一种混合提示学习与规则的领域命名实体识别方法

张　晗[1,2]，　张亚洲[2]，　徐秉智[2]，　张铖方[1]

(1. 四川警察学院 智能警务四川省重点实验室　四川 泸州 646099；
2. 郑州大学 网络空间安全学院　河南 郑州 450002)

摘要：基于提示的微调学习为改善针对特定领域的命名实体识别(named entity recognition，NER)任务的性能提供了一个新的研究方向，但现有的提示学习方法面临需要人工构造模板、提示信息冗长、提示模板固定等问题。针对以上问题，提出了一种结合提示学习与专家知识的领域命名实体识别方法。首先，通过引入 Bootstrapping 算法自动识别潜在的实体，并改进了在获取相同上下文未标注实体类别过程中字符串匹配算法以获取更多提示信息模板。其次，引入领域本体中的专家知识来解决提示信息的可靠性问题。同时，采用一阶谓词的形式表示提示信息来改善提示信息长度。最后，通过在金融与信息安全两个数据集上的实验，验证了该方法能够有效提高领域命名实体识别的性能。

关键词：提示学习；命名实体识别；自然语言处理；低资源
中图分类号：TP183　　**文献标志码**：A　　**文章编号**：1671-6841(2025)05-0031-08
DOI：10.13705/j.issn.1671-6841.2024040

A Hybrid Approach of Prompt-based Learning and Rules for Domain Specific Named Entity Recognition

ZHANG Han[1,2], ZHANG Yazhou[2], XU Bingzhi[2], ZHANG Chengfang[1]

(1. *Intelligent Policing Key Laboratory of Sichuan Province*, *Sichuan Police College*, *Luzhou* 646099, *China*;
2. *School of Cyber Science and Engineering*, *Zhengzhou University*, *Zhengzhou* 450002, *China*)

Abstract: Prompt-based fine-tuning was a new direction to improve the performance of domain specific named entity recognition (NER). However, the existing methods faced challenges such as the need of manual template construction, lengthy prompt information, and fixed prompt templates. To address these issues, a method combined prompt learning with expert knowledge was proposed in the field of domain specific named entity recognition. Firstly, by introducing the bootstrapping algorithm, potential entities were automatically identified. And the string matching algorithm used in the process of obtaining unannotated entity types from the same context was improved to obtain more prompt information templates. Next, expert knowledge from the domain ontology was introduced to address the reliability concerns associated with prompt information. Simultaneously, first-order predicate logic was used to represent prompt information and to improve the representation of prompt information. Finally, with experiments on finance dataset and information security dataset, the method was verified to improve the performance of domain specific named entity recognition effectively.

Key words: prompt based learning; named entity recognition; natural language processing; low resource

收稿日期：2024-03-04
基金项目：智能警务四川省重点实验室开放课题(ZNJW2024KFQN005)；河南省高等学校重点科研项目(24A520047)；河南省重大科技专项(231100210200)
第一作者：张晗(1985—)，女，博士研究生，主要从事自然语言处理研究，E-mail：zhang_han@zzu.edu.cn。
通信作者：张铖方(1990—)，男，硕士研究生，主要从事智能信息处理研究，E-mail：chengfangzhang@scpolicec.edu.cn。

0 引言

命名实体识别(NER)旨在从文本中提取各种类型的实体,其结果可用于其他复杂任务诸如关系提取[1]、领域知识图谱的构建[2-3]等。与通用领域的命名实体识别任务相比,特殊领域的命名实体识别经常面临着两方面的问题:领域标注数据缺乏;领域中的实体形式更加复杂,并非局限于传统 NER 定义的名词或者名词短语。

目前,基于提示的调优学习已经成为自然语言处理领域的新范式[4]。基于提示的调优学习可以通过改造下游任务和增加专家知识,使任务的输入和输出适合原始语言模型,从而在少样本场景中获得良好的效果。但是,目前基于提示的调优学习多应用于文本分类或文本生成领域[5-7],在命名实体识别领域的应用较少。

通过对文献[8-12]进行分析后发现,目前已经发表的基于提示学习的 NER 方法具有如下缺陷:1)需要人工构造提示信息模板[9],因此需要耗费大量的人力且容易出错;2)需要对序列中的每一个单词都构造提示信息[9-10],当输入序列较长时,会增加序列的长度,增加模型的计算复杂度;3)提示信息模板较为固定[11-12],在面对复杂类型实体时表现较差。

