

基于分布校正的小样本命名实体识别方法

浦震宇¹, 刘志伟¹, 黄勃^{1,2}, 何书锋³, 陈南希⁴, 郝文增⁵

(1. 上海工程技术大学电子电气工程学院, 上海 201620;

2. 华中农业大学农业智能技术教育部工程研究中心, 湖北 武汉 430070;

3. 中国地质调查局青岛海洋地质研究所, 山东 青岛 266071;

4. 中国科学院上海微系统与信息技术研究所, 上海 200050;

5. 天津国投津能发电有限公司, 天津 100032)

摘要: 在现代工业领域, 文本数据的感知和分析已成为推动智能制造和优化生产流程的重要手段。然而, 工业文本数据通常具有高专业性、多样性和复杂性等特点, 且标注成本较高, 因此传统的大规模标注方法难以适用。现有的小样本命名实体识别(NER)方法多采用原型网络对实体进行分类, 其中原型为属于同一类别的所有样本特征的平均值。然而, 这类方法对于支持集数据的敏感性较强, 容易出现样本选择性偏差的问题。为此, 提出基于分布校正的小样本命名实体识别(DC-NER)模型。采用跨度检测和实体分类两阶段方式进行小样本 NER 任务, 在第 1 个阶段, 利用实体抽取器选出候选实体, 在第 2 个阶段, 利用类型判别器将实体划分到预定义好的类别。为了解决小样本 NER 任务中利用少量标签样本难以捕捉到类别真实分布的问题, 进而使得类原型的计算不够准确, 提出一种利用源域数据中的有效信息校正目标域类别分布, 进而通过改进后的类别分布生成更多的样本用于构建更准确的原型, 从而提高其在小样本 NER 任务中的性能。在同领域数据集 Few-NERD 和跨领域数据集 Cross-NER 上的实验结果表明, DC-NER 在 F1 值上显著优于对比模型, 验证其在小样本 NER 中的有效性。

关键词: 工业文本数据; 命名实体识别; 小样本学习; 原型网络; 分布校正

中图分类号: TP391

文献标志码: A

DOI: 10.19678/j.issn.1000-3428.0070449

Few-Shot Named Entity Recognition Method Based on Distribution Calibration

PU Zhenyu¹, LIU Zhiwei¹, HUANG Bo^{1,2}, HE Shufeng³, CHEN Nanxi⁴, XI Wenzeng⁵

(1. School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai 201620, China;

2. Engineering Research Center for Intelligent Technology for Agricultural,

Ministry of Education, Huazhong Agricultural University, Wuhan 430070, Hubei, China;

3. Qingdao Institute of Marine Geology, China Geological Survey, Qingdao 266071, Shandong, China;

4. Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai 200050, China;

5. Tianjin SDIC JinNeng Electric Power Co., Ltd., Tianjin 100032, China)

【Abstract】 In the modern industrial sector, the perception and analysis of text data are essential for promoting intelligent manufacturing and optimizing production processes. However, industrial text data are typically characterized by high specialization, diversity, complexity, and annotation costs, making traditional large-scale annotation methods unsuitable. Existing few-shot Named Entity Recognition (NER) methods often use prototypical networks to classify entities, where the prototype is the average of the features of all samples belonging to the same category. However, these methods are highly sensitive to the support set data and prone to sample-selection bias. To address this, a few-shot NER model based on Distribution Calibration called DC-NER is proposed. A two-stage approach for span detection and entity classification is adopted for the few-shot NER task. In the first stage, an entity extractor is used to select candidate entities. In the second stage, a type discriminator assigns these entities to predefined categories. To address the challenge of capturing the true class distribution with a small number of labeled samples in few-shot NER, which leads to inaccurate class prototype calculations, a method is proposed that leverages useful information from the source domain data to calibrate the class distribution in the target domain. Thus, more samples are generated based on the refined class distribution to construct more accurate prototypes, thereby improving the performance in few-shot NER tasks. Experiments on both the in-domain dataset Few-NERD and the cross-domain dataset Cross-NER reveal that DC-NER significantly outperforms

基金项目: 华中农业大学农业智能技术教育部工程研究中心开放课题(ERCITA-KF002)。

作者简介: 浦震宇, 男, 硕士研究生, 主研方向为自然语言处理; 刘志伟, 硕士研究生; 黄勃(通信作者), 副教授; 何书锋, 高级工程师; 陈南希, 副研究员; 郝文增, 高级工程师。

收稿日期: 2024-10-09

修回日期: 2025-02-13

E-mail: huangbosues@sues.edu.cn

the comparative models in terms of F1 value, validating its effectiveness for few-shot NER.

【Key words】 industrial text data; Named Entity Recognition (NER); few-shot learning; prototypical network; distribution calibration

0 引言

在现代工业领域,随着信息化和智能化的发展,海量的工业文本数据被不断生成和积累。这些数据包含丰富的领域知识,如设备名称、工艺流程、操作规范和故障记录等,对其进行有效的感知和分析,对于提高生产效率、优化工艺流程和预防故障发生具有重要意义。然而,工业文本数据通常具有高专业性、多样性、复杂性等特点,使得传统的文本分析方法难以直接适用,特别是命名实体识别(NER)^[1]任务,在工业文本数据中面临着较大挑战。

NER 任务自提出开始经历了几次重大演变。早期的 NER 是基于规则和字典的方法。这类方法具有很好的可解释性,广泛适用于各种类型的语言和文本,当规则和字典能够充分覆盖文本中出现的命名实体时,这种方法可以获得较高的识别准确度。文献[2]构建了 HunFlair,是一个先进的生物医学命名实体识别工具,通过集成到 Flair NLP 框架,利用预训练的字符级语言模型完成对多种生物医学实体类型的高准确识别。文献[3]为了解决英文的 NER 系统无法应用于其他语言的问题,提出使用规则的马来文命名实体识别算法。文献[4]构建一种针对电子健康记录的字典,通过文本与字典进行匹配的方式取得了不错的效果。这种传统的 NER 方法高度依赖于规则和字典的质量和覆盖范围,需要不断地进行更新和维护,在处理新的领域或未知语言时通用性较差。

考虑到基于规则和字典的方法无法解决文本中存在的歧义和语法结构多样性等问题,研究人员探索采用机器学习的方式学习训练数据中的语言特征和规律,并根据不同的任务进行调整和优化。文献[5]基于条件随机场(CRF)^[6]模型构建一个输出平面或分层的捷克语 NER 系统。文献[7]提出一个用于识别社交媒体消息的系统,将多通道信息与 CRF 输出的综合单词表示合并到传统的双向长短期记忆(BiLSTM)网络中。文献[8]提出基于 PMV-LSTM 的中文医学领域命名实体识别方法,利用 Lattice LSTM 结构动态地将词汇信息和字符信息融合以解决语义缺失问题,通过条件随机场解码得到实体标签。文献[9]提出一种基于 BERT-BiLSTM-MHA-CRF 的中文命名实体识别方法,基

