MANNER：一个用于跨领域少样本命名实体识别的变分记忆增强模型

摘要

少样本命名实体识别（Few-Shot Named Entity Recognition, Few-Shot NER）在跨领域场景中面临着标注数据稀缺和领域分布差异带来的巨大挑战。本文基于 MANNER: A Variational Memory-Augmented Model for Cross Domain Few-Shot Named Entity Recognition 论文，复现了其提出的基于变分记忆增强的少样本学习框架。MANNER 模型通过引入变分记忆模块，结合少样本学习和跨领域适应技术，实现了任务相关特征的动态捕获和有效迁移，在低资源条件下显著提升了 NER 任务的性能。

本次复现从数据预处理、模型实现、训练优化到实验评估进行了系统的再现，并验证了论文中主要实验结果的可靠性。进一步地，我们针对中文数据集和特定领域数据集对模型进行实验，探讨了其在非英语环境中的适用性和局限性。复现结果表明，MANNER 模型在跨领域少样本 NER 任务中具有显著的性能提升，并为解决实际应用中的低资源问题提供了技术参考。本文的工作不仅加深了对 MANNER 模型的理解，也为未来相关研究提供了重要的借鉴意义。

关键词：命名实体识别；少样本学习；跨领域数据集；记忆增强模型

1 引言

命名实体识别是自然语言处理（Natural Language Processing, NLP）中的重要任务之一，旨在从非结构化文本中提取具有特定意义的实体，如人名、地名、组织名等。随着深度学习的快速发展，传统的 NER 方法已在许多标准领域和大规模标注数据上取得了显著进展 [5]。然而，在许多实际场景中，标注数据稀缺、标注成本高昂，尤其是在跨领域任务中，不同领域间标注数据的分布差异性更是进一步加剧了建模难度。这种背景下，研究如何在数据有限的条件下实现跨领域的少样本 NER 显得尤为重要 [4]。

MANNER 是一种旨在解决跨领域少样本 NER 问题的创新性方法 [2]。该模型通过引入变分记忆增强机制，结合少样本学习和跨领域适应能力，能够高效捕获领域间的共性特征，同时记忆特定领域的特有模式，为解决少样本 NER 提供了新的可能性。相比于传统方法，MANNER 在模型设计上考虑了更具普适性的变分推断和记忆增强机制，从理论到实践都具有显著意义。

本次研究选取 MANNER 作为复现目标，主要基于以下几方面的考虑：第一，该模型在当前跨领域少样本 NER 任务中表现出色，具备较高的学术研究价值；第二，论文中提出的

变分记忆增强模块和少样本学习框架具有较高的可拓展性，可为其他 NLP 任务提供参考；第三，深入复现并理解该模型，有助于掌握前沿算法的设计思想和实现细节，提升对少样本学习领域的研究能力。本次复现的意义不仅在于验证论文结论的可靠性，还在于通过实验进一步探索 MANNER 在实际场景中的适用性及其在中文等特定语言环境下的表现。这一过程将为未来相关研究提供数据支持和技术参考，也将为解决实际应用中的低资源 NER 问题积累经验。

2 相关工作

2.1 Few-Shot Named Entity Recognition

少样本命名实体识别（Few-Shot NER）旨在通过少量标注数据完成 NER 任务，是当前 NLP 领域的热门研究方向。许多方法基于元学习框架（如原型网络 [7]、匹配网络和基于梯度优化的 MAML [3]）来提升模型的少样本学习能力。这些方法通过模拟任务间的迁移，捕获任务间的共性特征，从而在新任务上快速适应。此外，研究者还提出了通过上下文增强的方法，如引入预训练语言模型（如 BERT [5]）以增强词语表示能力。然而，这些方法在跨领域泛化能力方面依然存在局限性，特别是在数据分布差异显著的领域中，表现仍需进一步提升。