事实上,构造提示信息模板是影响提示学习方法性能的关键因素[13]。因此,本文将更加关注提示信息的自动构造问题以及提示信息的可靠性问题。目前已有不少研究验证了专家知识的加入对提升模型在数据集上的可靠性有很大帮助[14-15]。鉴于以上思想,本文提出了一种结合提示学习与专家知识的 NER 方法,该方法的核心部分即提示信息构造模块,在这个模块中引入了 Bootstrapping 算法[16]来解决提示模式的固定问题,然后又引入领域本体中的专家知识来解决提示信息的可靠性问题。同时,本文还在该模块中引入了谓词逻辑理论,相较于采用自然语言序列作为提示信息的方法[9,11-12],谓词逻辑的序列更短,更易于被模型理解。此外,为了提高方法的运行效率,本文提出了一种基于 AC 自动机的高效文本匹配方法,以缩减使用 Bootstrapping 算法的时间。本文的主要贡献如下。

1)提出了一种基于提示学习与规则的领域命名实体识别方法,用于解决领域中的标注数据稀缺问题,该方法在低资源场景的信息安全和金融领域数据集上都表现出了最优的性能,这些改进主要来自可靠而多样的提示信息构造。

2)提出了一种提示信息自动构造方法,该方法首先通过 Bootstrapping 算法来获取实体的上下文模式,然后通过引入本体中的专家知识增加提示信息的可靠性,并采用一阶谓词公式的形式来表示提示信息,这样既解决了由于提示信息的加入所造成的输入序列过长问题,又易于对模型的理解。

3)提出了一种实体上下文模式在非结构化文本中的快速匹配方法,在 AC_BM 算法的基础上进行改进,将 Double-array trie 引入 AC_BM 算法,解决了 AC_BM 算法在中文领域的空间浪费及内存溢出问题。

1 相关工作

标注数据缺乏一直是领域命名实体识别所面临的一个大问题。最近,一种资源较少的下游任务方式——提示学习[17]引起了人们的兴趣。2021 年,Cui 等[9]提出采用 cloze prompt 解决小样本下的 NER 任务。对于一个句子,如果某个词组是实体,那么其对应的模板就是$<x_i:j>$ is a $<y_k>$;如果某个词组不是实体,那么其对应的模板为$<x_i:j>$ is not an entity。这种模板构造方法过于依赖人工经验,并且效果也难以保障。

为了解决上述问题,文献[10]利用同种类型实体相互预测以及构造空间映射,将 NER 任务的提示学习构造为无须模板的语言模型问题。Shen 等[8]构造了双槽多提示模板,将实体的定位与实体类型统一到一轮中,通过并行的方法提取所有可能的实体。利用二分图实现模板与实体的自动对应,自动形成提示模板。He 等[18]利用外部知识对实体类型对语义初始化并结合对比损失,将无须模板的提示嵌入句子表达中,也降低了模板构造与枚举过程的资源消耗。然而,由于领域数据集的稀缺,文献[8]中同类型实体的集合很难构建。Lee 等[11]在低资源场景下,通过在数据文本后提供额外的信息描述形成基于演示的学习,利用面向实体与面向实例的拼接提示模板来代替手工模板。有的研究从不同语言与不同任务的特点出发分别提出了跨语言与跨任务的基于提示学习的 NER 方法。Wang 等[19]利用不同语言作为监督与度量指标,实现了当时最先进的少样本跨语言 NER 任务效果。为了涵盖信息抽取过程中不同的模式特点,Lu 等[12]通过提示学习实现了语义知识共享以及任务迁移。与上述方法不同的是,文献[12]构造的查询模板中包含了类别信

息,训练的模型从文本序列中定位实体并补全模板。

2　方法

2.1　问题定义

假设有少量标注数据集 D,$x_i \in D$,有未标记文本集合 T,此时,通过将 D 中的 x_i 对 T 进行逐一匹配,获取一个候选规则集 r,$r_i = \{prefix_i, x_i, postfix_i\}$,将规则集 r 套用到未标记文本上,从而可以获取与 x_i 相同前后缀的实体 z_j,若 x_i 标签为 y_i,则 z_j 标签也为 y_i,会导致识别的实体有限,且会出现准确率高但召回率低的问题。因此,需要进一步加入领域本体作为辅助。从本体 O 中获取所有包含 y_i 类别关系的集合 rel,$rel_{i,j} = \{y_i, relation, y_j\}$,$y_i$ 为实体 x_i 对应的类别标签,y_j 表示本体 O 中其他类别标签,与 x_i 所在的文本进行相似度比较,选择其中置信度最高的规则 $rel_{i,j}$ 作为提示信息。将其与原文本一同作为预训练语言模型的输入,获取未标记实体 x_j 的标签 y_j。