于 BERT 预训练,通过 BiLSTM 网络结合多头注意力机制,从多个维度提取文本的语义特征,最后通过 CRF 层来获取全局最优的标签序列。

随着大数据时代的到来,越来越多的标注数据可以用于训练深度学习模型。这种方法可以从原始数据中学习到更加复杂和抽象的特征和表示,具有更强的灵活性和泛化能力。文献[10]提出一种结合 BiLSTM 网络和 CRF 的模型,用于完成序列标注任务。文献[11]提出一种 BiLSTM-卷积神经网络(CNNs)-CRF 的模型,通过卷积操作对字符进行编码,有效提升命名实体模型的性能。在中文领域,文献[12]提出一种融合字形信息的命名实体识别算法,将 CNN 获得的信息与双仿射编码层信息相融合,最后与 CNN 层的输出相加,得到最终预测结果。文献[13]提出一种基于 BERT-BiLSTM-CRF 的神经网络模型,用于提取生物医学领域的命名实体,解决了静态词向量表征语义方法准确率不高的问题。

在实际场景中,许多领域都存在着数据稀缺和标注困难等问题。为了解决这些问题,研究人员利用小样本学习技术,旨在仅有少量标注样本的情况下,实现模型的最佳识别效果。文献[14]提出一种用于学习具有相似性度量的样本之间关系的匹配网络。通过最小化损失函数来学习相似性度量函数,使得在给定支持集的情况下,模型能准确地预测查询样本的标签。文献[15]提出原型网络,在训练阶段,原型网络通过学习支持集中的样本来生成每个类别的原型向量。在测试阶段,对于一个查询样本,原型网络将其表示为一个特征向量,并计算它与每个类别原型向量之间的距离。文献[16]提出一种基于分解的元学习方法,通过引入与模型无关的元学习^[17]算法来训练跨度检测器,可以快速找到适应新实体类型的最佳模型,用于初始化参数。文献[18]提出一种两阶段的命名实体识别模型,在跨度抽取部分将标签转换为全局边界矩阵,使模型能够专注于显式边界信息。尽管这些方法都取得了不错的效果,但是仅依靠少量样本推测类的分布或原型仍然不够准确。

在现代工业领域,文本数据的感知和分析是推动智能制造和优化生产流程的关键。然而,工业文本数据的高专业性、多样性和复杂性,以及高昂的标

注成本,限制了传统大规模标注方法的适用性,使得小样本学习成为工业文本数据分析中的一个重要研究方向。现有的 NER 方法在处理小样本数据时,由于样本选择性偏差,难以准确捕捉类别的真实分布,导致识别的类别原型不准确。为了解决这一问题,本文提出一种基于分布校正的小样本命名实体识别 DC-NER 模型。DC-NER 模型通过两阶段方法来解决小样本敏感数据问题,利用实体抽取器识别候选实体,然后通过精准的距离度量函数,模型识别源域与目标域之间的相似类别,并据此校正目标域样本分布,生成更准确的类别原型。这种分布校正机制显著提高模型在小样本情况下的泛化能力,尤其是在处理工业文本数据的高专业性和复杂性背景下。

此外,DC-NER 模型在设计时考虑了跨领域应用的需求,通过在多个公开数据集上的评估证明其不同领域中的有效性和泛化能力。特别地,DC-NER 模型针对工业文本数据的特点进行优化,使其在处理这类数据时更有效。实验结果表明,DC-NER 在 F1 值方面优于对比模型。本文验证了 DC-NER 在小样本命名实体识别中的有效性。

1 相关工作

1.1 基于提示的方法

随着提示学习^[19]在深度学习领域中的应用,模型通过引入一种标准化的语言或模板,以指导模型的学习方向,从而提高模型的泛化能力。在提示学习中,通常会设计一些预定的提示或问题,要求模型根据这些提示或问题进行学习和推理。文献[20]提出一种软硬提示相结合的方法,使得模型可以在离散空间中自动搜索合适的模板。文献[21]设计一个双槽多提示模板,其中位置槽和类型槽分别提示定位和分类,并在训练期间通过扩展二部图匹配的方式为对应槽分配标签。现有的提示方法大多依赖人工设计的复杂模板,通过枚举的方法对所有模板进行测试,时间消耗较大。为了解决这类问题,文献[22]在放弃模板构建过程的同时保留了预训练模型的单词预测范式,以预测实体位置处与类相关的标签。

1.2 基于度量的方法

基于度量的小样本命名实体识别方法的核心思想是将样本表示为特征向量,并在特征空间中度量样本之间的距离或相似度。一般来说,相似的样本在特征空间中会更加接近,而不相似的样本则会更加远离。通过学习这种相似性度量,模型可以根据少量标注样本在特征空间中的分布来进行分类或聚类。文献[23]提出使用对比学习的策略,对高斯嵌

入进行建模并优化单词间的分布距离,有效缓解源域训练过程中的过拟合问题。文献[16]利用元学习策略依次解决少样本 NER 的跨度检测和实体分类问题,具有有效的参数初始化,这使得模型能够通过充分探索支持示例中包含的信息来快速适应新的实体类,并在实体分类阶段通过原型网络对测试样本进行分类。文献[24]为了充分考虑标签信息,提出利用对比学习的方式对样本和标签信息进行融合,以构造更加准确的原型,并提出一种新的跨度过滤策略来删除识别错误的实体跨度。

2 模型构建

为了解决小样本命名实体识别任务中基于少量样本计算的原型不能代表真实样本分布的问题,本文提出基于分布校正的小样本命名实体识别模型。该模型将源域数据中的相似特征与目标域的少量样本进行融合,可以获得更准确的类别分布。通过校正后的样本分布求出更准确的原型,可以提高实体的分类性能。模型主要分为两个部分:跨度检测器和类型判别器。DC-NER 模型的主要结构如图 1 所示(彩色效果见《计算机工程》官网 HTML 版,下同)。

图 1 展示了 DC-NER 模型的两个核心部分:跨度检测器和类型判别器,它们共同构成了模型特征提取和处理的骨架。

跨度检测器负责从输入文本中识别出潜在的实体边界,其嵌入层利用 BERT 模型将文本序列转换为包含语义信息的向量表示,这些向量捕捉了词汇的上下文关系和句法结构。双仿射编码层进一步对这些特征进行编码,为实体的开始和结束位置生成详细的特征表示,从而增强模型对实体边界的识别能力。最终,输出层基于全局边界矩阵的概念,对实体的位置信息和上下文语境之间的关联程度进行建模,以提高实体识别的准确性。

图 1 不仅展示了模型的结构,而且通过其组件和流程的设计,隐含地可视化了特征提取的有效性。每个组件都针对特定的特征提取任务进行优化,共同工作以实现命名实体的高精度识别。

2.1 跨度检测器

2.1.1 嵌入层

该部分利用 BERT^[25]将输入文本编码为高维空间中的向量表示。BERT 模型通过在大规模文本数据集上进行预训练,学习丰富的语言特征和深层的语义信息。与传统的词嵌入方式相比,BERT 模型能够捕捉更丰富的上下文信息。