2.2 Cross-Domain Named Entity Recognition

跨领域 NER 的挑战在于源领域和目标领域间的数据分布差异（domain shift），示例如图 ??所示。迁移学习和领域适应方法被广泛应用于这一任务 [4]，通过提取领域不变的特征或调整领域特定的特征分布来缓解数据分布差异。一些方法还利用对抗学习机制来对抗领域特定特征的影响 [6]，从而提升模型的泛化能力。然而，当目标领域标注数据稀缺时，现有方法的性能受到显著限制，如何在低资源条件下实现有效的领域迁移仍是一个挑战性问题。

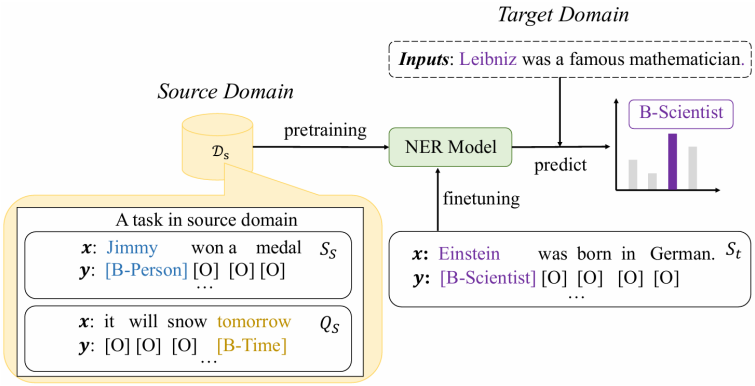

图 1. 跨领域 NER 示例

2.3 Memory-Augmented Neural Networks

记忆增强神经网络（Memory-Augmented Neural Networks, MANNs）通过外部记忆模块存储任务相关信息，能够提升模型处理复杂任务的能力 [1]。这类模型在少样本学习、强化学习等领域表现出了良好的适应性。然而，传统记忆增强模型的记忆管理和动态生成能力有限，

难以充分利用稀少的任务相关信息。在跨领域少样本 NER 任务中，设计能够灵活生成并有效利用任务相关记忆的模型是关键。

3 本文方法

3.1 本文方法概述

模型首先通过预训练语言模型为支持集和查询集的文本生成词嵌入表示，随后对于支持集的每个类，在 Memory 中检索出最相似的一个类，结合原有数据进行一个增强，得到支持集中的原型表示。查询集的词嵌入表示生成后，对其进行 span 的检测以及对检测到的 span 结合前一步得到的原型进行标注，得到最终的输出。在一次任务中，模型会现在源域数据中进行学习，随后利用少量目标领域的数据对模型进行微调，从而实现跨领域的命名实体识别。模型框架图如图 2所示：

