基于语义特征增强的中文小样本命名实体识别

司贺杰[1] 徐恕贞[1] 张嘉殷[1]
SI Hejie XU Shuzhen ZHANG Jiayin

摘 要 中文小样本命名实体识别资源有限，现有识别方法大多仅依赖模型内部的注意力机制，缺乏对标签语义信息的有效利用。在这一情况下，当命名实体与周围词语存在语义模糊时，模型难以精确界定命名实体的边界，降低了识别的准确性。为此，文章设计了一种基于语义特征增强的中文小样本命名实体识别方法。通过引入外部语义知识和预训练语言模型，对中文小样本中的命名实体进行深度语义挖掘，并生成语义特征增强后的实体表示。融入标签语义信息，对经过增强后的命名实体进行跨度识别，增强实体与非实体之间的注意力差异，提升整个命名实体识别的准确性。通过对比实验证明：该方法不仅提高了识别的准确性，还具备更高的识别普适性，为中文小样本命名实体识别提供了新的解决方案。

关键词 语义特征增强；小样本；实体识别；命名；中文

doi：10.3969/j.issn.1672-9528.2025.05.006

0 引言

命名实体识别（named entities recognition, NER）是自然语言处理（natural language processing, NLP）中的一个基础任务，其核心目标是从文本中识别出具有特殊意义或有指代性的词语，如人名、地名、组织机构名等。命名实体识别不仅在信息抽取、信息检索、机器翻译和问答系统等多种自然语言处理技术中扮演着至关重要的角色，还在智能客服、医疗电子病历识别、法律文书名词识别以及金融领域名词识别等多个领域有广泛的应用。然而，尽管其重要性不言而喻，命名实体识别仍然面临诸多挑战，特别是在处理中文文本时。

近年来，研究者们提出了一系列针对命名实体识别的有效方法。王昕岩等人[1]探讨了基于大语言模型的命名实体识别方法。该方法预训练阶段通常是基于大规模语料开展无监督学习。在此过程中，主要是对语言通用模式进行学习，而未能有效地融入标签语义信息。当命名实体和周围词语存在语义模糊的情况时，模型就难以精准判断哪些词语属于命名实体范畴，进而造成命名实体边界界定失准，最终导致识别错误。项恒等人[2]则采用双向长短期记忆网络（BiLSTM）与条件随机场（CRF）相结合的模型，针对航行通告这一特定领域的命名实体识别问题进行了深入研究。通过 BiLSTM 层对文本序列进行上下文信息的全局特征提取，并通过 CRF 层对输出序列进行标注规则的学习，从而实现了对命名实体的有效识别。该方法 CRF 层主要负责对输出序列进行标注规则的学习，这种学习依据已有的数据特征开展，并非基于标签语义信息本身。当命名实体与周围词语存在语义模糊时，该方法难以精确地界定命名实体的边界，也就无法准确识别出真正的命名实体。

为了克服上述不足，本文提出了一种基于语义特征增强的中文小样本命名实体识别方法。

1 基于语义特征增强的中文小样本增强后实体表示

在中文小样本命名实体识别的场景中，由于数据资源有限，小样本数据难以全面覆盖所有命名实体，导致模型学习不充分，难以准确捕捉词语间复杂的语义关系，容易出现误判[3]。为解决这一问题，本文提出了一种基于语义特征增强的方法。该方法通过引入外部语义知识和预训练语言模型进行深度语义挖掘，并生成语义特征增强后的实体表示，显著提升中文小样本命名实体识别的准确性和泛化能力。这不仅有助于改善模型在有限数据条件下的性能，还为后续的自然语言处理任务提供了更加可靠的基础。

设文本序列为：

$$T = \{w_1, w_2, \cdots, w_n\} \tag{1}$$

式中：w_i 表示第 i 个词语。

通过词嵌入模型，为每个词语生成对应的词表示：

$$\boldsymbol{E} = \{\boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_n\} \tag{2}$$