2.2　模型介绍

在本文所提出的方法中,提示信息模板构造是其核心部分。与之前提出的方法不同的是,本文为了保证提示信息的多样性和可靠性,采取了半监督方法 Bootstrapping 算法与专家知识相结合的方式,方法具体流程如图 1 所示。由图 1 可以看出,本文的模型主要包括了三个核心部分。

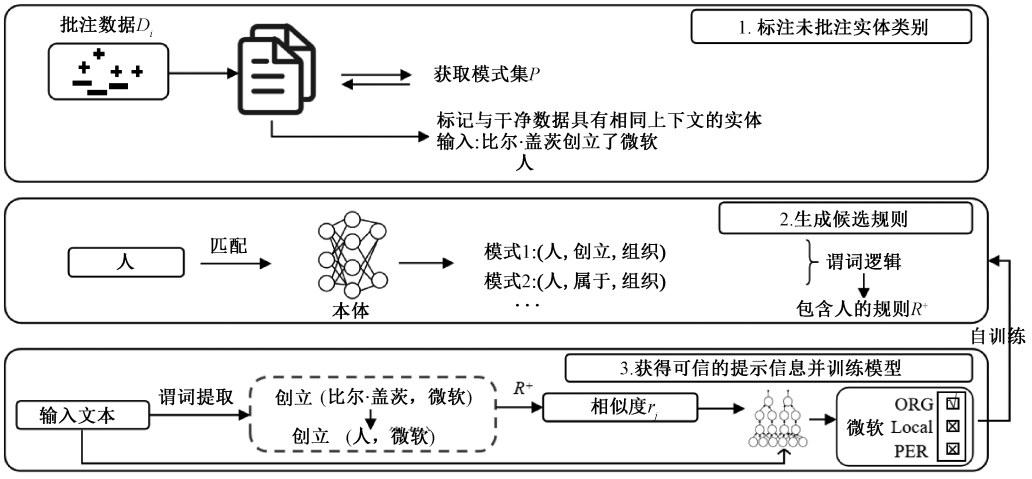

图 1　模型架构图

Figure 1　Model architecture diagram

1) 相同上下文未标注实体类别获取:该部分主要为本体中的可信任规则筛选提供匹配信息。它利用少样本标注数据集 D_i 获取规则模式集合 R,然后通过快速匹配算法,对句子 s_i 中的具有相同前后缀的未标记实例 x_j 标记类别 y_j。

2) 本体中的规则筛选:该部分主要用于筛选本体中的规则,将规则作为提示信息来标注句子 s_i 中不能通过规则模式匹配的实体类别。它将前一个组件提出的 y_j 作为输入,将本体中所有包含了 y_j 及其关系的三元组提取出来,组成集合 Reg_j。

3) 可信提示信息获取与模型训练:该部分主要用于从 Reg_j 中获取可信的规则作为提示信息。本文将引入数理逻辑中的概念,将 Reg_j 中所有的规则与输入句子进行对比,然后提取其中可信度最高的规则 Reg_{ji} 作为提示信息,与 s_i 一同交给预训练模型。然后将模型分类出的结果返回给第 2) 部分,通过自监督训练模型。

2.2.1　相同上下文未标注实体类别获取　在这一阶段,本文通过 Bootstrapping 算法将与标注数据集具有相同上下文的未标注实体标注出来。这部分的核心工作是字符串的快速匹配算法。传统的字符串匹配算法有 KMP 算法和 BM 算法,而本文的工作更类似于多模式查询匹配,AC 自动机算法更适用于本文的工作。文献[20]中采用了效率更高的 BM 算法代替了原来 AC 自动机算法中的 KMP 算法,更加提升了查询的速度。但是文献[20]主要针对的是英文模式匹配,这意味着构建英文词典的前缀树,只需要维护一个 26 叉树。但中文字符集的规模远远超过英文,由于前缀树的每个节点都需要创建一个与字符集等长的数组来快速访问子节点,直接将文献[20]用于中文模式匹配会造成空间的浪费并导致内存溢出问题。而 Double-array 结合了 array 查询效率高、list 节省空间的优点。因此,本文采用 Double-array trie(DATrie)与 BM 算法相结合的方法,构建一种高效的中文多模式查询算法(DAAC_BM 算法)。该算法过程如下所示。