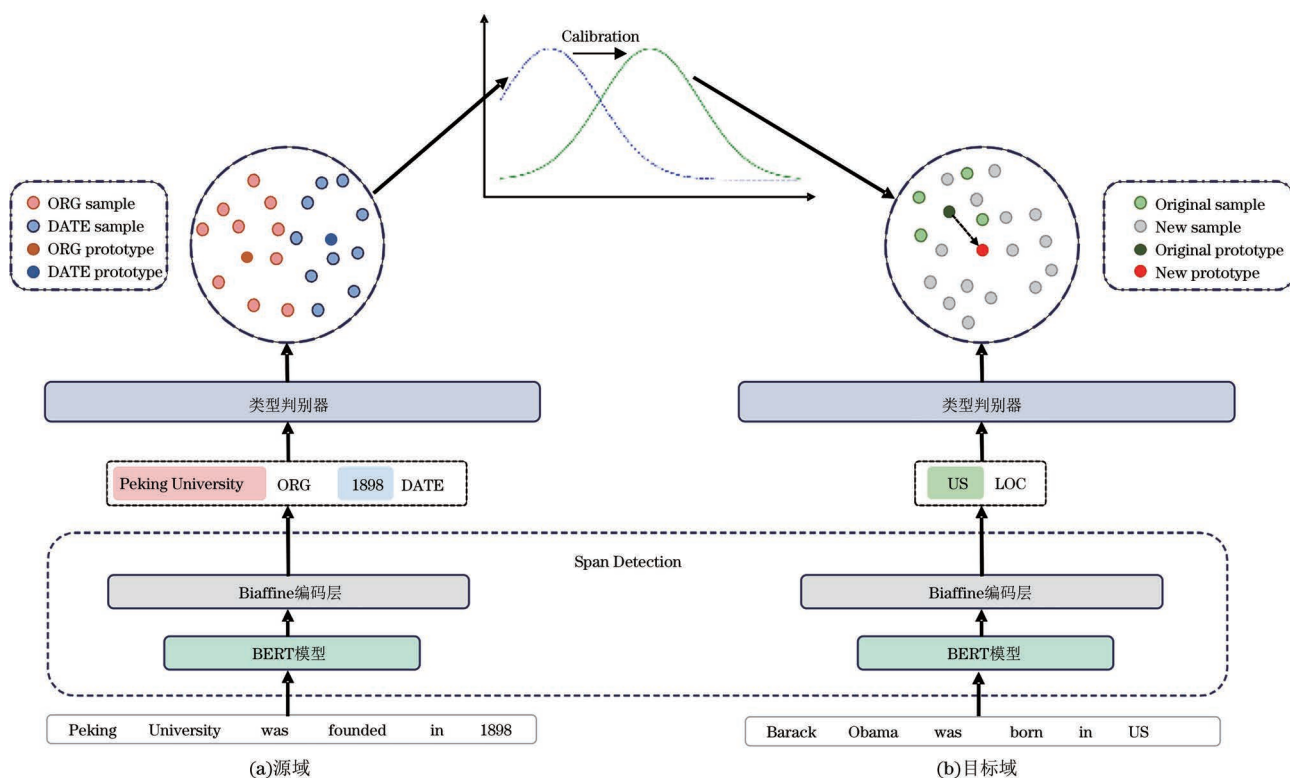


图 1 DC-NER 框架

Fig. 1 DC-NER architecture

BERT 模型的核心是 Transformer 架构,通过自注意力机制来捕捉序列中各元素之间的关系。这种机制使得模型能够识别出文本中长距离的依赖关系,并且能够为每个单词生成一个包含其上下文信息的向量表示。在本文模型中,BERT 作为嵌入层,负责将输入的文本序列转换为一系列高维的、包含丰富语义信息的向量。

对于一个长度为 n 的输入序列,BERT 模型通过其编码器对每个字符进行编码,最终获得 BERT 模型最后隐藏层的输出 $H = \{h_1, h_2, \dots, h_n\} \in R^{n \times h}$ 。这个输出作为输入序列的语义编码表示,其中 h 表示词向量的维度。这些向量不仅包含单词本身的语义信息,还包含其在句子中的语法和语义角色,为后续的实体识别任务提供了强有力的语义特征。语义编码表示如式(1)所示:

$$[h_1, h_2, \dots, h_n] = \text{PLM}([x_1, x_2, \dots, x_n]) \quad (1)$$

2.1.2 双仿射编码层

在命名实体识别任务中,Biaffine 双仿射编码层^[26-27]通常用于对输入序列中的标记进行分类,并将它们映射到预定义的命名实体类别中。Biaffine 层可以接受来自先前神经网络层的输入表示,并进一步对其进行编码和提取特征。这种编码过程涉及到前馈神经网络(FFNN),是一种基本的神经网络架构,它通过一系列全连接层传递信息,每层都包含

可以学习权重的神经元。在 FFNN 中,每个神经元的输出是其输入的加权和,经过非线性激活函数处理后的结果。这种结构使得 FFNN 能够学习输入数据中的复杂模式和关系。这些特征通常是输入序列中每个词或标记的语义表示,以及它们与周围词或标记的上下文信息。由于实体跨度的开始位置和结束位置的上下文片段不同,因此采用两个独立的前馈神经网络为实体的开始位置 x_{s_i} 和结束位置 x_{e_i} 分别创建不同的特征表示,可以更好地捕捉到实体的全局上下文信息,提高模型对实体边界的识别能力。特征表示如式(2)和式(3)所示:

$$h_s(i) = \text{FFNN}_s(x_{s_i}) \quad (2)$$

$$h_e(i) = \text{FFNN}_e(x_{e_i}) \quad (3)$$

式中: $h_s(i)$ 和 $h_e(i)$ 表示经过变换后的特征表示。随后,利用 Biaffine 编码器可以求出实体边界得分矩阵 P_x ,如式(4)所示:

$$P_x = h_s(i)^T W_a h_e(i) + W_b (h_s(i) \oplus h_e(i)) + b_m \quad (4)$$

式中: W_a, W_b 为可训练参数矩阵; b_m 表示偏置项。

2.1.3 输出层

为了增强模型对实体边界的建模能力,从而提高 NER 模型的性能和鲁棒性,本文基于全局边界矩阵的思想^[18]对实体在序列中的位置信息以及上下文语境之间的关联程度进行建模。通过学习

这些边界之间的关系,模型可以更准确地识别实体的边界,避免错误地分割实体或漏掉实体的一部分。通常来说,实体边界矩阵是一个方阵。在常规的下三角矩阵相关运算和理解中,通常下三角矩阵的元素满足行索引大于等于列索引的条件。然而,在实际情况中,可能受数据来源、数据处理过程中的误差或者其他原因的影响,出现了结束位置下标超过开始位置下标这种不符合常规逻辑的情况。为了避免这种异常数据对后续运算和分析造成不良影响,通常只选择上三角矩阵进行实际的运算。

$$\Omega_{s_j, e_j} = \begin{cases} 1, & s_j \leq e_j \wedge (s_j, e_j) \in M \\ 0, & s_j \leq e_j \wedge (s_j, e_j) \notin M \\ -\infty, & s_j > e_j \end{cases} \quad (5)$$

式中: s_j, e_j 表示第 j 个实体的开始和结束位置; Ω_{s_j, e_j} 表示实体 j 的全局得分; M 表示被判断为实体的跨度。

最终,模型通过交叉熵函数对模型损失进行计算,如式(6)所示:

$$L_{\text{span}} = -\log_a \left(1 + \sum_{1 \leq s_j \leq e_j \leq L} \exp((-1)^{\Omega_{s_j, e_j}} (\mathbf{P}_x)) \right) \quad (6)$$

