训练。其实现方法是在反向传播的过程中反转梯度方向，使 GRL 前向和后向的网络训练目标相反，以此来实现对抗的效果。具体来说，我们对前述的领域分类器和命名实体识别任务进行联合训练，对 Bert 分类器 token（[CLS]）中的隐藏状态h_{CLS}进行参数优化。

通过隐层向量h_{CLS}获取领域分类器的输出值的过程可以使用公式（9）表示：

$$d = softmax(W_d h_{CLS} + b) \tag{9}$$

领域分类器的损失函数可以表示为：

$$L_{TC} = -\frac{1}{m+n}\sum_{j=1}^{m+n} d_j log\hat{d}_j \tag{10}$$

联合训练的损失函数可以表示为：

$$L_{total} = L_{NER} - L_{TC} \tag{11}$$

2 实验

2. 1 实验数据

我们采用 CONLL[12]和 CrossNER[8]两种数据集进行实验，数据集的统计数据可见表 1。其中，CONLL 被广泛应用于命名实体识别任务中，它对路透社文章收集到的数据进行标注，包含四种实体标签，分别为人名（PER）、地名（LOC）、组织机构名（ORG）和其他（MISC）。CroossNER 是一个由 Liu 等人通过 wiki 数据收集和人工标注构建的命名实体识别数据集，包含政治、科学、音乐、文学和人工智能五种领域，在该数据集中，每个领域数据集的标签都由该领域的特定实体构成，比如政治领域中包含“政治家”、“政党”等实体类别。我们将这些领域特有的实体标签统一替换为“领域词”标签，最终实验数据包含人名、地名、组织机构名、领域词、其他五种标签类型。我们将 CrossNER 中的五个领域数据集分别看作目标领域进行了实验，当一种领域作为目标领域时，其余领域被作为源领域。

表 1 各领域数据统计

Tab.1 Data Statistics for Each Domain

领域	训练集	开发集	测试集
新闻	14987	3466	3684
政治	200	541	651
科学	200	450	543
音乐	100	380	456
文学	100	400	416
人工智能	100	350	431

2. 2 评价指标

我们通过 F1 值对模型的有效性进行评估，F1 值同时考虑了准确率（P）和召回率（R），计算公式如下：

$$P = \frac{正确识别的实体数量}{识别出的实体数量} \times 100\% \tag{12}$$

$$R = \frac{正确识别的实体数量}{测试数据中的实体数量} \times 100\% \tag{13}$$

$$F1 = \frac{2P * R}{P + R} \tag{14}$$

2. 3 实验设置

在两阶段领域自适应预训练中，训练过程遵循 BERT-base 模型中的参数设置，训练轮次为 3 次，在目标领域预训练阶段，使用全部可用的目标域数据进行训练，重复次数设置为 10，在源领域预训练阶段，训练数据为其余的五个源领域，重复次数为 2，由于新闻领域的数据量较大，可能会影响目标领域预训练的效果，因此从新闻领域中随机筛选出与目标领域数据等量的数据与其余源领域数据

混合进行源领域预训练。在融合多源领域贡献度加权的自适应命名实体识别模型中，分类任务和命名实体识别任务分别遵循 BERT-base 在下游任务微调中给出的参考参数，训练轮次为 5 次。

2. 4 基线模型

本文模型方法与以下经典和当前效果最好的模型（state-of-the-art，SOTA）进行了对比，以此来验证本文提出模型的有效性。

基于 BERT 的微调：在 BERT-base 的基础上使用任务训练数据在下游任务上进行微调，我们分别使用目标语料和所有语料混合进行了实验。

基于 BERT 的 INIT[13]：我们在 BERT 的基础上实现了 INIT 方法，首先用源领域数据训练模型，随后使用目标领域数据进行微调。

基于片段的预训练语言模型自适应（SOTA）[8]：Liu 等人提出在自适应预训练语言模型的阶段进行领域增强，同时在下游任务的微调中进行了三种模式的实验，其中两种取得的效果最好，所以我们与这两种方法进行了对比。第一种方法与微调的思路一致，第二种方法对源领域和目标领域模型进行联合训练。

2. 5 实验结果

我们对多个目标领域进行了实验来验证本文模型的适用性，当一种领域作为目标领域时，其余领域为源领域。将最终的模型与基于 BERT 的微调方法、基于 BERT 的 INIT 方法以及当前效果最好的基于片段的预训练语言模型自适应方法进行了对比，结果如表 2 所示。