步骤1：根据字典 N 生成前缀树 *trie*。将 *trie* 的 *root* 节点存入队列 *queue*，并初始化 *base* 和 *check* 数组。获得每个节点的好前缀偏移 *gsShift* 与坏字符偏移 *bcShift*。

步骤2：当 *queue* 不为空时，从 *queue* 中出队一个节点 *currentNode*，寻找其在 *base* 中对应的值 *baseVal*，并将 *currentNode* 的所有子节点加入队列。然后获得 *currentNode* 的父节点在 *base* 中的下标 *checkVal*，并将 *baseVal* 和 *checkVal* 存入 *base* 和 *check* 数组。

步骤3：将输入序列 X 分割为 n 值不同的 n-gram 子串组成的词语组集合 $\{phrases^1, phrases^2, \cdots, phrases^n\}$，对词语组 $phrases^k$ 中包含的词语 $phrases_i^k$ 进行匹配。

步骤4：$phrases_i^k$ 为字符集合 $\{c_i^1, c_i^2, \cdots, c_i^k\}$。初始化下标 j，通过 *base* 数组和 *check* 数组在 N 中查找模式 $\{c_i^j, c_i^{j+1}, \cdots\}$。若完整匹配 N 中一个词语，则 $phrases_i^k$ 匹配成功并跳转步骤6。

步骤5：根据步骤4中匹配结果，取 *gsShift* 和 *bcShift* 中较大值作为 j 的位移量，返回步骤4并重新计算 j。直到 j 小于0，则 $phrases_i^k$ 匹配完成。

步骤6：返回步骤3，直到 $\{phrases^1, phrases^2, \cdots, phrases^n\}$ 中 $phrases_i^k$ 全部完成匹配，并返回所有匹配成功的词语。

下面，本文对 DAAC_BM 算法与 KMP 算法、BM 算法以及传统的 AC_BM 算法分别比较时间复杂度和空间复杂度。

假设存在 k 个待匹配模式 P_k，m 为模式 P_k 中单个模式的长度，目标文本 T 长度为 n，字典的大小为 D。KMP 算法的每个模式匹配的时间复杂度为 $O(n)$，计算 *next* 数组为 $O(m)$。在多模式匹配情况下，总的时间复杂度为 $O(k*n)$（排除预处理时间）。BM 算法每个模式匹配的时间复杂度在 $O(n/m)$ 和 $O(m*n)$ 之间，则多模式下在 $O(k*n/m)$ 和 $O(k*m*n)$ 之间。改进后的算法可以分为在 T 上的移动与在 AC 树中查找节点两部分。在目标文本上匹配和移动时，最佳情况下的时间复杂度为 $O(n/\min(m))$，最差情况下为 $O(n*\max(m))$，平均为 $O(n/m)$。在 AC 树中查找时，如果每个节点只存储其后继子节点，则每个节点查找下一个节点的时间复杂度取决于查找算法，如二分查找为 $O(\log D)$，则总体时间平均复杂度为 $O((n*\log D)/m)$。如果每个节点存储的是字典长度，则查找下一个节点的时间复杂度为 $O(1)$。这与传统的 AC_BM 算法查找效率相同，总体时间复杂度为 $O(n/m)$。

在空间复杂度上，传统的 AC_ BM 算法的空间复杂度在 $O(D*\max(m))$ 和 $O(D*avg(m)*k)$ 之间。而改进后的算法主要包含两个数组，每个数组的大小在 $O(\max(m))$ 和 $O(avg(m)*k)$ 之间。因此，DATrie 的空间复杂度在 $O(\max(m)+D)$ 和 $O(avg(m)*k+D)$ 之间。相比之下，改进后的算法在保持与传统算法相同查找效率的同时解决了传统 AC_BM 算法中的空间消耗过大及内存溢出问题。

2.2.2 本体中的规则筛选 该部分主要对本体中的规则进行筛选，将实体类别 y_j 作为输入，从本体中提取出所有包含了 y_j 及其关系的三元组，组成集合 Reg_j。例如：

