doi:10.3772/j.issn.2095-915x.2022.03.002

基于参数迁移的领域命名实体识别方法

开放科学（资源服务）标识码（OSID）

孙新[1,2] 任翔渝[1] 郑洪超[1] 杨凯歌[1]

1. 北京理工大学计算机学院 北京 100081；
2. 富媒体数字出版内容组织与知识服务重点实验室 北京 100038

摘要：[目的 / 意义] 命名实体识别是自然语言处理领域中的基础任务，基于深度学习的方法在通用领域的命名实体中取得了显著成果，但在特定领域识别效果不佳。为了解决工业信息化领域标注数据不足，数据特征差异较大、模型难以扩展的问题，首先提出了一种基于 Transformer 的有限区间命名实体识别模型。[方法 / 过程] 采用预训练模型对文本进行分布式表示，然后利用基于有限区间的标注方法对输入序列进行标注，解决传统标注法在训练过程中可能导致的序列标注不一致的问题。在此基础上，引入迁移学习策略，采用参数共享的方式，将通用领域的命名实体识别模型迁移到工业信息化领域，并在工业信息化领域数据集上进行微调，最终获得在工业信息化领域上表现良好的模型。[结果 / 结论] 实验结果表明，本文提出的有限区间命名实体识别模型在工业信息化领域数据集上的准确率较基线模型提高了 8.7%，基于参数迁移的领域命名实体识别方法在人民日报语料和工业信息化领域数据集上的准确率和综合指标 F 值相较未使用迁移学习的模型分别提高了 3.1% 和 1.1%，证明了迁移策略的有效性。

关键词：命名实体识别；深度学习；迁移学习；预训练语言模型

中图分类号：G35；TP391

Domain Named Entity Recognition Method Based on Parameter Transfer Learning

SUN Xin[1,2] REN XiangYu[1] ZHENG Hongchao[1] YANG Kaige[1]

1. School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China;
2. The Key Laboratory of Rich-Media Knowledge Organization and Service of Digital Publishing Content, Beijing 100036, China

Abstract: [Objective/Significance] Named entity recognition is a fundamental task in natural language processing, and deep learning-based methods have achieved remarkable results in general domains, but not in specific domains. Aiming at the problems of insufficient labeling samples, quite differences in data features and difficulty in model expansion, this paper

基金项目 富媒体数字出版内容组织与知识服务重点实验室开放基金项目“基于模糊粗糙集理论的远程监督关系抽取研究”（ZD2021-11/06）。

作者简介 孙新（1975-），博士，副教授，研究方向为自然语言处理、人工智能，E-mail：sunxin@bit.edu.cn；任翔渝（1998-），硕士研究生，研究方向为自然语言处理、机器学习；郑洪超（1999-），硕士研究生，研究方向为自然语言处理、对话系统。

引用格式 孙新 , 任翔渝 , 郑洪超 , 等 . 基于参数迁移的领域命名实体识别方法 [J]. 情报工程 , 2022, 8(3): 13-27.

doi:10.3772/j.issn.2095-915x.2022.03.002

introduces a limited span-based transformer classifier for named entity recognition model (Span-based Transformer Classifier for Named Entity Recognition, STCNER). [Methods/Process] The model takes advantage of the features extraction of Encoder in Transformer and combines with the limited span-based labeling method, which solves the problem of the sequence labeling inconsistency caused by traditional labeling method in the training process. On this basis, then introduce the transfer learning strategy which adopt the parameter sharing method to transfer the named entity recognition model in general domains to the specific domains. After fine-tuning it on the domain-specific dataset, the model performs well in specific domain. [Results/Conclusions] The experimental results show that the accuracy of STCNER model is 8.7% higher than the baseline model on the dataset in the industrial informatization field. Compared with the model without transfer learning, the accuracy and F-scores are improved by 3.1% and 1.1% respectively on the corpus of People's Daily and the data set in the industrial informatization field, which proves the effectiveness of the transfer strategy.

Keywords: Named entity recognition; deep learning; transfer learning; pre-trained language model

引言

随着互联网、移动终端的广泛使用，信息从单一的媒体逐步转变为文本、图像、音频、视频等多媒体的形式。信息的载体呈现富媒体化，信息的获取和传播也以更多样、灵活、开放的方式进行，这使得传统提取信息的方式面临着巨大的挑战。利用人工智能技术对富媒体数据进行处理、分析，挖掘出富媒体数据间深层次的关系，实现对信息获取的自动化、智能化具有重要的应用价值。命名实体识别（Named Entity Recognition，NER）是自然语言处理（Natural Language Processing，NLP）领域中的基础任务，旨在从文本中识别具有特定意义的实体。命名实体识别在信息抽取、问答系统、机器翻译等任务中起着重要的作用。

命名实体识别技术经历了基于规则的方法、基于机器学习的方法到深度学习方法的演变。在早期的基于规则和基于机器学习的命名实体识别方法都取得了较好的识别效果，但这些方法都耗费人力且具有较高的局限性。深度学习的方法可以自动捕获输入句子的特征，具有较强的泛化能力，受到学者们的广泛关注，主要包括基于双向长短期记忆网络(Bidirectional Long Short-Term Memory, BiLSTM)[1]，基于条件随机场 (Conditional Random Field, CRF)[2]，以及基于 BiLSTM-CRF 模型的命名实体识别 [3] 等方法。2017 年 Ashish Vaswani 等 [4] 提出 Transformer 模型，该模型使用基于注意力机制的方式，被广泛应用在众多自然语言处理任务中。Yan 等 [5] 针对命名实体识别任务对 Transformer 模型进行了改进，取得了较为理想的识别效果。此外，也有研究工作尝试通过预训练语言模型来提高命名实体识别的性能 [6, 7]。