表 2 本文提出模型和基线模型的实验结果

Tab.2 Overall Results of Proposed Model Compared with Baseline Models

模型	政治	科学	音乐	文学	人工智能	平均值
Fine-tuned BERT（target only）	66.56	65.28	65.44	57.86	54.77	61.98
Fine-tuned BERT（all data）	76.98	64.89	81.29	68.35	69.67	72.24
BERT-base INIT	70.23	70.69	80.40	68.42	67.23	71.39
Pre-train then Fine-tune	72.73	72.03	71.05	70.99	69.12	71.18
Jointly Train	73.20	67.07	68.86	67.65	65.86	68.53
MDAIL	**79.42**	**72.78**	**82.96**	**72.81**	**74.90**	**76.57**

从表 2 中可以看出：

1） 与当前表现最好的模型相比，我们所提出的模型在五个领域上分别提高了 6.22%、0.7%、11.91%、1.73% 和 5.78%。证明了该模型有效且适用于多个不同领域。

2） 我们分别使用目标领域和所有语料混合进行了实验，从结果可以看出，不经过处理，直接使用其他领域的数据集对目标领域数据进行数据扩充，有时可以取得很大的提升，比如在音乐领域，但有时会产生负迁移效果，比如对于科学领域，由于分布不同，多个领域混合，可能效果反而会下降。而本文方法有效地缓解了这一问题。

2. 5 消融实验

为了验证每种结构的有效性，我们进行了消融实验。将去掉源领域预训练和目标领域预训练分别表示为 w/o 源领域预训练和 w/o 目标领域预训练。同样，将去掉领域层级贡献度加权和样本层级贡献度加权分别表示为 w/o 领域贡献度和 w/o 样本贡献度。实验结果如表 3 所示。

表 3 消融实验结果

Tab.3 Results of Ablation experiments

模型	政治	科学	音乐	文学	人工智能	平均值
MDAIL	79.42	72.78	82.96	72.81	74.90	-
w/o 源领域预训练	78.90	72.18	82.19	72.76	73.29	-0.71
w/o 目标领域预训练	78.19	71.95	81.07	71.05	72.24	-1.67
w/o 领域贡献度	77.60	72.43	82.15	72.32	72.24	-1.67
w/o 样本贡献度	78.10	71.58	81.72	70.13	70.97	-2.07

从消融实验的结果可以看出：

1） 分别去掉领域自适应预训练和两种层级的贡献度加权后，模型性能出现了不同程度的下降，因此通过消融实验的结果验证了两阶段领域自适应预训练和两种层级贡献度加权都是十分必要的。

2） 从两阶段领域自适应的消融实验中，可以看出，去掉目标领域预训练后性能下降得更多，说明目标领域预训练对于模型效果的提升更多，但源领域预训练这一步骤同样是必要的。

3） 从两种层级贡献度的消融实验中可以看出，领域贡献度和样本贡献度都大大提升了命名实体识别的性能，去掉样本层级贡献度后模型性能下降得更多，说明样本层级贡献度对模型更重要一些。

3 相关工作

由于越来越多语料资源匮乏的新兴领域存在文本分析及构建领域知识图谱的需求，针对领域自适应的命名实体识别研究逐渐成为重点内容。这项研究的基本思路是利用源领域来提升目标领域的模型效果。对此，Mou 等人提出了 INIT 方法，首先使用源领域语料训练模型，用训练好的参数启动目标模型，并使用目标领域语料继续训练[13]。Lee 等人提出联合训练源领域模型和目标域模型，共享参数[14]。Yang 等人提出只对部分层的参数进行共享，同时将 CRF 层分别分配给源领域和目标域[15]。但当前的大多数研究中都保持了单一源领域的设置，并未考虑全部能应用的标注数据集。Wang 等人探索了多源领域自适应，他们参照了多任务训练的思路，将不同的领域考虑为不同的任务，为每个领域构建不同的线性层和 CRF 层，进行共同训练[16]。这种方法的问题是当领域过多时，会导致模型的复杂度大幅增加，增加了训练的成本。

在分类任务中，针对多源跨领域方法的研究，已经取得了一定的进展，Li 等人针对情感分类问题提出了特征级和分类器级两种融合方法[17]。Wu 等人在分类任务中将分类器分解为公有部分和领域特有部分进行训练[18]。Chen 等人基于先前的工作提出 Mans 模型结构，利用领域鉴别器来提升领域的识别能力[19]。与 NER 任务不同，分类任务的数据标签在不同领域中也是相同的，而不同领域中的 NER 标签可能存在很大的差异。针对这一问题，Wang 等人将标签类别统一为通用的人名、地名、组织机构名，但这样做的缺陷是丧失了可能是对特定领域命名实体识别最重要的领域名词识别。