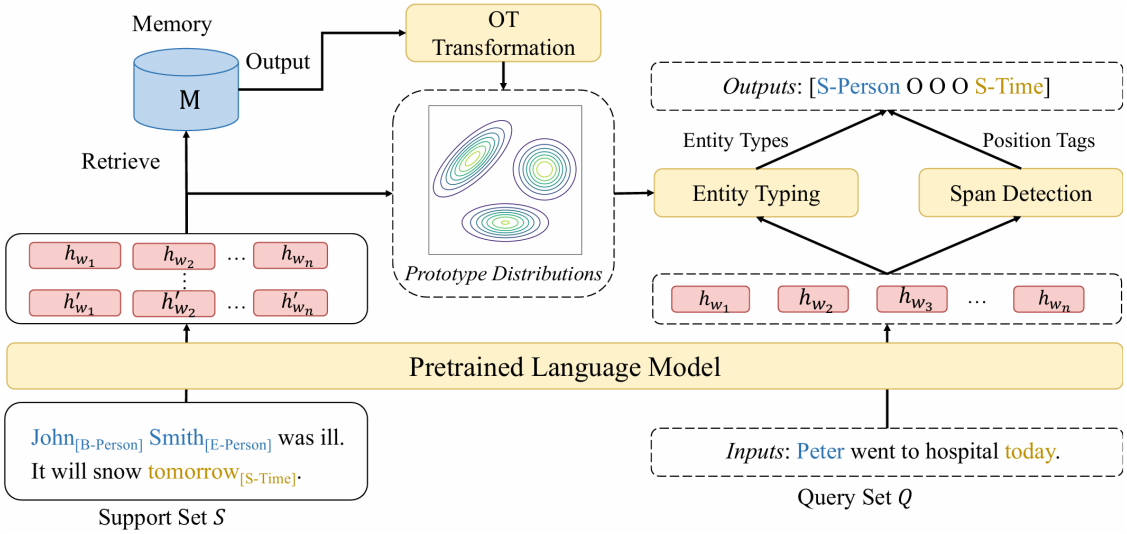

图 2. 模型框架图

3.2 记忆增强模块

在一次训练任务中，数据被划分为支持集（Support Set）、查询集（Query Set）以及验证集（Validation Set），模型在支持集中学习新的数据特征，随后用查询集来优化模型，完成一定的训练轮数后使用验证集来评估模型性能。在 Few-Shot 设置中，支持集中只有少量有标记数据，因此从支持集中获得的原型可能不够准确且不具有代表性，于是论文作者利用外部记忆模块来存储来自源域的实体类型信息，从而增强目标域中的 Few-Shot 支持集。

具体来说，作者将记忆模块表示为：M，其中包含对应源域中不同实体类型的键值对。键是不同的实体类型，值是属于响应实体类型的 token 表示。他们利用最佳传输策略在记忆模块中检索和处理信息。对于每个类，首先根据 OT 距离在记忆模块中检索出与其最相似的实体类型 k：

$$k^* = \arg\min_{k' \in \mathcal{E}} \mathcal{W}(\mathrm{M}_{k'}, \mathrm{H}_k) = \arg\min_{k' \in \mathcal{E}} \min_{\mathbf{T} \in \sum(\frac{1}{m}\mathbf{1}_m, \frac{1}{n_k}\mathbf{1}_{n_k})} \langle \mathrm{C}, \mathrm{T} \rangle \tag{1}$$

其中 $\mathbf{H}_k = f_\theta(\mathcal{S}^{(k)}), \mathcal{S}^{(k)} = \{x_{k,1}, ..., x_{k,n_k}\}$ 表示属于实体类型 k 的 token 的上下文表示，f_θ 是 token 编码器，例如 BERT，$\mathrm{M}_{k'}$ 表示存储在记忆模块中的 token 表示，C 是一个代价矩阵，用来存储检索的与被检索的 token 的距离。被检索出来的信息表示为 M_{k^*}，随后通过下列公式将检索到的信息进行投影:

$$\hat{h}_i = \arg\min_{h \in \mathbb{R}^D} \sum_j \mathbf{T}_k^*(i,j) \cdot c(h, \mathrm{H}_{k,j}) \tag{2}$$

再将其与原有支持集进行结合，得到支持集的增强表示, 此处引入变分推断来为原型分布建模，为每个实体类生成高斯分布来描述该类的特征:

$$p_\theta(\mathbf{Z} \mid \mathcal{S}, \mathrm{M}) = \prod_{k \in \mathcal{E}} p_\theta(z_k \mid \mathcal{S}^{(k)}, \mathrm{M_{k^*}}) = \prod_{k \in \mathcal{E}} \mathcal{N}(z_k \mid g_\theta(\hat{\mathbf{H}}_k, \mathbf{H}_k), \sigma_1^2 \mathbf{I}) \tag{3}$$