无论是传统机器学习还是深度学习的方法，都依赖大量标注数据来训练模型，而现有研究对少量标注数据的特定领域命名实体识别问题探讨较少 [8]。标注样本不足是特定领域命名实体识别任务的最大问题，同时，由于特定领域命名实体识别方法通常是根据各自领域的特点

设计的，还面临着数据特征差异较大、模型难以扩展等问题。迁移学习（Transfer Learning）为这一问题的解决提供了可行思路。迁移学习利用领域相似性，在领域之间进行数据共享和模型共建，其目的是迁移已有的知识，用以解决目标领域中仅有少量甚至没有标注数据的学习问题[9]，在一定程度上解决标注语料不足或缺失的问题。

本文研究适用于工业信息化领域的命名实体识别方法，首先针对传统命名实体识别模型在工业信息化领域数据集上识别效果不佳的问题，提出基于 Transformer 的有限区间命名实体识别模型（Span-based Transformer Classifier for Named Entity Recognition，STCNER），使用预训练语言模型对文本进行分布式表示，采用基于有限区间的标注方法对输入序列进行标注，确定命名实体的边界并判断其类别，解决传统标注法在训练过程中可能导致的序列标注不一致的问题，提高领域命名实体识别的准确性。进一步，针对领域标注样本不足的问题，引入迁移学习策略，提出基于参数迁移的跨领域命名实体识别方法（Trans-STCNER）。采用参数共享的方式，以新闻领域的人民日报数据集和工业信息化领域的数据集分别作为源域和目标域，将通用领域训练得到的模型迁移到工业信息化领域，并在工业信息化领域数据集上进行微调。通过人民日报语料和工业信息化领域数据集上的对比实验验证了迁移策略的有效性。

1 相关工作

命名实体识别任务能够从非结构文本中提取出预先定义的实体，通常包括人名、地名、机构名、时间、专有名词等。常用的方法有基于规则的方法、基于机器学习的方法，以及基于深度学习的方法。

基于规则的方法是早期的主流方法，它首先需要语言学家手工构造规则模板，或者借助机器自动地制定或者生成规则，然后再从文本中寻找与之匹配的字符串。这类方法移植性差，且人工构建规则的成本极高。基于机器学习的方法把实体识别看作序列标注问题，通过学习大量的标注语料，提取特征并训练模型，相对于基于规则的方法更为通用，并且移植性更好。基于机器学习的常用方法包括：条件随机场（Conditional Random Fields，CRF）、隐马尔可夫模型（Hidden Markov Model，HMM）、支持向量机（Support Vector Machine，SVM）等。其中条件随机场为命名实体识别提供了一个灵活的、全局最优的标注方法，它在对某个位置进行标注的时候考虑了标注标签之间的顺序，可以利用到已经标注的信息，在学术界得到了广泛认可。但是，基于传统机器学习的命名实体识别方法的特征仍然依靠人工定制。

基于深度学习的方法借助神经网络自动学习特征并训练序列标注模型，取得了卓越的效果。Peng 等[2]在实验中利用单词边界标记的信号进行分词，联合 LSTM 和 CRF 训练模型取得了较高的识别准确率。Peter 等[10]提出的深度学习模型 TagLM，以及融合注意力机制并结合条件随机场的 Attention-BiLSTM-CRF 模型[11]均在开放域命名实体识别任务中取得了较好的结果。

随着深度学习的发展，Google 团队在 2017 年 6 月提出了 Transformer 模型[4]，因其引入的

doi:10.3772/j.issn.2095-915x.2022.03.002

Self-Attention 机制、残差前馈网络和多头机制等结构，使其在语义特征提取、长距离特征捕获、高并行计算等方面表现突出。这些特性使得 Transformer 更适用于机器翻译任务，以及在语言模型任务中用来建模序列、作为预训练语言模型中的特征提取器 [12]。BERT（Bidirectional Encoder Representation Transformer）模型 [12] 将预训练步骤同深度双向 Transformer 相结合，克服了传统特征提取器的缺点，在 NLP 的许多问题中取得了出色的成绩。

深度学习的引入和发展使命名实体识别任务在通用领域上取得了良好效果，但在特定领域由于标注数据获取成本高或数量规模较小，方法表现不佳。迁移学习是利用已获取的知识对相关领域的不同问题进行求解的一种机器学习方法，其核心思想就是根据源域（source domain）和目标域（target domain）之间的相似性，以度量为准则增大两个领域之间的相似性，研究如何把源域的知识迁移到目标域上。迁移学习广泛应用诸如计算机视觉、情感分类、命名实体识别等任务中。

Yosinski 等 [13] 对深度神经网络的可迁移性进行了研究，研究结果表明，对于一个深度神经网络而言，随着网络层数的不断加深，网络对特定任务的依赖性就越强，而浅层网络一般只能学习到一些通用的大致的特征。基于该理论，学者们提出了许多深度迁移模型，核心思想是根据一个度量准则制定损失函数，并通过训练模型减小损失函数值来缩小两个领域的差异。例如 DDC（Deep Domain Confusion）模型 [14] 就在预训练好的 AlexNet 网络的特征层后加入了一个领域适配层来度量网络对源域和目标域的判别能力，经过对源域和目标域领域适配层的输出计算一个最大均值差异（Maximum Mean Discrepancy，MMD），通过最小化这个最大均值差异值来缩小两个域之间的分布差异。该方法思路简单，是迁移学习领域的经典方法。这些研究成果促进了迁移学习在多任务、多领域中的应用。

冯建周等 [15] 提出了一种基于迁移学习的细粒度实体分类方法，结合 BiLSTM 模型和注意力机制实现迁移学习。类似的，武惠等人 [16] 也提出了一种基于迁移学习和深度学习的 TrBiLSTM-CRF 模型。这些模型采用双向长短期神经网络对单词表示进行上下文编码，其编码能力有限且无法并行，没有充分融合上下文语义信息对文本特征进行提取。