影子经纪人利用永恒蓝和双脉冲星这两个漏洞进入受害者的机器。

在这句话中，本文先通过 Bootstrapping 算法将实体“影子经纪人”标记类别为“威胁主体”。然后采用 KMP 算法从网络安全领域本体 UCO 中获取所有含有“威胁主体”的三元组。因为缺乏比较完善的中文网络安全领域本体，所以，本文采用了最简单的翻译方式。如图2所示，这些三元组构成了“threat agent”的关系集合 Reg_{attack}。

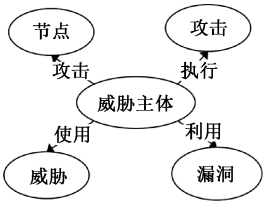

图2 UCO 中所有含有“威胁主体”的三元组

Figure 2 All triples containing "threat agent" in UCO

2.2.3 可信提示信息获取 接下来，本文将要从 Reg_{attack} 中获取可信的规则作为提示信息。由于 Reg_{attack} 中都是以三元组形式构建的集合，本文将其转换为一阶谓词逻辑公式集合 $Predict_{attack}$。同样，也要把句子转换成一阶谓词逻辑的形式。一个句子的核心成分为主语、谓语和宾语。谓语类似于谓词逻辑公式中的谓词，主语和宾语就相当于谓词逻辑中的项。因此，第一步，本文先对输入文本进行词性标注，提取句子中所有动词。第二步，对句子进行语法依存关系分析，提取名词主语。最后，本文再提取动词的直接宾语。以“影子经纪人利用永恒蓝和双脉冲星这两个漏洞进入受害者的机器”这句话为例。首先提取两个动词“利用”和“进入”；其次，提取句子中的名词主语“影子经纪人”；最后提取“利

用”和“进入”的直接宾语,分别是“漏洞”和“机器”。因此,可以构建出该句子的谓词公式分别是“利用(影子经纪人,漏洞)”和“进入(影子经纪人,机器)”。在 2.2.1 的工作中,已经确定了影子经纪人的类型为“威胁主体”,因此,谓词公式可转换为“利用(威胁主体,漏洞)”“进入(威胁主体,机器)”。

本文将从句子中得到的谓词逻辑公式集合 $Predict_{\text{sentence}}$ 与 $Predict_{\text{attack}}$ 进行比较,获取相似度比较高的谓词公式,本文采用的是余弦相似度比较方法。如果 $Predict^{i}_{\text{sentence}}$ 与 $Predict^{j}_{\text{attack}}$ 的相似度大于阈值 λ,则 $Predict^{j}_{\text{attack}}$ 被选为可信赖的提示规则,然后将其转变为普通句子的形式,与输入句子一同输入预训练语言模型中。

2.2.4 模型训练　借鉴同类相吸引异类相排斥的观点,本文认为模型要达到好的训练效果应当使得同类别距离最小,不同类别距离最大。

有实体集合 E,其中,$E_B \in E$ 表示通过 Bootstrapping 算法标注的实体集合,$E_M \in E$ 表示通过模型标注的实体集合。由于 Bootstrapping 算法具有高准确率,因此,本文将 E_B 中属于同类别的实体向量求和取均值作为该类别中心,设 $e_i \in E_M$,y_i 为预测标签,$\hat{y}_i$ 为类别 y_i 的中心,Y 为所有类别集合,则目标函数 L_{target} 可表示为

$$L_{\text{target}} = \min\left\{ distance(e_i, \hat{y}_i) + \frac{1}{\sum_{y_j \in Y, j \neq i}^{Y} distance(e_i, \hat{y}_j)} \right\}。 \tag{1}$$

3　实验与分析

本文设计了三个实验来验证本文方法的优越性。首先,比较本文方法与其他先进的命名实体识别方法在数据集上的表现;其次,利用不同批注样本数量的数据集训练模型,验证本文方法在低资源场景数据集上的适应性;最后,比较了 AC_BM 算法与 DAAC_BM 算法在关键字集上的空间消耗。

3.1　实验设置与基线模型

实验使用了金融与网络安全两个领域数据集。金融数据集(Dataset Ⅰ)包含了从多个金融网站的新闻、财报、文章摘要等信息中提取的 7 521 个句子,批注了人名(PER)、产品名(PRO)、公司名(COM)、地名(LOC)、组织(ORG)和时间(TIME)6 种实体类型。网络安全领域数据(Dataset Ⅱ)主要来自 Freebuf 网站和乌云漏洞数据库,共 13 428 条数据,包含了 PER、LOC、ORG、软件名(SW)、网络术语(RT)和漏洞编号(VUL)6 类实体类型。