前述可知，当前跨领域命名实体识别任务存在的主要问题是 1）采用单一源领域，未考虑可用的其他领域数据对目标领域模型存在的促进作用。2）在多源跨领域 NER 中，没有考虑到领域特定词的识别。针对以上问题，我们将 NER 标签统一为人名、地名、组织机构名、领域名词、其他实体五种类别，提出了融合多源领域贡献度加权的自适应命名实体识别模型，通过领域层级和样本层级的贡献度参数评估样本对目标领域的贡献度。此外还引入了对抗训练，提升模型效果。

4 结论

本文提出一种融合多源领域贡献度加权的自适应命名实体识别模型，并提出了一种两阶段领域

自适应预训练方法，两阶段领域自适应预训练解决了目标领域的大规模数据依赖问题，以较小的资源成本实现了同样的领域自适应效果。在融合多源领域贡献度加权的自适应模型中，通过多个源领域来提升目标领域的模型效果，同时通过计算领域层级和样本层级的贡献度参数，并加权到命名实体识别模型中，来更一步地提升领域自适应效果。本文模型在多个目标领域的实验中都超越了当前性能最好的方法（SOTA），并可应用于低资源领域下的上游任务中来提升性能。在未来的工作中，我们将讨论如何通过领域选择挑选出最适应目标领域的多个源领域，希望能够通过领域选择，进一步提升领域适应效果。

参考文献：

[1] Grishman R .Information Extraction[J]. Intelligent Systems IEEE, 1999, 30(5).

[2] Woods W A . Semantics and Quantification in Natural Language Question Answering[J]. Advances in Computers, 1978, 17:1-87.

[3] Bahdanau D , Cho K , Bengio Y . Neural Machine Translation by Jointly Learning to Align and Translate[J]. Computer Science, 2014.

[4] Yadav V , Bethard S . A Survey on Recent Advances in Named Entity Recognition from Deep Learning models. 2019.

[5] H Daumé Iii. Frustratingly Easy Domain Adaptation[J]. ACL, 2009.

[6] Peng N , Dredze M . Multi-task Domain Adaptation for Sequence Tagging[J]. 2016.

[7] Jia C, Zhang Y. Multi-Cell Compositional LSTM for NER Domain Adaptation[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 5906-5917.

[8] Liu Z, Xu Y, Yu T, et al. CrossNER: Evaluating Cross-Domain Named Entity Recognition[J]. arXiv preprint arXiv:2012.04373, 2020.

[9] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019: 4171-4186.

[10] Aharoni R, Goldberg Y. Unsupervised Domain Clusters in Pretrained Language Models[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 7747-7763.

[11] Ganin Y, Lempitsky V. Unsupervised domain adaptation by backpropagation[C]//International conference on machine learning. PMLR, 2015: 1180-1189.

[12] Sang E T K, De Meulder F. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition[C]//Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003. 2003: 142-147.

[13] Mou L, Meng Z, Yan R, et al. How Transferable are Neural Networks in NLP Applications?[C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. 2016: 479-489.

[14] Lee J Y, Dernoncourt F, Szolovits P. Transfer Learning for Named-Entity Recognition with Neural Networks[C]//Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). 2018.

[15] Yang Z, Salakhutdinov R, Cohen W W. TRANSFER LEARNING FOR SEQUENCE TAGGING WITH HIERARCHICAL RECURRENT NETWORKS[J].

[16] Wang J, Kulkarni M, Preoţiuc-Pietro D. Multi-domain named entity recognition with genre-aware and agnostic inference[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 8476-8488.

[17] Li S, Zong C. Multi-domain sentiment classification[C]//Proceedings of ACL-08: HLT, Short Papers. 2008: 257-260.

[18] Wu F, Huang Y. Collaborative multi-domain sentiment classification[C]//2015 IEEE International Conference on Data Mining. IEEE, 2015: 459-468.

[19] Chen X, Cardie C. Multinomial Adversarial Networks for Multi-Domain Text Classification[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). 2018: 1226-1240.

A Multi-Source Domain Adaptation Approach in Named Entity Recognition

LI Jiarui, LIU Jian, CHEN Yufeng*, XU Jinan, ZHANG Yujie

(School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China)

Abstract: Domain adaptation is a general way to solve the problem of lacking training data, which can be applied to all kinds of Natural Language Processing tasks. At present, the research of domain adaptation for Named Entity Recognition (NER) usually follows the setting adapting from single source domain to single target domain. Though it can achieve certain results for the target domain which is close to the source domain, it has great limitations for the target domain which is not highly related to the source domain. To solve this problem, a multi-source domain adaptation named entity recognition model based on importance weighting is proposed, which promotes the performance of the target domain through the knowledge of multiple domains, models the importance of the target domain according to different domains and their samples, and weights to the named entity recognition model, so as to achieve better model adaptability. Experiments on several target domains show that the performance of the model is better than the state-of-the-art method, which verifies the effectiveness of the model.

Keywords: Named Entity Recognition; domain adaptation; importance weighting; multi-source