面向特定领域的命名实体识别方法通常是根据各自领域的命名实体特点而设计的，难以扩展到其他领域，而且，现有方法没有很好地利用文本的语义表示，无法充分融合语义上下文。此外，标注数据不足依然是个突出的问题。因此，本文引入迁移学习的方法，充分利用迁移学习能够从大量标注数据领域的数据和模型中学习到一定的泛化的语言学知识的能力，以期在少量标注数据领域获得更好的模型。

2 基于 Transformer 的有限区间命名实体识别

针对传统命名实体识别模型在工业信息化领域识别效果不佳的问题，提出基于有限区间的 Transformer 命名实体识别模型（STCNER），模型结构如图 1 所示。

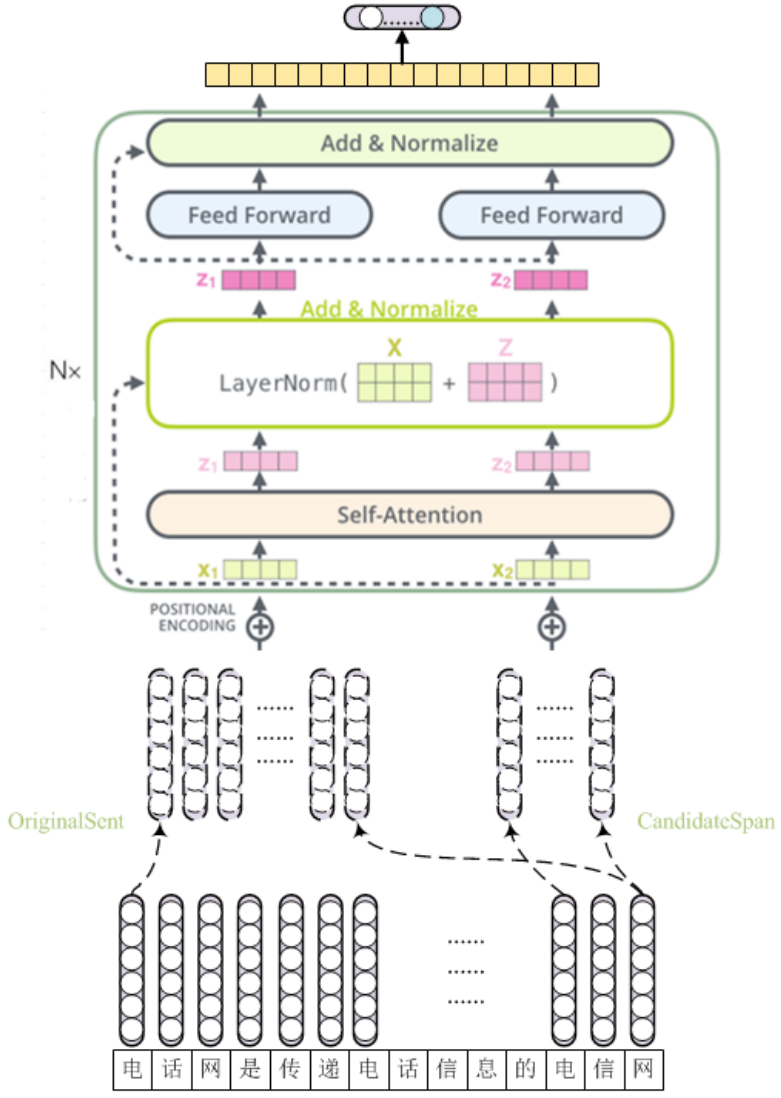

图 1　**STCNER** 模型结构

首先，在输入部分，采用 BERT 预训练模型对输入的文本进行字级别的动态向量表示（即同一个字在不同的上下文中会有不同的向量表示）。同时，考虑到 Attention 机制无法考虑字之间的相对位置关系，输入部分引入了位置编码，将词语的 Embedding 向量加上其位置向量作为最后的输出，这使得同一个字在句子的不同位置具有不同的向量表示。由于基于 Attention 机制的 Transformer 可以更加充分地融合上下文语义信息对文本特征进行提取，本文利用其中的 Encoder 部分作为模型的主体，对句子进行编码。为了解决序列标注过程中的不一致问题，模型中只使用 Transformer 的 Encoder 部分用于对输入进行语义抽取，Encoder 部分的输出再通过一个全连接层和 softmax 层，得到最后对应每个类别的概率分布。

在训练过程中采用一种基于有限区间的标注方法，将序列标注问题转化为分类问题，可

doi:10.3772/j.issn.2095-915x.2022.03.002

有效解决序列标注任务中序列标注不一致的问题。同时将基于 Attention 机制的编码器和有限区间的分类方法相结合，也可以克服传统的序列标注方法在对某个 token 进行标注时，无法考虑输入文本的全局语义信息的问题。

STCNER 模型由四个模块构成：词向量表示模块、基于有限区间的候选命名实体抽取模块、编码器模块和分类模块。

（1）词向量表示模块

词向量表示模块对句子进行分布式表示，生成连续且稠密的词向量。本文采用 BERT 模型进行词向量表示，使其既可以捕获实体的上下文语义，又可以分别对词语和字符级别的特征进行抽取，从而最大程度上得到贴近实体含义的词向量。

（2）基于有限区间的候选命名实体抽取模块

采用一种基于有限区间的标注方法对句子向量进行实体标注与抽取，确定命名实体的边界并判断其类别，将命名实体识别这一序列标注问题转化为分类问题，是对下游分类模块的预处理。

（3）编码器模块

编码器模块的输入为经过 BERT 训练后得到的词向量与候选命名实体的词向量的拼接。本文采用的编码器模块由四个结构相同的 Transformer Encoder 网络格串行连接而成，用于对输入进行语义抽取。