实验中的 5 个基准模型分别如下。

1) BERT-tagger:在 BERT 模型的基础上利用标签分类器进行微调的方法[21]。

2) BERT-BiLSTM-CRF:一种利用 BERT 模型实现上下文关联,结合双向长短时记忆网络进行微调的方法[22]。

3) BERT-RDCNN-CRF:该方法针对网络安全领域的数据集,在 BERT 模型的基础上增加了基于残差空洞卷积网络与 CRF 层[23]。

4) PWII-BERT:一种将提示引导词汇信息集成到 BERT 模型的中文命名实体识别方法[24]。

5) TemplateNER:一种低资源场景下无模板的提示学习方法[10]。

本实验中使用的预训练模型为 BERT-base-Chinese 模型,训练优化器为 Adam,学习率为 1×10^{-5}。模型输入的最大句子长度为 256,批大小设置为 32。实验的硬件环境为 64 GB 显存的 NVIDIA GTX 3090 显卡、64 GB 内存。

在验证集上取得最佳 $F1$ 值的前提下,将本文中在 Dataset Ⅰ和 Dataset Ⅱ上使用的阈值 λ 分别设定为 0.64 与 0.58。本文所提方法在处理信息安全领域数据集时用的本体为 UCO,处理金融领域数据集时用到的本体为 FTHO。

本实验中采用常用的精确率(*precision*, P)、召回率(*recall*, R)、$F1$ 分数作为评价标准。计算每个类别的 P、R、$F1$,累加后取平均值作为总体数据集的评估值。

3.2　结果分析

3.2.1 与基线模型的对比实验　通过多轮训练,每种方法在两个数据集上实验时最佳性能结果如表 1 所示。观察表 1 可以发现,本文方法在两个数据集上的表现总体超过基线模型。在引入了句法与表层特征提取模型后,BERT-BiLSTM-CRF 和 BERT-RDCNN-CRF 都可以较大提高模型的识别效果。同时,PWII-BERT、TemplateNER 以及本文方法这些基于提示学习的微调方法,总体表现更好,特别是在召回率上。这可能是由于提示学习与预训练任务形式更相似,在低资源场景下具有较好的泛化能力。本文方法在去除本体知识后,在两个数据集上的 P、R 和 $F1$ 都明显下降,验证了本体知识作为提示信息的有效性。本文方法在 R 与 $F1$ 两个指标上高于其他方法,这得益于本体知识的

加入为模型提供了更全面、准确的提示信息。而在 Dataset Ⅰ的 P 指标上,PWII-BERT 模型略高于本文所提出的方法,这可能是由于文献[24]将词级特征注入提示信息中。总体来说,本文的方法在金融与网络信息安全的两个领域数据集上,取得了最好的综合效果。

表 1 各方法在金融与网络安全领域数据集上的表现性能

Table 1 Performance evaluation of methods on datasets in the finance and cybersecurity domains

方法	Dataset Ⅰ			Dataset Ⅱ		
	P	R	$F1$	P	R	$F1$
BERT-tagger	0.749 1	0.740 1	0.740 1	0.820 3	0.807 1	0.810 6
BERT-BiLSTM-CRF	0.782 9	0.803 4	0.795 1	0.847 1	0.890 2	0.873 9
BERT-RDCNN-CRF	0.775 2	0.788 7	0.783 2	0.856 9	0.880 7	0.871 3
PWII-BERT	**0.801 9**	0.833 4	0.817 4	0.883 6	0.898 7	0.890 5
TemplateNER	0.796 7	0.802 5	0.798 8	0.845 7	0.891 4	0.862 1
本文方法(-O)	0.750 1	0.747 6	0.748 1	0.824 5	0.852 7	0.841 9
本文方法	0.799 1	**0.845 1**	**0.820 5**	**0.884 1**	**0.902 7**	**0.892 3**