3.3 源域中训练

在源域训练首先要得到每个类的原型表示。首先结合记忆模块和支持集得到原型的先验分布:

$$p_\theta(\mathbf{Z} \mid \mathcal{S}, \mathrm{M}) = \prod_{k \in \mathcal{E}} p_\theta(z_k \mid \mathcal{S}^{(k)}, \mathrm{M_{k^*}}) = \prod_{k \in \mathcal{E}} \mathcal{N}(z_k \mid g_\theta(\hat{\mathbf{H}}_k, \mathbf{H}_k), \sigma_1^2 \mathbf{I}) \tag{4}$$

随后结合源域的支持集和源域的查询集得到原型的后验分布，训练时用后验分布进行分类。

$$q_\theta(\mathbf{Z}_s \mid \mathcal{S}_s, \mathcal{Q}_s) = \prod_{k \in \mathcal{E}} q_\theta(z_{s,k} \mid \mathcal{S}_s^{(k)}, \mathcal{Q}_s^{(k)}) = \prod_{k \in \mathcal{E}} \mathcal{N}(z_{s,k} \mid g_\theta(f_\theta(\mathcal{Q}_s^{(k)}, \mathcal{S}_s^{(k)})), \sigma_2^2 \mathbf{I}) \tag{5}$$

在计算损失的时候采用变分下界（Evidence Lower BOund，ELBO）作为对数似然的下界估计，损失函数定义如下，$\mathcal{D}_{KL}$ 指 KL 散度，这里用来衡量变分后验分布与模型先验分布之间的差距，值越小代表效果越好。第二项是对数似然期望，计算模型输出的参数和原型的对数似然期望，值越大表示模型对数据的解释能力越强，故 $\mathcal{L}_{ELBO}$ 越大，模型效果越好。

$$\mathcal{L}_{ELBO} = -\mathrm{D}_{KL}[q_\theta(\mathbf{Z}_s \mid \mathcal{S}_s, \mathcal{Q}_s) \parallel p_\theta(\mathbf{Z}_s \mid \mathcal{S}_s, \mathrm{M})] + \sum_{o \in \mathcal{D}_s} \mathbb{E}_{q_\theta(\mathbf{Z}_s \mid \mathcal{S}_s, \mathcal{Q}_s)}[\log p_\theta(y, a, e, \mid x, \mathbf{Z}_s)] + \text{const} \tag{6}$$

3.4 目标域微调

模型在源域数据中完成训练后，在给定的少量目标域数据的训练下进行微调，目标是实现对目标域的数据进行分类标注，整体流程与训练无异，但是在 Few-Shot 设置下，模型无法得到目标域的查询集，也就无法得到后验分布。于是用先验分布当作原型的分布，损失函数也变为如下表示，指在给定目标域支持集和记忆模块后得到的原型分布条件下，观测给定的原型到支持集的对数似然，对数里面值越大，表明模型能更好的拟合目标域数据，添加对数后由于内部项是概率，所以原有的最大化目标也变为最小化。

$$\min_\theta \mathbf{E}_{p_\theta(\mathbf{Z}_t \mid \mathcal{S}_t, \mathrm{M})}[\log p_\theta(\mathcal{S}_t \mid \mathbf{Z}_t)] \tag{7}$$

4 复现细节

4.1 与已有开源代码对比

本篇论文的代码实现有在 github 上开源，本次复现采用其提供的源代码，实现了其在跨领域数据集上的少样本命名实体识别。同时针对原文在记忆增强部分中引入的变分推断部分，尝试更改为基于正交化的变分推断，引入正交化约束，提升隐变量的解耦能力，使其能更有效地捕获独立特征。

4.2 实验环境搭建

实验运行环境所需的第三方库：
Python3.8+、PyTorch(2.4.0+cu124)、Transformers>=4.15.0、pot=0.7.0、seqeval
实验运行环境所需的内置库可根据项目源码进行安装。

4.3 使用说明

运行前需要下载跨领域数据集(here)与实体类型集(here)，然后将其放入项目中的‘data’文件夹下，随后还需下载编码器：‘bert-base-uncased’（here），并将其放入项目目录下。做好上述准备后，就可以运行下述命令运行项目，下述命令为在 Ontonotes 数据集上进行 1-shot 实验：

```
bash scripts/ontonotes_1shot.sh
```