（4）分类模块

分类模块的输入为编码器模块的输出，使用全连接层和 softmax 对候选命名实体进行分类，完成命名实体识别任务。

2.1 命名实体抽取模块

首先，定义一组标签，$Lable=\{E_1, E_2, \ldots, E_n, O\}$，其中 E_i 表示当前知识库中的实体类别，n 为预测标签序列长度。命名实体抽取模块的任务是对于文本中所有规定区间范围内（记为 L）的序列预测它们的标签，即对于给定包含 T 个字的句子 $X=\{w_1, w_2, \ldots, w_T\}$，其中 w_i 表示第 i 个字，$i=1, 2, \ldots, n$，预测一组带标记的输出 Y：

$$Y=\{(i,j,l)\,|\,1\le i\le j\le T; j-i+1\le L; l\in Lables\}$$

其中 i 和 j 表示这个句子中字的位置，l 表示该区间的标签，L 表示区间的最大长度，Y 表示输出。举例如图 2 所示，其中 N 表示通信网。

Input	电 1 话 2 网 3 是 4 传 5 递 6 电 7 话 8 信 9 息 10 的 11 电 12 信 13 网 14
Output	(1,2,O),(2,3,O),(3,4,O),...,(13,14,O), (1,3,N),(2,4,O),(3,5,O),...,(12,14,N), (1,4,O),(2,5,O),(3,6,O),...,(11,14,O)

图 2 有限区间的序列标注

命名实体抽取模块的输入是词向量表示模块的输出，即一个句子的向量表示，然后直接从原始向量表示中根据位置截取候选区间，对候选区间所表示的实体进行类别标注与抽取。

该方法的优点是提取的搜索空间可以随句子长度线性缩小，远远小于标记法。此外，由于概率是由区间大小来决定的，因此该模型能够在预测之前考虑所有目标词，从而避免了标注不一致的问题。

2.2 编码器模块

编码模块采用的是 Transformer 中的堆叠 Encoder 层结构。

（1）位置编码（Positional Encoding，PE）

在数据预处理时将每个词向量与一个位置

编码求和，加入相对位置信息，该信息可以确定每个单词的位置，或者是序列中不同单词之间的距离。相对位置编码计算如下所示：

$$PE_{(pos,2i)} = \sin(pos / 10000^{2i/d_{model}}) \quad (1)$$

$$PE_{(pos,2i+1)} = \cos(pos / 10000^{2i/d_{model}}) \quad (2)$$

其中 pos 是当前字在句子中的位置标号，i 是指位置向量中每个值的分量。

（2）Self-attention

常规的 Attention 可以被解释为一个查询序列 Query 和一个被查询的序列，用 Key-Value 键值对表示。通过 Query 序列与 Key 计算相似度得到 Value 对应的权重，再用权重和对应的 Value 计算得到新的 Value 值。将 Query、Key、Value 分别简写为 Q、K、V，则有：

$$attention\_output = Attention(Q,K,V) \quad (3)$$

（3）Feed-forward networks

第二个子层是全连接层，该层对多头注意力的每个头的输出进行进一步的前馈网络处理，如公式（4）所示，其中 W_1，W_2，b_1，b_2 是可学习的参数。

$$FFN(x) = \max(0, xW_1 + b_1)W_2 + b_2 \quad (4)$$

2.3 分类模块

分类模块的输入为句子与候选命名实体向量经过编码器输出的特征向量表示 h，输出为候选实体对应各个分类的概率值。分类模块首先将 h 输入全连接层，然后通过 softmax 将输出候选实体对对应到各个分类的概率值。计算如公式（5）所示：

$$o=softmax(Wh+b) \quad (5)$$

其中 o 是预测分类向量，W 是权重矩阵，b 是偏置向量。

3 基于共享参数的迁移学习策略

STCNER 模型使用预训练模型能够提升工业信息化领域命名实体识别任务的效果，同时，通过有限区间的标注方法能够解决针对传统 BIO 标注方式导致的序列标注不一致的问题。但是，工业信息化领域标注数据不足依然是个突出的问题。

为解决这一问题，在 STCNER 模型的基础上引入迁移学习策略，提出一种深度迁移的算法（Trans-STCNER），进一步提高工业信息化领域命名实体识别的准确率。Trans-STCNER 采用基于共享参数的迁移策略，该策略能够充分地将在拥有大量标记样本的领域中训练的模型应用于标注数据较少的特定领域，同时在最大程度上降低特定领域模型训练时对标注数据数量的依赖，使得工业信息化领域的命名实体识别模型能在较少数据量的情况下实现收敛，达到较好的训练效果。

基于参数迁移的领域命名实体识别模型 Trans-STCNER 以 STCNER 模型作为基础模型进行深度迁移，源域和目标域共享模型的权重，将全连接层作为领域适配层。通过适配层的输出计算最大平均差异（Maximum Mean Discrepancy，MMD）距离表示出源域和目标域之间的差异，误差由分类误差和最大平均差异距离误差构成。训练的过程即是最小化源域和目标域差距的过程，自动学习一个联合训练的表示来优化分类和域的自适应性，从而实现更好的迁移效果。

这里，采用具有大量标注信息的人民日报语料库作为源域，用具有少量标注信息的工业

doi:10.3772/j.issn.2095-915x.2022.03.002

信息化领域数据作为目标域进行迁移。具体方法是：首先利用新闻领域的命名实体识别数据对模型进行预训练；其次将训练得到的模型参数融入到训练参数相同的工业信息化领域的命名实体识别模型中，并在其中加入位置信息；最后利用当前工业信息化领域的命名实体识别数据对模型进行微调，实现对工业信息化领域的命名实体识别性能的提升。基于参数迁移的命名实体识别方法 Trans-STCNER 的流程示意如图 3 所示。