2.2 类型判别器

为了克服模型在少样本场景中面临的挑战,即少量样本分布难以代表整个类的样本分布。本文利用源域与目标域样本相似的特征进行分布校正。在命名实体识别任务中,类别差距越小的类在特征空间中的距离也越近。假设类别分布服从高斯分布,那么属于该类的均值和方差代表了其在空间中的中心位置和分布形状。基于这个观察,如果能够衡量两个类别之间的相似性,就可以将基类中学到的相似信息扩展到目标域,从而帮助模型更准确地推断目标域类别的真实分布。

2.2.1 源域信息统计

模型通过在具有丰富样本多样性的源域数据集上进行全面训练,能够捕获更全面且精细的特征,并且其特征分布更贴近于真实世界的分布。假设在源域数据集中, i 为其中的一个特定的类,包含样本 $X = \{\mathbf{x}_j\}_{j=1}^{n_i}$, 其中, n_i 表示属于类别 i 的样本特征总数。因此,可以通过式(7)计算该类别在特征空间中的均值 $\boldsymbol{\mu}_i$ 。这个均值是通过将类别 i 下所有样本特征向量的各个维度进行算术平均得到的。

$$\boldsymbol{\mu}_i = \frac{\sum_{j=1}^{n_i} \mathbf{x}_j}{n_i} \quad (7)$$

由于特征向量在空间中是多维的,为了更准确地理解这些特征之间的相互关系,因此使用协方差 $\boldsymbol{\Sigma}_i$ 代表类别 i 的方差,如式(8)所示:

$$\boldsymbol{\Sigma}_i = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (\mathbf{x}_j - \boldsymbol{\mu}_i)(\mathbf{x}_j - \boldsymbol{\mu}_i)^T \quad (8)$$

因此, $N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ 可以代表源域数据集中类别 i 的分布。

2.2.2 目标域分布校正

在小样本学习中,考虑到目标域中样本数量较少,为了使得目标域中类的空间分布更加接近高斯分布,本文使用图基变换对目标域中的支持样本和查询样本进行特征变换。

$$\mathbf{x}' = \begin{cases} \mathbf{x}^\lambda, & \lambda \neq 0 \\ \log_a(\mathbf{x}), & \lambda = 0 \end{cases} \quad (9)$$

式中: $\mathbf{x}'$ 表示经过变换后的向量; λ 是一个超参数。

对于查询样本 $\bar{\mathbf{x}}$, 计算它与源域中每个类别均值向量 $\boldsymbol{\mu}_i$ 的距离,将求得的所有结果存入集合 D_s 中,在这里选用欧几里得距离作为距离度量函数,如式(10)所示:

$$D_s = \{ -\|\boldsymbol{\mu}_i - \bar{\mathbf{x}}\|, i \in C_b \} \quad (10)$$

式中: C_b 为源域中的类别集合。

最后,通过挑选集合 D_s 中值最大的 K 个类别作为参考,即可得到与目标域数据样本最相关的 K 个类别。

$$D_{\text{top-}K} = \{ i \mid \|\boldsymbol{\mu}_i - \bar{\mathbf{x}}\| \in \text{top-}K(D_s) \} \quad (11)$$

式中: 集合 $D_{\text{top-}K}$ 包含目标域样本与源域中最相关的 K 个类; $\text{top-}K(\cdot)$ 是一个运算符。

本文选择欧几里得作为距离度量函数,用于计算查询样本与源域中每个类别均值向量之间的距离。欧几里得距离函数是最常用的距离度量之一,它提供了一种直观的方法来衡量样本之间的直线距离。这种距离度量方法简单直观,易于理解和计算,在许多机器学习任务中都有广泛的应用。在本文模型中,欧几里得作为距离函数,有助于准确地识别源域与目标域之间的相似类别,并据此校正目标域样本分布,生成更准确的类别原型。

最终,通过式(12)和式(13)对目标域分布进行校正。

$$\boldsymbol{\mu}' = \frac{\sum_{i \in C_b} \boldsymbol{\mu}_i + \bar{\mathbf{x}}}{K + 1} \quad (12)$$

$$\boldsymbol{\Sigma}' = \frac{\sum_{i \in C_b} \boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_i}{K} + \alpha \quad (13)$$

式中: $\boldsymbol{\mu}'$ 表示校正后的目标域样本均值,将目标域中的样本 $\bar{\mathbf{x}}$ 与源域中相似类别的样本均值结合起来,

以 $K+1$ 为权重进行平均,从而得到校正后的均值,这种方法有助于减少因目标域样本数量不足而造成的估计偏差; Σ' 表示校正后的目标域样本方差; α 是一个调整参数,用于控制新样本对方差的影响,这种方法有助于在目标域样本较少时,调整方差的估计,使其更加稳定和可靠。

2.2.3 原型分类

由于校正后的样本分布更符合真实的数据分布,因此模型通过校正后的分布 $N(\mu', \Sigma')$ 生成 N 个新的样本,并存储在集合 D_T 中。

$$D_T = \{(\mathbf{x}, y) | \mathbf{x} \sim N(\mu', \Sigma')\} \quad (14)$$

此时,根据目标域新生成的样本可以计算目标域类别 k 中更准确的原型表示 c_k ,如式(15)所示:

$$c_k = \frac{1}{N} \sum_{(\mathbf{x}, y_k) \in D_k} f_\phi(\mathbf{x}) \quad (15)$$

模型通过 Softmax 函数计算类别概率,并通过损失函数 L_γ 优化模型参数。

$$p_\phi(y=k | \mathbf{x}) = \frac{\exp(-d(f_\phi(\mathbf{x}), c_k))}{\sum_{k'} \exp(-d(f_\phi(\mathbf{x}), c_{k'}))} \quad (16)$$

$$L_\gamma = \sum -\log_\alpha p_\phi(y=k | \mathbf{x}) \quad (17)$$

3 实验结果与分析

3.1 数据集

本文实验在同领域和跨领域的 5 个数据集上展开,表 1 展示了各个数据集的统计数据,包含数据集的领域、实体类别数量和数据规模。

表 1 数据集统计信息

Table 1 Dataset statistics information 单位:个

数据集	领域	类别数	数据量
Few-NERD	维基百科	66	188.2×10^3
GUM	维基百科	11	3.5×10^3
OntoNotes	通用	18	76.7×10^3
CoNLL-2003	新闻	4	20.7×10^3
WNUT2017	社交媒体	6	5.7×10^3

3.1.1 Few-NERD 数据集

Few-NERD^[28]是第一个专门用于小样本 NER 任务的数据集,按照 N -way K -shot 形式划分,包含 8 种粗粒度实体类型和 66 种细粒度实体类型,且每个实体标签均为粗粒度与细粒度相结合的层级结构。其中,Intra 为按照粗粒度实体进行分类,训练集与测试集之间的相关性较低,使得该任务更具挑战性;Inter 为按照细粒度类型进行划分,每个集合都涉及到所有粗粒度实体类。