4.4 创新点

原文中的变分推断的核心思想是假设潜在变量 z 的后验分布较难直接求解，于是使用变分分布来近似真是后验分布，并通过优化变分下界来作为训练策略。基于正交化的变分推断旨在通过引入正交化约束，使隐变量的各个维度相互独立，从而提高表示的解耦性。具体来说，就是计算损失时多考虑一些约束，在代码实现中，就是为 μ_{prior} 和 $\mu_{posterior}$ 添加正交化约束：$\mathcal{L}_{orth} = \left\|\mathrm{M}^T\mathrm{M} - \mathrm{I}\right\|_F$，其中 M 是隐变量的矩阵表示。

5 实验结果分析

论文方法 MANNER 与其他 baseline 的总体性能如表 1所示，其中数据集有 Ontonotes（通用领域），WNUT（社交媒体），GUM（采访新闻等），CoNLL（新闻），跨领域数据设置是两个数据集用作训练，一个用于验证，一个用于测试。例如评估在 Ontonotes 上的性能，将 WNUT 和 GUM 作为训练集，CoNLL 作为验证集。MANNER 在 Ontonotes 数据集上取得的 F1 分数平均在 43.61，我复现的实验使用改进后的损失计算在 Ontonotes 上的 1-shot 和 5-shot 的表现如图 3训练日志所示，取得的 F1 分数分别是 44.55% 和 58.77%，与原论文的表现相比有细微提升。

表 1. 在 Cross-Dataset 上，MANNER 和其他 baseline 的总体性能（F1 分数）其中 † 是在 [4] 中提到的，‡ 是在 [6] 中得到的。

Models	1-shot				5-shot			
	Ontonotes	WNUT	GUM	CoNLL	Ontonotes	WNUT	GUM	CoNLL
TransferBERT†	3.46 ± 0.54	2.71 ± 0.72	0.57 ± 0.32	4.75 ± 1.42	35.49 ± 7.60	11.08 ± 0.57	3.62 ± 0.57	15.36 ± 2.81
SimBERT†	13.99 ± 0.00	5.18 ± 0.00	6.91 ± 0.00	19.22 ± 0.00	21.12 ± 0.00	8.20 ± 0.00	10.63 ± 0.00	32.01 ± 0.00
Matching Network†	15.06 ± 1.61	17.23 ± 2.75	4.73 ± 0.16	19.50 ± 0.35	8.08 ± 0.47	6.61 ± 1.75	5.58 ± 0.23	19.85 ± 0.74
ProtoBERT†	6.67 ± 0.46	10.68 ± 1.40	3.89 ± 0.24	32.49 ± 2.01	13.59 ± 1.61	17.26 ± 2.65	9.54 ± 0.44	50.06 ± 1.57
CONTaiNER	32.96 ± 0.91	16.45 ± 0.92	10.81 ± 0.45	34.09 ± 0.94	48.62 ± 0.64	27.50 ± 0.58	24.31 ± 0.66	58.63 ± 1.56
L-TapNet+CDT†	15.17 ± 1.25	20.80 ± 1.06	12.04 ± 0.65	44.30 ± 3.15	20.95 ± 2.81	23.30 ± 2.80	11.65 ± 2.34	45.35 ± 2.67
DecomposedMetaNER‡	34.13 ± 0.92	25.14 ± 0.24	17.54 ± 0.98	46.09 ± 0.94	44.55 ± 0.90	31.02 ± 0.91	31.36 ± 0.91	58.18 ± 0.87
MANNER	**43.61 ± 0.48**	**28.54 ± 0.69**	**23.17 ± 0.20**	**49.06 ± 1.37**	**58.37 ± 0.62**	**35.86 ± 1.42**	**40.86 ± 0.96**	**64.84 ± 0.51**