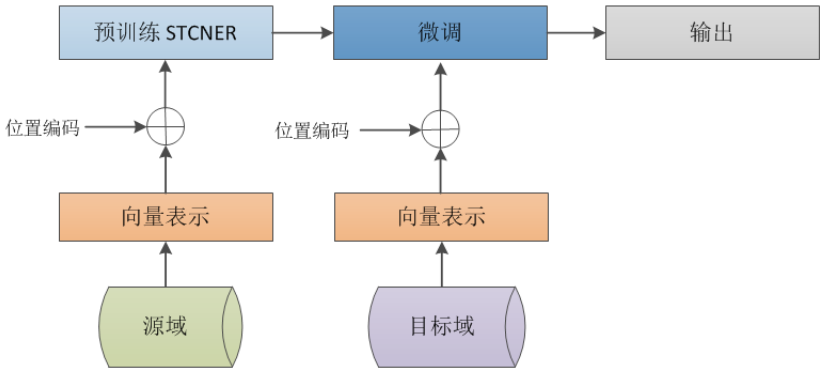

图 3　基于参数迁移的命名实体识别方法 **Trans-STCNER** 的流程示意图

网络包括两个流向，一个是新闻领域的命名实体识别数据，另一个是工业信息化领域的命名实体识别数据，两个流向的神经网络共享权值。STCNER 模型中的全连接层在这里作为领域适配层、不共享权值，并通过这两个适配层的输出计算出一个域损失值，这里通过计算源域和目标域之间的最大均值差异 MMD 距离作为域损失值。在训练模型的过程中，通过最小化 MMD 距离来缩小源域和目标域之间的差异，以实现更好的迁移效果。

（1）模型输入

模型每个流向的输入由词向量和位置向量相加而成。词向量采用 BERT 预训练模型生成，位置向量则采用 Transformer 模型中所用的正余弦位置编码。

（2）最大均值差异 MMD

最大均值差异（Maximum Mean Discrepancy，MMD）是一种核学习方法，它度量的是在再生希尔伯特空间 (Reproducing Kernel Hilbert Space，RKHS)[17] 中两个分布的距离。MMD 通常用于度量两个分布之间的相似度，是迁移学习中使用频率最高的度量方法，两个随机变量的 MMD 平方距离的计算如公式（6）所示：

$$MMD^2(X,Y)=\left\|\sum_{i=1}^{n_1}\phi(x_i)-\sum_{j=1}^{n_2}\phi(y_j)\right\|^2 \quad (6)$$

其中 $\phi(\cdot)$ 是映射函数，用于把原变量映射到 RKHS 中。希尔伯特空间是对于函数的内积完备的，而再生核希尔伯特空间是具有再生性的希尔伯特空间。将平方展开后，RKHS 中的内积就可以转换成核函数，所以最终 MMD 可以直接通过核函数进行计算。简单来说 MMD 就是求两堆数据在 RKHS 中的均值的距离。

基于参数迁移的跨领域命名实体识别模型使用 MMD 的平方计算新闻领域和工业信息化领域的数据在 RKHS 的均值差异，当作是两部分数据的差异。计算如公式（7）所示：

$$MMD^2(X_S,X_T)=\left\|\frac{1}{|X_S|}\sum_{x_s\in X_S}\phi(x_s)-\frac{1}{|X_T|}\sum_{x_t\in X_T}\phi(x_t)\right\|^2 \quad (7)$$

其中 X_S 表示源域数据，X_T 表示目标域数据，$|X|$ 表示 X 域的数据量。

实际中，在一个 batch 的训练里采用的计算如公式（8）所示：

$$MMD^2[F,X,Y]=\frac{1}{m(m-1)}\sum_{i\neq j}^{m}k(x_i,x_j)+\frac{1}{n(n-1)}\sum_{i\neq j}^{n}k(y_i,y_j)-\frac{2}{mn}\sum_{i,j=1}^{m,n}k(x_i,y_j) \quad (8)$$

其中 X 和 Y 分别表示源域和目标域，m 和 n 分别表示源域和目标域数据集的大小，我们选择的核函数 k 为高斯核函数，以计算两个点在高维空间中的距离。

（3）网络的损失函数

经过对深度网络迁移的研究，使用领域适配层来实现源域数据和目标域数据自适应的方法，使得数据分布更加接近，从而进一步提高网络的迁移效果。

模型中采用的损失函数如公式（9）所示，网络的损失由分类误差和MMD距离误差构成。

$$\ell=\ell_C(D_s,y_s)+\lambda\ell_A(D_s,D_t) \quad (9)$$

其中，ℓ 表示网络的最终损失值，$\ell_C(D_s, y_s)$ 表示网络在源域上的常规分类损失（这与普通的深度网络完全一致），它来度量预测值和真实值的差异，$\ell_C(D_s, D_t)$ 表示网络的自适应损失，即 MMD 距离。公式（9）中的 λ 是权衡两部分的权重参数。

4 实验结果及分析

为了验证本文提出的 Trans-STCNER 模型的性能，设计以下两组实验：（1）不同命名实体识别方法的对比实验，本文首先将 STCNER 模型与其他模型进行对比，分别从模型对比、Encoder 层数和词向量表示方法层面验证 STC-NER 模型的命名实体识别效果；（2）迁移学习的性能试验，研究不同因素对迁移学习效果的影响，以验证迁移学习的有效性。

4.1 实验数据及设置

实验采用数据的是 1998 年人民日报命名实体识别数据集，数据集中的训练集和测试集划分以及语料的识别种类、标签设置和格式信息如表 1 所示。

表 1 人民日报命名实体识别数据集

类型	训练语料	测试语料
总数	46364个句子	4365个句子
识别种类	人名，地名，机构名	人名，地名，机构名
标签设置	PER，LOC，ORG	PER，LOC，ORG
格式	IOB，首B，中I，非O	IOB，首B，中I，非O