3.1.2 Cross-NER 数据集

为了评估本文模型在不同领域的粗粒度 NER 数据集上的表现,本文在 4 个命名实体识别的公开数据集上进行相关实验,包括 GUM^[29]数据集、OntoNotes^[30]数据集、CoNLL-2003^[31]数据集和 WNUT2017^[32]数据集。其中,GUM 数据集的语料源自维基百科,具有极高的多样性和覆盖面,不仅包含新闻报道,还涵盖博客文章、用户评论以及文学作品等多种形式的文本。这种多样性为研究人员提供了丰富的语境来训练和测试模型,但同时也带来了挑战,因为不同文本类型之间的语言风格和用词习惯可能大相径庭。此外,维基百科的条目通常包含专业知识,这对模型理解上下文和专业术语的能力提出了更高的要求。OntoNotes 是一个大规模的多语言语料库,它不仅覆盖广泛的领域,还针对多个自然语言处理任务提供了详细的标注,如词性标注、句法分析和语义角色标注。然而,由于其标注的复杂性和多样性,对模型的学习和泛化能力提出了挑战。研究人员需要开发能够处理复杂结构和多层次标注信息的算法。CoNLL-2003 数据集专注于新闻领域的语料,其经过专家的详细手工标注,因此在质量上具有很高的可靠性。这一数据集特别适用于评估模型在新闻文本上的命名实体识别能力。然而,新闻文本通常具有明确的结构和专业术语,这对模型捕捉行业特定模式的能力构成了挑战。此外,新闻报道的时效性要求模型能够适应随时间变化的语言使用。WNUT2017 数据集收集自社交媒体、微博和其他用户生成的内容,这些文本通常包含口语化表达、非标准语法和大量噪声,如拼写错误和表情符号。这些特点使得 WNUT2017 成为一个更具挑战性的实验场景,因为它们要求模型不仅能够处理非正式语言,还要能够识别和过滤噪声。此外,用户生成的内容通常涉及广泛的主题和个人表达,这增加了模型理解和分类实体的难度。

3.2 实验设置

3.2.1 对比模型

为验证所提模型的有效性,本文在同领域实验中选择目前效果较好的几个模型进行比较。

1)ProtoBERT^[15]。该模型采用 BERT 作为编码器,通过计算样本与各类别原型之间的距离作为划分依据。初始学习率设置为 1×10^{-3} ,并在每训练 2 000 个 Epoch 后减半。在 BERT 模型中使用批量归一化来加速训练并提高泛化能力。使用 Adam 优化器,动量参数为 0.9~0.999。

2) NNShot^[33]。该模型使用 BERT 进行词嵌入,并在分类时使用 k -近邻算法。在训练监督 NER 模型时,使用交叉熵损失函数。对于 BERT 模型,使用了 Adam 优化器,学习率设置为 5×10^{-5} 。

3) StructShot^[33]。该模型在测试时使用维特比解码的标签转移概率来改进 NNShot 模型。

4) CONTaiNER^[23]。该模型通过对比学习优化了实体之间的分布距离,并利用高斯分布嵌入来区分标记类别。

5) ESD^[34]。该模型将命名实体识别任务分解为一系列基于实体跨度预测的子任务,并为非实体类设计多个原型。对于 BERT 编码器学习率设置为 2×10^{-5} ,对于其他模块,学习率设置为 5×10^{-4} ,Dropout 为 0.1,最大跨度长度为 8。

6) DecomMeta^[16]。该模型通过元学习策略训练模型来学习良好的参数初始化,通过找到一个良好的嵌入空间,可以更好地区分来自不同实体类的文本跨度表示。学习率为 3×10^{-5} ,批大小设置为 32,序列最大长度设置为 128,Dropout 概率保持为 0.1。

7) TadNER^[24]。该模型构建一种类型感知对比学习策略来构建原型,并引入一种类型感知跨度过滤策略用于识别错误跨度。

8) SpanProto^[18]。该模型引入全局边界矩阵来促使模型学习显式边界信息,并通过一种新的损失方法用于去除错误识别的实体跨度。

在跨领域实验中,除选取上述 ProtoBERT、DecomMeta、SpanProto 模型以外,本文还选择了如下模型进行对比。

1) Matching Network^[14]。该模型将 ProtoBERT 模型中待测样本与类原型的距离度量换成了相似度度量。

2) L-TapNet+CDT^[35]。该模型在条件随机场中引入了标签转移机制,并提出一种使用标签语义增强的自适应投影网络来解决源域与目标域标签差异的问题。

3.2.2 实验环境与参数设置

本实验基于 PyTorch 框架进行代码编写,运行于 Python 3.8 环境下。所使用计算机配置环境的硬件参数为处理器 Intel i5-13490F、显卡 NVIDIA GeForce GTX 3090。详细参数如表 2 所示。

本文通过一系列实验来确定模型的最佳参数配置,以确保模型在训练过程中既能有效学习数据模式,又能保持良好的泛化能力。以下是本文对关键参数进行调整和分析的结果。

表 2 参数设置

Table 2 Parameter setting

参数	值	选择原因
Epoch	100	100 个 Epoch 平衡了学习效果和训练时间,避免过拟合
Batch Size	8	较小的 Batch Size 提高了泛化能力,尽可能减少训练时间
Learning Rate	5×10^{-5}	此学习率可以保证本模型稳定收敛
Dropout	0.15	在减少过拟合和保持模型性能之间取得最佳平衡
λ	0.5	与 α 一起有效调整正则化项,提高泛化能力
α	0.35	与 λ 一起有效调整正则化项,提高泛化能力
句子最大长度	128	平衡计算效率和模型性能,处理大多数实际情况
Biaffine 隐藏层维度	150	提供足够参数学习特征表示,同时保持计算效率

1) Epoch。测试了 50、100 和 150 个 Epoch,发现 100 个 Epoch 能够平衡模型的学习效果和训练时间,使模型有足够的时间学习数据中的模式,同时避免过拟合。

2) Batch Size。本文测试了 4、8 和 16 的不同配置。实验结果表明,较小的 Batch Size 有助于提高模型训练过程中的泛化能力,尽管这可能会增加训练时间。因此,本文选择 8 作为批次大小。

3) Learning Rate。本文测试 1×10^{-5} 、 5×10^{-5} 和 1×10^{-4} 等不同的值。实验结果表明, 5×10^{-5} 的学习率有助于模型在训练过程中稳定收敛。

4) Dropout。为了防止过拟合,引入 Dropout 机制,并测试了 0.1、0.15 和 0.2 的不同 Dropout 率。实验结果表明,0.15 的 Dropout 率在减少过拟合和保持模型性能之间取得了最佳平衡。

5) 正则化参数。引入 2 个正则化参数 λ 和 α ,实验结果表明, $\lambda=0.5$ 和 $\alpha=0.35$ 的设置能够有效地调整模型中的正则化项,以进一步提高模型的泛化能力。

6) 句子最大长度。考虑到计算效率和模型性能的平衡,本文将句子的最大长度设置为 128。这一设置有助于减少不必要的计算量,同时确保模型能够处理大多数实际情况。

7) Biaffine 隐藏层维度。Biaffine 隐藏层维度设置为 150,这一选择是基于实验和性能测试的结果。本文测试了 100、150 和 200 的不同配置,发现 150 的维度能提供足够的参数来学习丰富的特征表