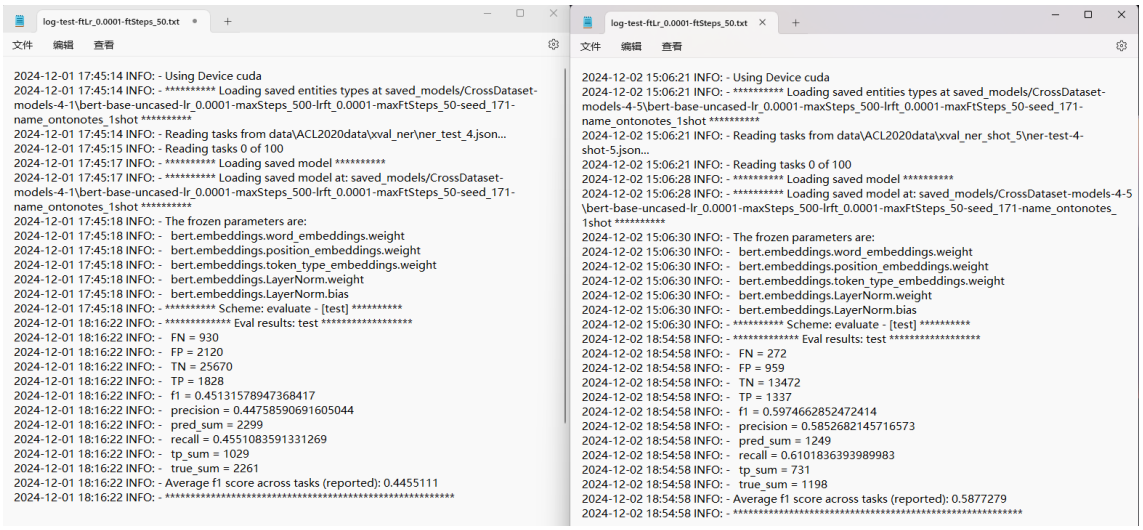

图 3. 复现结果图

6 总结与展望

在本次复现中，我成功复现了 MANNER 模型，而且在跨领域少样本命名实体识别任务数据集中验证了复现的有效性，同时针对原有的损失计算方式，加入了正交化约束，使得隐变量的各个维度相互独立，从而提高表示的解耦性，使得模型能更有效地捕获每个类的独立特征。

然而模型的性能提升并没有很明显，原因可能是正交化约束与 Few-Shot NER 任务并不适配，亦或是正交化强度并未调整到合适水平。文章也提出 MANNER 仅用外部记忆模块来增强目标域中实体类型分类模块的性能，但对 FS-NER 任务来说，更有效的提升还是可能在跨度（span）检测中，能否利用外部记忆模块来提升 span 检测中的性能，这是一个可行的研究方向。

参考文献

[1] Rajarshi Das, Manzil Zaheer, Siva Reddy, and Andrew McCallum. Question answering on knowledge bases and text using universal schema and memory networks. *arXiv preprint arXiv:1704.08384*, 2017.

[2] Jinyuan Fang, Xiaobin Wang, Zaiqiao Meng, Pengjun Xie, Fei Huang, and Yong Jiang. Manner: A variational memory-augmented model for cross domain few-shot named entity recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4261–4276, 2023.

[3] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

[4] Yutai Hou, Wanxiang Che, Yongkui Lai, Zhihan Zhou, Yijia Liu, Han Liu, and Ting Liu. Few-shot slot tagging with collapsed dependency transfer and label-enhanced task-adaptive projection network. *arXiv preprint arXiv:2006.05702*, 2020.

[5] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2. Minneapolis, Minnesota, 2019.

[6] Tingting Ma, Huiqiang Jiang, Qianhui Wu, Tiejun Zhao, and Chin-Yew Lin. Decomposed meta-learning for few-shot named entity recognition. *arXiv preprint arXiv:2204.05751*, 2022.

[7] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.