此外，针对工业信息化领域的命名实体识别任务，本小节选择工业信息化语料作为实验数据，语料中包含了三个子领域的实体：计算机软件，算法和通信网。数据集中训练集和测试集的划分以及语料的识别种类、标签设置和格式信息如表 2 所示。

计算机软件、算法和通信网三个子领域的语料具体例子如图 4 所示。

实验采用正确率（Precision）、召回率（Recall）、和 F 值（F-measure）三个评价指标对命名实体识别结果进行性能分析。计算如下所示：

doi:10.3772/j.issn.2095-915x.2022.03.002

$$正确率（P）=\frac{系统正确识别的实体个数}{系统识别的实体个数}\times 100\% \quad （10）$$

$$召回率（R）=\frac{系统正确识别的实体个数}{语料库中的实体个数}\times 100\% \quad （11）$$

$$F=\frac{P*R(\beta^2+1)}{R+\beta^2*P} \quad （12）$$

其中 β 表示权重，当 β=1 时表示正确率与召回率二者同样重要，当 β>1 时表示正确率重于召回率，反之召回率重于正确率。这里取 β=1，因此，F 值计算如公式（13）所示：

$$F=\frac{2\times P\times R}{R+P}\times 100\% \quad （13）$$

表 2　工业信息化命名实体识别数据集

类型	计算机软件	算法	通信网
语料总数	1604	1295	844
训练总数	1000	810	527
测试总数	604	485	317
识别种类	系统软件，应用软件	经典算法、安全算法、机器学习算法	有线网、无线网
标签设置	SYS，APP	CLA、SEC、MAC	WED、WLE
格式	IOB，首B，中I，非O	IOB，首B，中I，非O	IOB，首B，中I，非O

计算机软件

UNIX	是	上	世	纪	70	年	代	问	世	的	分	时	系	统
B-SYS	O	O	O	O	O	O	O	O	O	O	B-SYS	I-SYS	I-SYS	I-SYS

算法

决	策	树	是	常	见	的	监	督	学	习	算	法
B-MAC	I-MAC	I-MAC	O	O	O	O	B-MAC	I-MAC	I-MAC	I-MAC	I-MAC	I-MAC

通信网

有	限	电	视	网	络	具	有	高	带	宽	的	特	点
B-WED	I-WED	I-WED	I-WED	I-WED	I-WED	O	O	O	O	O	O	O	O

图 4　工业信息化领域语料示例

4.2　STCNER模型的实验结果

为了验证本文提出的命名实体识别模型 STCNER 在工业信息化领域数据集上的效果，本小节设计了三组实验，分别从模型对比、Encoder 层数和词向量表示方法层面验证 STCNER 模型的命名实体识别效果。模型的参数设置如表 3 所示。

（1）不同模型在工业信息化数据集上的对比实验

为了证明 STCNER 模型的有效性，在工业信息化数据集上与以下经典模型进行比较，实验结果如表 4 所示。

表 3　参数设置

参数类型	参数值
Embedding dimension	100
Sample_maxlen	20
Zero_padding	True
Learning rate	0.0003
Batch size	16
Epoch	100
Encoder blocks	6
Dropout rate	0.3
smoothing	0.1
Warmup step	4000

表 4 STCNER 模型与传统模型的实验结果对比

模型	准确率（%）	召回率（%）	F1值（%）
LSTM	19.8	33.2	24.8
BiLSTM	50.7	67.4	57.9
LSTM+CRF	42.9	56	48.6
BiLSTM+CRF	61.4	76.9	68.3
Random Initialization	51.3	63.1	56.6
STCNER- Word2vec	54.6	68.9	60.9
STCNER -ELMO	67.8	70.4	69.1
STCNER-BERT	70.1	78.5	74.1

由表 4 给出的实验结果可以看出，基于 BiLSTM 的模型要比基于 LSTM 的模型效果好，在二者的基础上加入 CRF 可有效提高模型的性能，而 STCNER 模型的效果最优。

由于 LSTM 只能考虑到文本的单向信息，对文本的语义表征是不够的，所以实验效果不如 BiLSTM 模型。基于 LSTM 和 BiLSTM 的命名实体识别模型在预测时可能存在标注不一致的问题，而 CRF 模型在序列标注时可以考虑前面的标注信息，以解决标注不一致的问题。实验结果表明，在 LSTM 和 BiLSTM 模型中引入 CRF 进行序列标注可以进一步提升模型效果。

STCNER 模型则将序列标注问题转化为基于有限区间的分类问题，这不仅解决了标注不一致的问题，也进一步缩小了模型输出的搜索空间，同时在做标注时可以充分考虑全局信息，可以进一步提高精度，实验效果也证明了这一点，STCNER 模型的效果要优于已有的 LSTM、BiLSTM、LSTM+CRF 和 BiLSTM+CRF 模型。

为了研究不同的词向量表示方法对实验结果的影响，实验还选择了随机初始化、Word2vec[18]、ELMO[19] 和 BERT 这四种现在主流的词向量表示方法来进行对比实验，其中随机初始化是完全上下文无关的静态词向量表示方法，Word2vec 是考虑了词语的上下文信息的静态词向量表示方法。ELMO 和 BERT 都是可以充分表征词语的上下文语义和句法信息的动态词向量表示方法，区别在于 ELMO 只能考虑到词语的单向语义信息，而 BERT 能够充分融合词语的上下文两个方向上的语义信息。

实验结果如表 4 所示。实验结果证明了在 STCNER 模型中采用 BERT 预训练模型进行向量表示的优势，采用了 BERT 的 STCNER 模型在工业信息化领域数据集上的精确率、召回率和 F 值都要高于采用其他三种词向量表示的 STCNER 模型。