示,同时保持计算效率。

3.2.3 评价指标

本实验采用 F1 值 (F_1) 进行评价。精确率 (P)、召回率 (R) 和 F_1 计算如式 (18)~式 (20) 所示:

$$P = \frac{N_{TP}}{N_{TP} + N_{FP}} \quad (18)$$

$$R = \frac{N_{TP}}{N_{TP} + N_{FN}} \quad (19)$$

$$F_1 = \frac{2 \times P \times R}{P + R} \quad (20)$$

式中: N_{TP} 为模型正确预测正类样本为正类的数量; N_{FP} 为模型将负类样本预测为正类的数量; N_{FN} 为模型将真实标签为正类的样本预测为负类的数量。

3.3 对比实验

3.3.1 同领域实验

表 3 和表 4 分别展示了 DC-NER 与对比模型在 Few-NERD Intra 和 Inter 数据集上的实验结果,加粗表示最优数据(下同)。图 2 和图 3 以柱状图的形式更直观地展示了本模型与对比模型在 Few-NERD 数据集下的 F1 值。本文提出的 DC-NER 在大多数情况下均优于对比模型,且相比于基线模型 SpanProto,在 (1~2)-shot 情况下提升更为明显。在 (5~10)-shot 设置下,标签样本较多,相比 (1~2)-shot 情况下对原型的计算更准确。从原因上看,传统的命名实体识别模型基于大量标签数据进行训练,可以更准确地反映样本的真实分布。当模型仅使用单个样本确定类型原型时,模型容易学习到样本的特征而不是更广泛的类别分布。此时,利用源域数据中学到的相似性知识对目标域的少量特征进行数据增强,可以对新类样本分布进行校正。

表 3 在 Few-NERD Intra 数据集上的 F1 值

Table 3 F1 value on Few-NERD Intra dataset %

Model	(1~2)-shot		(5~10)-shot	
	5-way	10-way	5-way	10-way
ProtoBERT	23.45	19.76	41.93	34.61
NNShot	31.01	21.88	35.74	27.67
StructShot	35.92	25.38	38.83	26.39
CONTaiNER	40.43	33.84	53.70	47.49
ESD	41.44	32.29	50.68	42.92
DecomMeta	52.04	43.50	63.23	56.84
TadNER	60.78	55.44	67.94	60.87
SpanProto	52.19	44.03	67.76	59.97
DC-NER	56.22	49.71	70.54	62.37

表 4 在 Few-NERD Inter 数据集上的 F1 值

Table 4 F1 value on Few-NERD Inter dataset %

Model	(1~2)-shot		(5~10)-shot	
	5-way	10-way	5-way	10-way
ProtoBERT	44.44	39.09	58.80	53.97
NNShot	54.29	46.98	50.56	50.00
StructShot	57.33	49.46	57.16	49.39
CONTaiNER	55.95	48.35	61.83	57.12
ESD	66.46	59.95	74.14	67.91
DecomMeta	68.77	63.26	71.62	68.32
TadNER	64.83	64.06	72.12	69.94
SpanProto	71.30	65.24	77.47	73.94
DC-NER	74.61	67.19	79.27	76.32

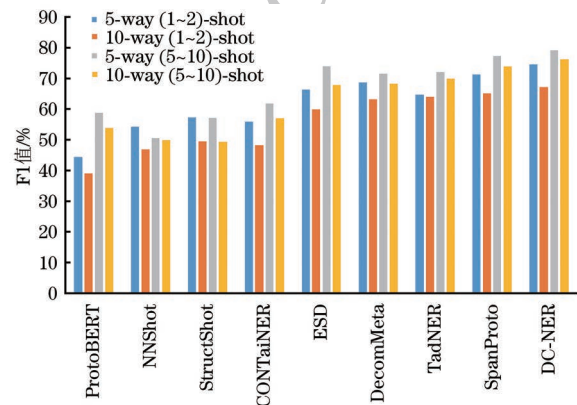


图 2 DC-NER 与对比模型在 Few-NERD Inter 数据集下实验结果

Fig.2 Experimental results of DC-NER and comparison models under Few-NERD Inter dataset

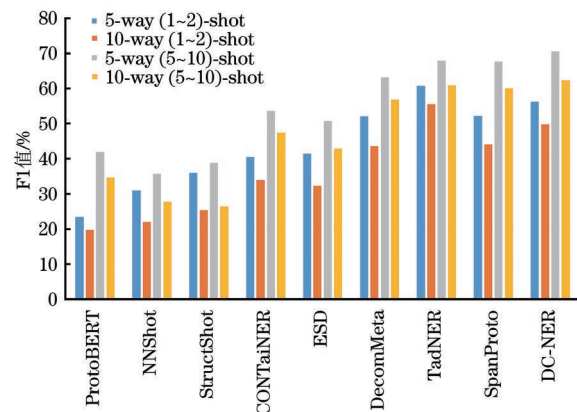


图 3 DC-NER 与对比模型在 Few-NERD Intra 数据集下实验结果

Fig.3 Experimental results of DC-NER and comparison models under Few-NERD Intra dataset

图 4 直观地呈现了模型在数据集上的 F1 值随样本数量变化的实验结果。实验结果表明,对

于不同的数据集,需要采用不同的采样策略。总体而言,在采样数为 200~400 个时,DC-NER 模型表现出较好的性能。然而,随着样本数量的增加,模型往往会学习过多的噪声信息,而非目标领域特定类别的真实分布。该结果强调了适当应用分布校正技术的重要性,可以有效利用支持集中的相似样本来优化目标领域原型的分布表示,从而显著提高模型性能。

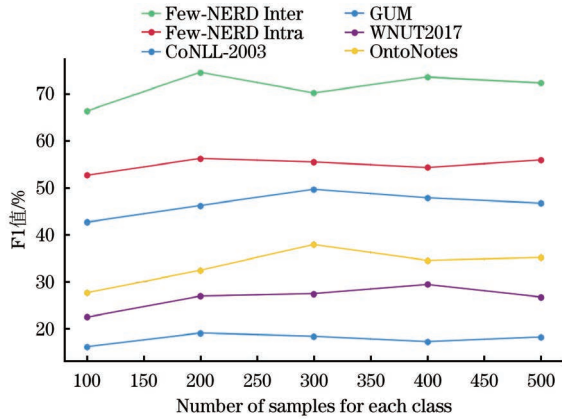


图 4 模型性能随采样数量变化的实验结果

Fig.4 Experimental results of model performance with varying sampling sizes

3.3.2 跨领域实验

为了进一步验证所提模型在跨领域数据下的泛化性能,本文在 CrossNER 数据集上进行实验,结果如表 5 和表 6 所示。图 5 和图 6 以折线图的形式直观地展示 DC-NER 模型与对比模型在跨领域的 4 个数据集下的 F1 值。从实验结果来看,DC-NER 模型在所有条件下均优于对比模型。在 1-shot 和 5-shot 设置下,相对于基线模型 SpanProto,该模型的性能分别提升了 2.28 和 1.59 百分点,这表明通过分布校正使得目标域样本的分布更加准确,从而获得更准确的原型。在少量样本的情况下,尤其是单个样本的情况下,每个类别的原型通常就是该类别中唯一的样本表示,使得异常值或噪声样本可能会显著偏移原型位置,导致原型不准确。随着分布校正策略生成新的样本数量增多,每个样本对原型的贡献相对减少,从而减少了噪声样本的影响。本文进一步分析 WNUT2017、OntoNotes、GUM 和 CoNLL-2003 数据集的标签,尽管这些数据来自不同领域,但它们包含的实体类型存在重叠。即使语料库不同,但由于标签相似,编码后的特征也会相似,这意味着可以利用源域数据中的相似特征对目标域的少量样本进行更准确的分布校正,从而提高分类的准确性。