式中：$\boldsymbol{e}_i$ 为 w_i 的词向量。

为进一步增强实体表示，引入丰富的语义特征，这些特征包括词义关联、句法依赖以及针对特定领域的知识。这

1. 郑州科技学院信息工程学院 河南郑州 450064

些特征可以通过外部知识库和句法分析工具来获取[4]。利用WordNet等词典，能够为每个词语找到其同义词、反义词和上位词等，构建出词义关联网络。设$S(w_i)$表示w_i的同义词集合，则词义关联特征可以表示为：

$$\boldsymbol{f}_{s,i}=\frac{1}{S(w_i)}\sum_{s\in S(w_i)}\boldsymbol{e}_s \tag{3}$$

式中：$\boldsymbol{f}_{s,i}$表示由同义词集合构成的词义关联特征。

通过句法分析工具，可以解析出文本中的句法依赖关系。设$D(w_i)$表示与w_i有直接依赖关系的词语集合，则句法依赖特征可以表示为：

$$\boldsymbol{f}_{D,i}=\frac{1}{D(w_i)}\sum_{d\in D(w_i)}\boldsymbol{e}_d \tag{4}$$

对于特定领域的数据集Resume，构建了领域词典，并据此提取领域特定知识特征[5]。设$G(w_i)$表示与w_i相关的领域知识集合（如专业术语、职位名称等），则领域特定知识特征可以表示为：

$$\boldsymbol{f}_{G,i}=\frac{1}{G(w_i)}\sum_{d\in G(w_i)}\boldsymbol{e}'_d \tag{5}$$

式中：$\boldsymbol{e}'_d$表示领域知识d的词向量，可通过领域特定的词嵌入模型获得。

将基础特征和语义特征进行结合，得到增强后的实体表示：

$$\boldsymbol{r}_i=\boldsymbol{e}_i+\lambda_1\boldsymbol{f}_{s,i}+\lambda_2\boldsymbol{f}_{D,i}+\lambda_3\boldsymbol{f}_{G,i} \tag{6}$$

式中：λ_1、λ_2、λ_3表示权重系数，用于量化不同特征对实体表示的影响程度。通过上述设计，将语义特征引入了中文小样本命名实体识别的实体表示中，从而增强了模型的识别能力。

2 融入标签语义的命名实体跨度识别

虽然引入外部语义知识和预训练语言模型增强了实体表示，能识别出实体，但仅识别实体本身是不够的。完整的命名实体识别还需要准确确定实体的边界（跨度）。然而，不同类型的命名实体（如人名、地名、机构名等）在语义和结构上有各自的特点，传统方法仅依赖模型内部的注意力机制，缺乏对标签语义信息的有效利用，难以准确判断实体跨度。为此，本研究提出融入标签语义信息辅助命名实体跨度识别，以提升识别准确性[7]。标签语义信息为命名实体类型提供了额外的上下文，这些信息有助于模型更深入地理解文本中的实体，更精确地界定边界。融入标签语义信息后，模型能够更细致地分析文本特征，减少误判。这种方法不仅能有效识别实体的存在，还能精准确定其实体和结束位置，使命名实体识别的结果更加精准可靠，为自然语言处理中的信息抽取、关系抽取等相关任务提供更精确的数据支撑[8]。

每个命名实体类型（如人名、地名、机构名等）都蕴含着丰富的语义信息，这些信息对于准确识别实体及其边界至关重要。为充分利用这些语义信息，本文为每个命名实体类型的标签构建了相应的语义表示[9]。这一构建过程可以通过两种主要途径实现：首先，利用预训练的词嵌入模型，将标签名称转换为高维的词向量表示，这些词向量能够捕捉到标签名称的语义特征；其次，通过查阅外部知识库（例如WordNet），获取标签的同义词、上位词等相关信息，并基于这些信息构建标签的语义网络，从而进一步丰富标签的语义表示[10]。设标签集合为$L=\{l_1,l_2,\cdots,l_m\}$，其中，l_j表示第j个标签。

在命名实体识别的任务中，不仅需要确定实体是否存在，还需要确定其跨度。为了实现这一目标，本研究提出融入标签语义信息。在传统的序列标注模型基础上，融入标签语义信息来辅助识别实体的跨度。设文本序列为$T=\{w_1,w_2,\cdots,w_n\}$，经过模型处理后，得到每个位置i上对应每个标签l_j的得分$s_{i,j}$。为了融入标签语义，对这些得分进行了修改，引入了标签语义表示的加权项，计算公式为：

$$s'_{i,j}=s_{i,j}+r_i\cdot l_j \tag{7}$$

式中：l_j表示标签语义。

在得到修改后的得分$s'_{i,j}$后，使用$s'_{i,j}$进行标签解码，以确定文本序列中每个位置的标签。通过这一步骤，可以准确地识别出命名实体的起始和结束位置，从而确定其实体的跨度。最后，根据识别出的跨度信息，将文本中的实体和非实体进行区分，增强实体与非实体之间的注意力差异，完成命名实体识别的任务。