采用 BERT 进行词向量要优于其他几种方法的原因在于：

1）BERT 能够充分的融合词语的上下文语义信息，而 ELMO 只能融合词语的单向语义信息，再将两个方向上的语义信息进行简单的拼接来得到词语的上下文信息的近似表示，难以

doi:10.3772/j.issn.2095-915x.2022.03.002

对词语的上下文信息进行充分表示。

2）BERT 能够在考虑输入文本的上下文信息的情况下，对其中的每个词语的语义信息进行动态表示，这一点是 BERT 远优于随机初始化和 Word2vec 的特点，能够解决一词多义的现象。

（2）Encoder 层数对模型效果的影响

工业信息化领域的语料规模较小，为了避免过拟合问题的发生而影响模型的效果，对比 STCNER 中采用不同的 Encoder 层数的模型效果，以确定最适合任务的 Encoder 层数。同时，为了更充分地对比模型参数量在不同规模的语料上训练效果的差异，在大规模的人民日报语料和小规模的工业信息化语料上分别采用 1~6 个不同的 Encoder 层进行训练，通过 F 值的差异来确定最优的 Encoder 层数。实验结果如表 5 所示。

表 5 不同 **Encoder** 层数在不同数据集上的 **F1** 值

Encoder层数	人民日报（%）	工业信息化（%）
1	46.0	62.8
2	51.2	63.5
3	53.5	71.5
4	65.7	74.1
5	71.5	70.2
6	72.2	64.3

从表 5 中可以看出，小规模的工业信息化语料在 Encoder 层数为 4 时 F 值最高，更多 Encoder 层数意味着更多的参数量、使得模型发生了过拟合。而大规模的人民日报语料的 F 值则随着 Encoder 层数的增加一直在增加，说明参数量的增加还未产生过拟合的问题。通过这两个不同规模语料的对比，可以发现不同规模的数据集对模型参数量的要求是不同的。小规模的语料必然要求模型的结构和参数量都不能太复杂，否则很容易产生过拟合的问题。而大规模语料则要求模型具有足够的结构复杂性和参数量，这样才能充分地提取数据集中的特征，以提高模型性能。

（3）错误识别实例分析

进一步分析错误识别的实例，一个错误识别的实例如图 5 所示。对于句子“非对称数字用户环路是一种新的数据传输方式”，其中“非对称数字用户环路”属于通信网中的“有线网”实体。由于候选区间的长度设置问题，未能有效识别出这个长度为 9 的实体，导致整个句子中每个词的标签都是“O”。分析其原因，STCNER 模型做预测时，为了减小候选区间的搜索空间，模型需要预先设定候选区间的长度范围。这个范围无法自动获得，使得区间长度的确定需要依赖经验公式，因此可能会导致长度超过该范围的领域实体识别出现错误。

非	对	称	数	字	用	户	环	路	是	一	种	新	的	数	据	传	输	方	式
O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O

图 5 通信网实体错误识别实例

4.3 Trans-STCNER迁移学习策略的实验结果

为了验证本文提出的迁移学习策略在工业信息化领域命名实体识别任务中的效果，本节设计了三组实验，对比了是否引入迁移学习的STCNER模型的实验效果，以及不同的源域对模型迁移效果的影响。模型的参数设置如表6所示。

表 6 参数设置

参数类型	预训练参数值	微调参数值
Embedding dimension	256	256
Sample_maxlen	50	50
Zero_padding	True	True
Learning rate	0.003	0.0003
Batch size	32	16
Epoch	100	50
Encoder blocks	6	6
Dropout rate	0.5	0.5
smoothing	0.1	0.1
Warmup step	5000	4000

（1）迁移学习对工业信息化领域的命名实体识别的影响

为了验证Trans-STCNER中基于参数的迁移学习以及加入MMD距离的迁移学习对命名实体识别模型的有效性，首先将Trans-STCNER模型分为三个主要部分：STCNER模型、迁移学习和加入MMD距离的迁移学习。在STCNER模型的基础上，依次加入其余两个模块进行实验对比，实验结果如表7所示。

实验结果表明，在STCNER模型中加入基于参数的迁移学习方法，虽然在召回率上有所降低，但在精确率和综合指标F值上要优于STCNER模型，这也证明了迁移学习策略的有效性。进一步，在基于参数迁移的STCNER模型中加入MMD距离，以MMD距离作为领域差异的度量标准，通过减小MMD值使得源域和目标域的分布更加接近，实验结果表明这种策略虽然在准确率和召回率上的表现都不是最优的，但F值是最高的，这也在一定程度上说明这种方法的有效性。

表 7 引入迁移学习和 MMD 距离的 STCNER 模型的实验对比

模型	准确率（%）	召回率（%）	F1值（%）
STCNER	70.1	78.5	74.1
STCNER+Transfer Learning	74.2	75.3	74.7
STCNER+Transfer Learning+MMD	73.2	77.4	75.2

这种策略之所以不能取得较大程度的性能提升，分析其原因可能有以下两方面：首先是源域数据集的规模较小，人民日报语料在迁移学习中作为源域数据集可能包含的数据量偏少，致使模型无法得到充分的预训练。其次是源域和目标域的分布情况不同，虽然在模型中引入MMD距离缩小了两个领域之间的分布差异，但由于只是在单层上引入了适配层，可能无法充分地缩小两个领域之间的分布差异。

（3）不同的源域数据对模型迁移效果的影响

为了研究不同的源域数据对迁移学习方法的影响，这里选取了MSRA、ResumeNER、WeiboNER和人民日报四种不同的命名实体识别语料作为源域数据集，它们具有不同的规模、不同的文本领域和不同的标注方案。数据集的统计信息如表8所示。

doi:10.3772/j.issn.2095-915x.2022.03.002

表 8 不同数据集的信息

数据集	样本总数	标签	格式
ResumeNER	4760	NAME\EDU\TITLE\CONT\LOC\ORG\RACE\PRO	BMEO
WeiboNER	1890	PER.NOM\PER.NAM\LOC.NOM\LOC.NAM\ORG.NOM\ORG.NAM\GPE.NAM	BIO
MSRA	46365	PER\LOC\ORG	BIO
人民日报	50729	PER，LOC，ORG	BIO