表 5 各模型在跨领域 1-shot 的实验结果

Table 5 Experimental results of each model in the cross-domain 1-shot

Model	F1 值			
	CoNLL-2003	GUM	WNUT2017	OntoNotes
Matching Network	19.50	4.73	17.23	15.06
ProtoBERT	32.49	3.89	10.68	6.67
L-TapNet+CDT	44.30	12.04	20.80	15.17
DecomMeta	46.09	17.54	25.14	34.13
SpanProto	46.92	16.40	27.67	35.86
DC-NER	49.66	19.03	29.37	37.90

表 6 各模型在跨领域 5-shot 的实验结果

Table 6 Experimental results of each model in the cross-domain 5-shot

Model	F1 值			
	CoNLL-2003	GUM	WNUT2017	OntoNotes
Matching Network	19.85	5.58	6.61	8.08
ProtoBERT	50.06	9.54	17.26	13.59
L-TapNet+CDT	45.35	11.65	23.30	20.95
DecomMeta	58.18	31.36	31.02	45.55
SpanProto	58.59	34.86	30.22	46.97
DC-NER	60.81	36.27	31.64	48.29

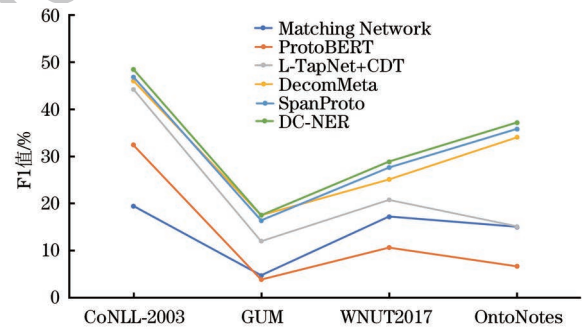


图 5 DC-NER 与对比模型在跨领域数据集 1-shot 下的实验结果

Fig.5 Experimental results of DC-NER and comparison models under cross-domain dataset 1-shot

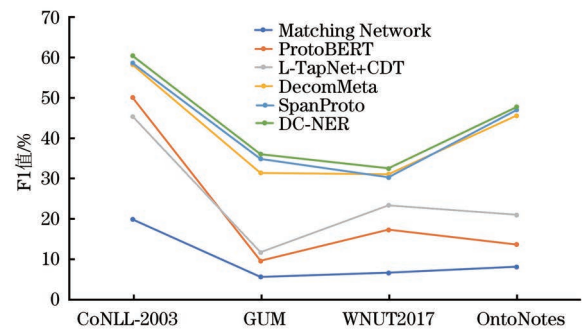


图 6 DC-NER 与对比模型在跨领域数据集 5-shot 下的实验结果

Fig.6 Experimental results of DC-NER and comparison models under cross-domain dataset 5-shot

3.4 消融实验

为验证 DC-NER 模型主要组成部分的有效性, 本文在 Cross-NER 数据集上设置 3 个实验。

1) w/o. Entity Span Extractor 表示模型去除实体跨度抽取器, 并使用基于单个词分类的网络结构。

2) w/o. Prototypical Network 表示模型不使用原型网络进行分类, 而使用 k -近邻算法进行分类。

3) w/o. Distribution Calibration 表示模型不使用校正后的样本分布生成新样本, 只选用支持集中样本进行原型计算。

表 7 展示了 DC-NER 模型在 CoNLL-2003、GUM、WNUT2017 和 OntoNotes 数据集下的消融实验结果。从表 7 可以看出, 模型的任意一个组件都发挥着一定作用。当模型去除实体跨度抽取器后, 模型的 F1 值下降了 8.14 个百分点, 这说明基于跨度相比于基于单词的特征提取更有效。当模型去除原型网络后, 模型性能下降了 2.18 个百分点, 这表明原型网络相比于 k -近邻算法对待测样本的分类更加准确。在去除分布校正模块后, 模型的性能下降了 4.5 个百分点, 这说明利用单个样本估计类别分布的效果较差。通过源域数据中学到的相似性特征增加目标域样本, 进而校正目标域分布, 可以实现更好的识别效果。

表 7 消融实验结果

Table 7 Ablation experimental results

Methods	F1 值			
	CoNLL-2003	GUM	WNUT2017	OntoNotes
DC-NER	49.66	19.03	29.37	37.90
w/o. Entity Span Extractor	36.24	11.52	23.61	32.04
w/o. Prototypical Network	46.76	15.94	24.35	37.28
w/o. Distribution Calibration	42.84	15.29	25.34	34.51

4 结束语

在现代工业环境中, 对文本数据的有效感知和分析是提升生产效率和优化工艺流程的关键。然而, 由于工业文本数据的高专业性和复杂性, 传统的大规模标注方法在实际应用中面临诸多挑战。本文提出了一种基于分布校正的小样本命名实体识别模型。该模型创新性地将任务分解为跨度检测和实体分类两个阶段。在实体分类阶段, 通过精准的距离

度量函数识别源域与目标域间的相似类别, 并据此校正目标域样本分布, 生成更准确的类别原型。实验结果表明, 该模型在小样本实体命名识别中具有一定的有效性。下一步将采用更大的预训练模型作为词向量编码器, 进一步提升模型的表现力。此外, 本研究基于英文数据集进行相关实验, 考虑到中文与英文在语法结构、词汇构成和表达习惯上存在显著差异, 且缺乏明显的边界, 后续将对中文语料进行小样本命名实体识别, 从而推动小样本学习在多语言环境下的应用和发展。

参考文献

- [1] LI J, SUN A, HAN J L, et al. A survey on deep learning for named entity recognition [J]. IEEE Transactions on Knowledge and Data Engineering, 2020, 34(1): 50-70.
- [2] WEBER L, SANGER M, MUNCHMEYER J, et al. HunFlair: an easy-to-use tool for state-of-the-art biomedical named entity recognition [EB/OL]. [2024-09-01]. <https://arxiv.org/pdf/2008.07347>.
- [3] ALFRED R, LEONG L C, ON C K, et al. Malay named entity recognition based on rule-based approach [J]. International Journal of Machine Learning and Computing, 2014, 4(3): 300-306.
- [4] QUIMBAYA A P, MÚNERA A S, RIVERA R A G, et al. Named entity recognition over electronic health records through a combined dictionary-based approach [J]. Procedia Computer Science, 2016, 100: 55-61.
- [5] KONKOL M, KONOPÍK M. CRF-based Czech named entity recognizer and consolidation of Czech NER research [C] // Proceedings of the 16th International Conference on Text, Speech, Berlin, Germany: Springer, 2013: 153-160.
- [6] 张红民, 颜鼎鼎, 田钱前. 改进时空图卷积网络的视频异常检测方法 [J]. 光电工程, 2024, 51(5): 240034.
ZHANG H M, YAN D D, TIAN Q Q. Improved spatio-temporal graph convolutional networks for video anomaly detection [J]. Opto-Electronic Engineering, 2024, 51(5): 240034. (in Chinese)
- [7] LIN B Y, XU F F, LUO Z Y, et al. Multi-channel BiLSTM-CRF model for emerging named entity recognition in social media [C] // Proceedings of the 3rd Workshop on Noisy User-Generated Text. San Diego, USA: Association for Computational Linguistics, 2017: 160-165.
- [8] 陈雪松, 朱鑫海, 王浩畅. 基于 PMV-LSTM 的中文医学命名实体识别 [J]. 计算机工程与设计, 2022, 43(11): 3257-3263.
CHEN X S, ZHU X H, WANG H C. Chinese medical named entity recognition based on PMV-LSTM [J]. Computer Engineering and Design, 2022, 43(11): 3257-3263. (in Chinese)
- [9] 夏成魁, 李少波. 基于 BERT-BiLSTM-MHA-CRF 的中文命名实体识别方法 [J]. 计算机与数字工程, 2023, 51(9): 2087-2091, 2102.
XIA C K, LI S B. Chinese named entity recognition method based on BERT-BiLSTM-MHA-CRF model [J]. Computer and Digital Engineering, 2023, 51(9): 2087-2091, 2102. (in Chinese)
- [10] HUANG Z, XU W, YU K. Bidirectional LSTM-CRF models for sequence tagging [EB/OL]. [2024-09-01]. <http://de.arxiv.org/pdf/1508.01991>.
- [11] MA X Z, HOVY E. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF [EB/OL]. [2024-09-01].