这种方法不仅能够有效识别出实体的存在，还能精确确定其实体的边界，从而使命名实体识别的结果更加精确可靠。这一改进为自然语言处理中的信息抽取、关系抽取等相关任务提供了更加准确的数据支撑，有助于提升这些任务的性能和效果。

3 对比实验

3.1 实验数据

为验证本文提出的基于语义特征增强的识别方法在中文小样本命名实体识别任务中的可行性，选择MSRA、Weibo、Resume、Medicine和OntoNotes五个中文数据集作为实验数据集。在所选择的实验数据集中，MSRA为新闻数据集，含组织、人物、地点3类实体；Weibo为微博数据集，含组织、人物、地点、政治4类实体，分命名实体（NE）和泛指实体（NM）；Resume为经济数据集，注释了8类命名实体；Medicine为医学领域数据集，含非实体及治疗方式、身体部位、疾病症状3类实体；OntoNotes为新闻多语言语料库，包含多种文本注释和中文命名实体标签。各数据集的

实体数量及分布如表 1 所示，其中训练集和测试集的规模均较小，符合小样本识别任务的特点。

表 1 各数据集实体数量记录表

序号	类型	训练 /kB	测试 /kB
（1）	MSRA	18.23	1.73
（2）	Weibo	15.24	3.12
（3）	Resume	25.63	3.91
（4）	Medicine	16.15	3.06
（5）	OntoNotes	12.36	2.15

实验参数设置方面，考虑到小样本的特点，采用了编码层 12 层、注意力头数 12、字符隐层大小 768、实体隐层 64 的配置，并使用 Adam 优化器进行训练。实验环境为 Win10 系统，i9-9900 CPU，64 GB 内存，使用 Python 编程。

3.2 实体识别效果对比

为体现实验的客观性和可对比性，选择两种对照方法：一是基于大语言模型的识别方法（对照 I 组）；二是基于 BiLSTM-CRF 的识别方法（对照 II 组）。以 MSRA 数据集为例，选取其中一句话：“微软在北京发布新项目，由比尔·盖茨主导。”作为测试样本，其中“微软（组织）”“比尔·盖茨（人物）”“北京（地点）”为实体，其余部分为非实体。通过对比不同识别方法在相同测试样本上对实体部分和非实体部分的注意力分布情况，来评估并证明不同方法的优劣。3 种识别方法的识别结果如图 1~3 所示。

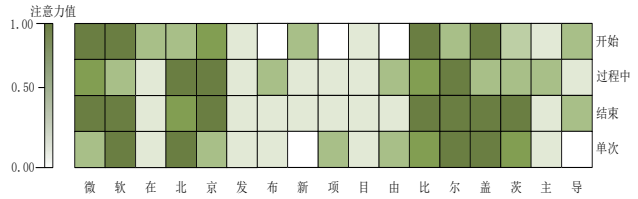

图 1 本文方法识别结果

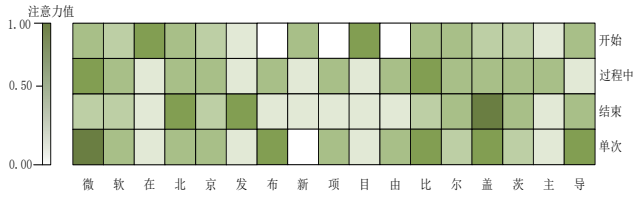

图 2 对照 I 组识别结果

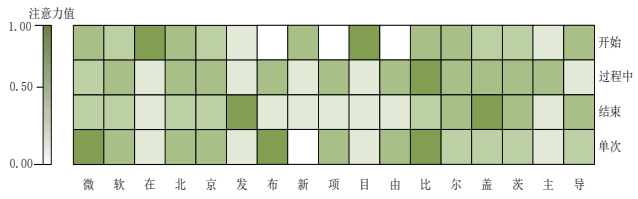

图 3 对照 II 组识别结果

从图 1~3 所示识别结果可以看出，本文方法识别结果中，“微软”“比尔·盖茨”“北京”等实体的注意力分布较为集中，非实体部分的注意力分布较为分散，证明了本文方法的有效性。相比之下，对照 I 组和对照 II 组识别方法识别结果中，实体部分与非实体部分的注意力分布差异不如本文方法显著。通过上述实验结果可以证明，本文提出的基于语义特征增强的识别方法在中文小样本命名实体识别任务中表现出色，能够减轻对非实体部分的关注，将更多的注意力值分配给实体部分，从而提高实体识别的效果。