将以上四种不同数据集作为源域进行迁移的实验结果如表 9 所示。由实验结果可知，在小规模的数据集 ResumeNER 和 WeiboNER 上的表现要远低于在 MSRA 和人民日报语料上的表现，甚至要远低于不加入迁移学习策略的 STCNER 模型的表现，可能的原因有两点：一是数据集的规模较小，无法充分对模型进行预训练；二是数据集之间的数据分布差异较大，反而阻碍了模型在目标域上的训练过程。

表 9 采用不同源域的实验结果对比

数据集	准确率（%）	召回率（%）	F1值（%）
ResumeNER	53.5	61.3	57.1
WeiboNER	69.5	49.0	57.5
MSRA	70.3	76.2	73.1
人民日报	73.2	77.4	75.2

实验结果证明基于人民日报的迁移学习策略是表现最好，这是因为人民日报与其他数据集相比具有更大的数据规模，同时数据分布与目标数据集的分布差异最小。

5 结论

针对传统标注法在训练过程中可能导致的序列标注不一致的问题，本文提出了基于有限区间的命名实体识别模型 STCNER。首先采用基于有限区间的标注方法对文本的向量表示进行截取，抽取候选命名实体并标注其类别；其次以 Transformer 中的 Encoder 作为模型基本架构，利用其中的 Self-Attention 机制对文本与候选命名实体的特征进行充分抽取；最后通过全连接层输出候选命名实体的类别。在工业信息化领域数据集上的对比实验表明了 STCNER 模型的有效性。

针对工业信息化领域命名实体识别任务中标注样本不足、模型训练效果不佳的问题，本文在 STCNER 模型的基础上提出一种深度迁移的算法 Trans-STCNER，采用参数共享的方式，将在通用领域的命名实体识别数据集上预训练得到的模型迁移到工业信息化领域，并在领域命名实体识别数据集上进行微调，以获得在工业信息化领域数据集上表现良好的模型。在人民日报语料和工业信息化领域数据集上的对比实验表明了本文提出的迁移策略的有效性。

本文重点研究了领域命名实体识别标注数据不足，数据特征差异较大、模型难以扩展的问题，提出适用于工业信息化领域的命名实体识别方法，具有一定研究意义。综合实验结果来看，后续工作可以考虑把候选区间的长度作为学习参数，动态地调整长度范围，进一步提升实体识别效果。此外，实验验证了本文为提出的迁移策略对模型的效果有一定的提升，未

来随着迁移学习、无监督学习技术的发展，可以尝试深度迁移方法来达到更好地迁移效果。

参 考 文 献

[1] Panchendrarajan R, Amaresan A. Bidirectional LSTM-CRF for named entity recognition[C]. Proceedings of the 32nd Pacific Asia Conf. on Language, Information and Computation. 2018: 531-540.

[2] Peng N, Dredze M. Improving named entity recognition for Chinese social media with word segmentation representation learning[C]. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics. 2016: 149-155.

[3] 高甦 , 金佩 , 张德政 . 基于深度学习的中医典籍命名实体识别研究 [J]. 情报工程 , 2019, 5(1): 113-123.

[4] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C]. Proceedings of the Annual Conference on Neural Information Processing Systems. California: NeurIPS. 2017: 5998-6008.

[5] Yan H, Deng B, LI X, et al. TENER: Adapting Transformer Encoder for Named Entity Recognition [J]. arXiv preprint arXiv:1911.04474, 2019.

[6] 琚生根 , 李天宁 , 孙界平 . 基于关联记忆网络的中文细粒度命名实体识别 [J]. 软件学报 , 2021, 32(8): 2545-2556.

[7] 丁龙 , 文雯 , 林强 . 基于预训练 BERT 字嵌入模型的领域实体识别 [J]. 情报工程 , 2019, 5(6): 65-074.

[8] 石教祥 , 朱礼军 , 望俊成 , 等 . 面向少量标注数据的命名实体识别研究 [J]. 情报工程 , 2020, 6(4): 37-50.

[9] 庄福振 , 罗平 , 何清 , 等 . 迁移学习研究进展 [J]. 软件学报 , 2015, 26(1): 26-39.

[10] Peters M E, Ammar W, Bhagavatula C, et al. Semi-supervised sequence tagging with bidirectional language models [C]. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017: 1756-1765.

[11] Luo L, Yang Z, Yang P, et al. An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition[J]. Bioinformatics, 2018, 34(8): 1381-1388.

[12] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding [C]. Proceedings of Annual Conference of the North American Chapter of the Association for Computational Linguistics. Minnesota: NAACL-HLT. 2019: 4171-4186.

[13] Yosinski J, Clune J, Bengio Y, et al. How transferable are features in deep neural networks? [J]. In Advances in neural information processing systems, 2014(27): 3320–3328.

[14] Tzeng E, Hoffman J, Zhang N, et al. Deep domain confusion: Maximizing for domain invariance [J]. arXiv preprint arXiv:1412.3474, 2014.

[15] 冯建周 , 马祥聪 . 基于迁移学习的细粒度实体分类方法的研究 [J]. 自动化学报 , 2020, 46(8): 1759-1766.

[16] 武惠 , 吕立 , 于碧辉 . 基于迁移学习和 BiLSTM-CRF 的中文命名实体识别 [J]. 小型微型计算机系统 , 2019(6): 1142-1147.

[17] 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.

[18] Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space[J]. arXiv preprint arXiv:1301.3781, 2013.

[19] Peters M E, Neumann M, Iyyer M, et al. Deep contextualized word representations[C]. Association for Computational Linguistics. 2019: 2227-2237.