- <https://arxiv.org/pdf/1603.01354>.
- [12] 陈鹏, 马洪彬, 周佳伦, 等. FCG-NNER: 一种融合字形信息的中文嵌套命名实体识别方法[J]. 重庆理工大学学报(自然科学), 2023, 37(12):222-231.
- CHEN P, MA H B, ZHOU G L, et al. FCG-NNER: a Chinese nested named entity recognition method fused with glyph information[J]. Journal of Chongqing University of Technology (Natural Science), 2023, 37(12):222-231. (in Chinese)
- [13] 许力, 李建华. 基于 BERT 和 BiLSTM-CRF 的生物医学命名实体识别[J]. 计算机工程与科学, 2021, 43(10): 1873-1879.
- XU L, LI J H. Biomedical named entity recognition based on BERT and BiLSTM-CRF[J]. Computer Engineering and Science, 2021, 43(10):1873-1879. (in Chinese)
- [14] VINYALS O, BLUNDELL C, LILLICRAP T, et al. Matching networks for one shot learning[C]//Proceedings of the 30th International Conference on Neural Information Processing Systems. New York, USA: ACM Press, 2016: 3630-3638.
- [15] SNELL J, SWERSKY K, ZEMEL R. Prototypical networks for few-shot learning [C] // Proceedings of the 31st International Conference on Neural Information Processing Systems. New York, USA: ACM Press, 2017: 4080-4090.
- [16] MA T T, JIANG H Q, WU Q H, et al. Decomposed meta-learning for few-shot named entity recognition[EB/OL]. [2024-09-01]. <https://arxiv.org/pdf/2204.05751>.
- [17] FINN C, ABBEEL P, LEVINE S. Model-agnostic meta-learning for fast adaptation of deep networks[EB/OL]. [2024-09-01]. <https://arxiv.org/pdf/1703.03400>.
- [18] WANG J N, WANG C Y, TAN C Q, et al. SpanProto: a two-stage span-based prototypical network for few-shot named entity recognition [C] // Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. San Diego, USA: Association for Computational Linguistics, 2022: 3466-3476.
- [19] LI X L, LIANG P. Prefix-Tuning: optimizing continuous prompts for generation[EB/OL]. [2024-09-01]. <https://arxiv.org/pdf/2101.00190>.
- [20] LAYEGH A, PAYBERAH A H, SOYLU A, et al. ContrastNER: contrastive-based prompt tuning for few-shot NER [C] // Proceedings of the 47th Annual Computers, Software, and Applications Conference. Washington D. C., USA: IEEE Press, 2023: 241-249.
- [21] ASHOK D, LIPTON Z C. PromptNER: prompt for named entity recognition[EB/OL]. [2024-09-01]. <https://arxiv.org/pdf/2305.15444>.
- [22] MA R T, ZHOU X, GUI T, et al. Template-free prompt tuning for few-shot NER[EB/OL]. [2024-09-01]. <https://arxiv.org/pdf/2109.13532>.
- [23] DAS S S S, KATIYAR A, PASSONNEAU R J, et al. CONTaiNER: few-shot named entity recognition via contrastive learning[EB/OL]. [2024-09-01]. <https://arxiv.org/pdf/2109.07589>.
- [24] LI Y Q, YU Y, QIAN T Y. Type-aware decomposed framework for few-shot named entity recognition[EB/OL]. [2024-09-01]. <https://arxiv.org/pdf/2302.06397>.
- [25] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[EB/OL]. [2024-09-01]. <https://arxiv.org/pdf/1810.04805>.
- [26] DOZAT T, MANNING C D. Deep biaffine attention for neural dependency parsing[EB/OL]. [2024-09-01]. <https://arxiv.org/pdf/1611.01734>.
- [27] YU J T, BOHNET B, POESIO M. Named entity recognition as dependency parsing[EB/OL]. [2024-09-01]. <https://arxiv.org/pdf/2005.07150>.
- [28] DING N, XU G W, CHEN Y L, et al. Few-NERD: a few-shot named entity recognition dataset[EB/OL]. [2024-09-01]. <https://arxiv.org/pdf/2105.07464>.
- [29] ZELDES A. The GUM corpus: creating multilayer resources in the classroom[J]. Language Resources and Evaluation, 2017, 51(3): 581-612.
- [30] PRADHAN S, MOSCHITTI A, XUE N W, et al. Towards robust linguistic analysis using OntoNotes[C]//Proceedings of the 17th Conference on Computational Natural Language Learning. San Diego, USA: Association for Computational Linguistics, 2013: 143-152.
- [31] ERIK F, SANG T K. Introduction to the CoNLL-2003 shared task: language-independent named entity recognition [C] // Proceedings of the 17th Conference on Natural Language Learning. San Diego, USA: Association for Computational Linguistics, 2003:142-147.
- [32] DERCZYNSKI L, NICHOLS E, VAN ERP M, et al. Results of the WNUT2017 shared task on novel and emerging entity recognition[C]//Proceedings of the 3rd Workshop on Noisy User-generated Text. San Diego, USA: Association for Computational Linguistics, 2017: 140-147.
- [33] YANG Y, KATIYAR A. Simple and effective few-shot named entity recognition with structured nearest neighbor learning[EB/OL]. [2024-09-01]. <https://arxiv.org/pdf/2010.02405>.
- [34] WANG P Y, XU R X, LIU T Y, et al. An enhanced span-based decomposition method for few-shot sequence labeling[EB/OL]. [2024-09-01]. <https://arxiv.org/pdf/2109.13023>.
- [35] HOU Y T, CHE W X, LAI Y K, et al. Few-shot slot tagging with collapsed dependency transfer and label-enhanced task-adaptive projection network[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. San Diego, USA: Association for Computational Linguistics, 2020: 1381-1393.

文字编辑 薛晋栋
栏目编辑 赖玉玲