3.3 识别普适性对比

为进一步验证本文识别方法在不同数据集上的识别性能，选取了 MSRA、Weibo、Resume、Medicine 和 OntoNotes 这五个数据集进行中文小样本命名实体识别实验。MSRA 数据集属于新闻领域，实体类型（人名、地名、机构名）标注规范，适合通用场景。Weibo 数据集属于社交媒体文本，噪声多、实体边界模糊，考验模型抗干扰能力。Resume 数据集属于简历领域，实体类型（教育背景、公司名等）结构化，适合垂直场景。Medicine 数据集属于医学文献，专业术语密集，测试领域自适应能力。OntoNotes 数据集属于多领域混合，实体类型丰富（如法律、金融），验证跨领域迁移能力。随机抽取以上 5 个数据集中共 11 572 698 条数据，构成本次测试使用的数据集。

在实验中，采用 F_1 值作为评价指标，该指标综合了精确率（Precision）和召回率（Recall）两个重要性能指标，能够全面评估识别性能。F_1 值的计算公式为：

$$F_1 = 2 \times \frac{P+R}{PR} \tag{8}$$

式中：P 表示精确率；R 表示召回率。

其中，精确率表示模型预测为正类的样本中，实际为正类的比例。召回率表示实际为正类的样本中，被模型正确预测为正类的比例。其计算公式为：

$$P = \frac{T_P}{T_P + F_P} \tag{9}$$

$$R = \frac{T_P}{T_P + F_N} \tag{10}$$

式中：T_P 表示实际为正类且被预测为正类的样本数；F_P 表示实际为负类但被预测为正类的样本数（误报）；F_N 表示实际为正类但被预测为负类的样本数（漏报）。

通过精准率与召回率的联合分析，可更全面地诊断本文设计的方法在真实场景中的表现，避免因单一指标误导决策。

将 3 种识别方法识别结果的 F_1 值记录，如图 4~6 所示。对比图 4~6 所示的结果可以看出，本文识别方法在每个数据集上的 F1 值均表现出最优效果，而对照 I 组和对照 II 组的

F_1 值则普遍偏低。特别是对照 II 组在 Resume 数据集上的 F_1 值明显低于其他四种数据集，这进一步凸显了本文方法的优势。这一优势得益于该方法引入的语义特征增强机制，使其能够捕捉到更深层次的语言结构，如词义关联、句法依赖以及领域特定的知识，从而在不同数据集上均实现了 F_1 值的显著提升。

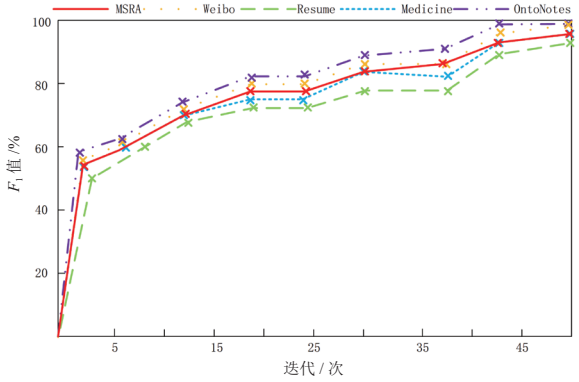

图 4 本文方法识别结果 F_1 值曲线图

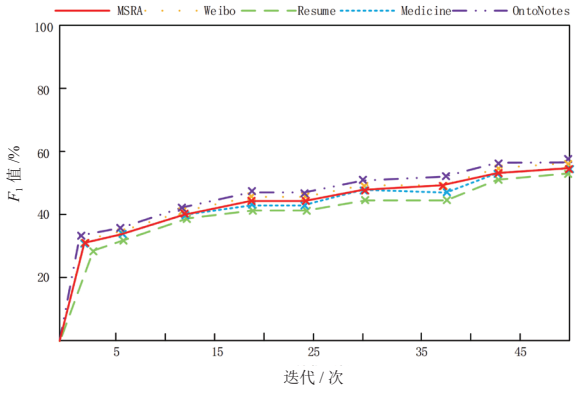

图 5 对照 I 组识别结果 F_1 值曲线图

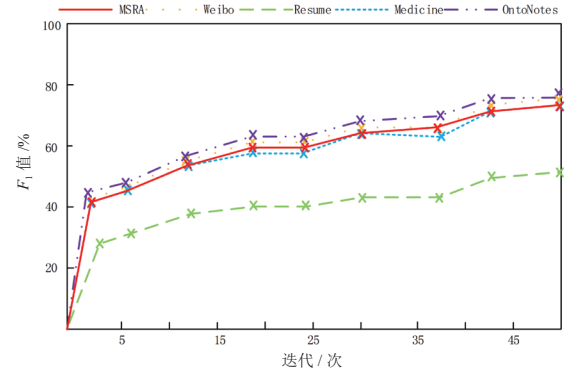

图 6 对照 II 组识别结果 F_1 值曲线图

4 结语

本文提出了一种基于语义特征增强的中文小样本命名实体识别方法，该方法的核心在于克服现有方法在面对小样本数据时性能受限的挑战。实验验证显示，即便在训练资源有限的情况下，该方法仍能维持高水准的识别准确率，并且广泛适用于识别多种类型的命名实体。尽管本文的方法在中文小样本命名实体识别领域取得了显著成果，但仍面临一些挑战。特别是在处理含有复杂语义结构和高度歧义的命名实体时，模型的识别精确度尚需提升。此外，研究更高效的模型架构以减少计算成本，是后续研究的关键方向。

参考文献：

[1] 王昕岩，陈建，邓曦．基于大语言模型的命名实体识别方法研究 [J]. 通信与信息技术，2024(6): 109-112.

[2] 项恒，杨明友，李猛．基于 BiLSTM-CRF 的航行通告命名实体识别研究 [J]. 计算机科学，2024, 51(S2): 125-130.

[3] 曹倩雯，李维，林伯韬，等．融合旋转位置编码与掩码条件随机场的钻井工程命名实体智能识别方法 [J]. 石油科学通报，2024, 9 (5): 750-763.

[4] 田泽庶，刘春雨，张云婷，等．基于软提示微调和强化学习的网络安全命名实体识别方法研究 [J]. 通信学报，2024, 45 (10): 1-16.

[5] 廖天正，林晓兰，陈永辉，等．基于 BiLSTM-CRF 的命名实体识别在临床电子病历的研究与应用 [J]. 电脑知识与技术，2024, 20 (28): 17-19.

[6] 刘洋，唐海，朱梦涵，等．基于视频资源与 WoBERT-AT-BiLSTM-CRF 的命名实体识别方法 [J]. 智能计算机与应用，2024, 14 (10): 63-69.

[7] 刘彤，石昌岭，倪维健．面向生物医学命名实体识别和规范化的多粒度特征融合 [J]. 计算机系统应用，2024, 33 (11): 237-246.

[8] 刘悦，姚文轩，刘大晖，等．高质量文本数据驱动的命名实体识别加速镍基单晶高温合金材料知识发现 [J]. 金属学报，2024, 60 (10): 1429-1438.

[9] 翟姗姗，余华娟，陈健瑶，等．基于多维特征分析的戏曲类方志文献命名实体识别研究 [J]. 情报学报，2024, 43 (9): 1094-1104.

[10] 陈述，张超，陈云，等．基于命名实体识别的水电工程施工安全规范实体识别模型 [J]. 中国安全科学学报，2024, 34 (9): 19-26.

【作者简介】

司贺杰（1993—），女，河南驻马店人，硕士，助教，研究方向：知识图谱、高等教育。

徐恕贞（1990—），女，河南南阳人，硕士，助教，研究方向：数据挖掘。

张嘉殷（1997—），女，河南永城人，硕士，助教，研究方向：教学教改。

（收稿日期：2024-12-09 修回日期：2025-05-09）