注:-O 为本文方法在去除本体知识后的模型;黑体为最优值。

3.2.2 *在低资源情况下的性能实验* 为了验证在少样本情况下的模型效果,本文随机选择比例为10%、20%、30%、50%、100%的训练数据集对模型训练。不同模型的 $F1$ 性能结果如表 2 所示。由表 2 可以得出,随着训练数据集的减少,BERT-tagger 与 BERT-BiLSTM-CRF 模型的性能会有明显下降。在训练数据集较多时,PWII-BERT 与本文方法效果相近,但当可用数据降低到 20%与 10%时,本文的方法表现更好。TemplateNER、PWII-BERT 方法由于提示学习方法的特点,$F1$ 下降趋势并不太明显。总体来说,本文方法总体保持最高的 $F1$ 值,尤其是在20%与 10%的情况下。

表 2 不同比例训练集下各模型的 $F1$ 值

Table 2 $F1$ scores of different models with different training set ratios

模型	Dataset Ⅰ					Dataset Ⅱ				
	100%	50%	30%	20%	10%	100%	50%	30%	20%	10%
BERT-tagger	74.01	71.98	70.21	67.71	62.42	81.06	79.68	78.51	74.28	69.42
BERT-BiLSTM-CRF	79.51	77.72	76.39	72.22	66.11	87.39	84.49	82.41	78.63	73.11
BERT-RDCNN-CRF	78.32	76.41	74.39	73.12	70.41	87.13	85.63	83.39	80.12	76.41
PWII-BERT	81.74	80.32	78.57	76.32	71.48	89.05	**87.42**	**86.27**	82.32	77.80
TemplateNER	79.88	78.40	77.08	75.92	73.39	86.21	84.96	83.08	81.02	76.59
本文方法	**82.05**	**80.74**	**79.03**	**78.01**	**75.14**	**89.23**	87.23	85.93	**83.41**	**78.62**

注:黑体为最优值。

3.2.3 *DAAC_BM 算法与原始 AC_ BM 算法的空间对比实验* 该实验比较了 AC_BM 算法和 DAAC_BM 算法使用 6 个中文关键字集(set of keywords, SK)构造的查找树与双数组在空间上的表现。通过 Python 实现 AC_BM 算法与改进的算法,并分别统计 4 个关键字集形成的查找树与双数组占用的内存空间信息。关键字集统计信息与算法内存占用结果见表 3。

从表 3 可以发现,DAAC_BM 算法在所有关键字集上测试的内存占用低于 AC_BM 算法,验证了 DAAC_BM 算法的有效性。特别是当关键字集合较大时,改进后 AC_BM 算法占用的内存变化并不大,这是因为双数组的紧凑结构可以避免不必要的空白节点造成的空间资源浪费。在可接受的预处理时间消耗范围内,DAAC_BM 算法结合了 AC_BM 算法与 Double-array trie 的优点,在内存占用方面表现更为高效。

表 3　关键字集统计信息与各算法占用内存

Table 3　Statistical information of keywords set and the memory occupancy with different algorithms

关键字集	包含内容	关键字数量/个	平均长度/字符	AC_BM 算法占用内存/MB	DAAC_BM 算法占用内存/MB
SK1	国家及地区名	240	6.1	248.62	1.328
SK2	世界城市名	3 795	3.5	1 659.32	5.296
SK3	财经词语	3 830	4.1	2 177.23	5.238
SK4	国内外人名	6 992	2.6	2 067.26	6.359
SK5	Dataset Ⅰ中过滤出的关键字集	4 162	3.7	1 969.43	5.317
SK6	Dataset Ⅱ中过滤出的关键字集	7 214	3.4	2 631.67	6.867

4　结语

本文提出了一种结合提示学习与专家知识的领域命名实体识别方法。引入 Bootstrapping 算法自动识别潜在的实体,并改进了相同上下文未标注实体类别获取过程中字符串匹配的算法以获取更多提示信息模板。引入领域本体中的专家知识来解决提示信息的可靠性问题,采用一阶谓词公式的形式表示提示信息来改善提示信息长度。实验表明,本文方法与其他基准模型相比,在金融与信息安全领域命名实体识别任务上能够达到更加优越的性能。考虑中文分词的特点,未来工作需要进一步研究如何有效利用词级信息,以提升模型在中文数据上的识别效果。

参考文献:

[1] YANG L Y, YUAN L F, CUI L Y, et al. FactMix: using a few labeled in-domain examples to generalize to cross-domain named entity recognition[EB/OL]. (2022-08-24)[2023-12-16]. https://doi.org/10.48550/arXiv.2208.11464.

[2] 李明键,李卫军,王海荣.融合关联信息与 CNN 的实体识别研究[J].郑州大学学报(理学版),2023,55(5):53-59.
LI M J, LI W J, WANG H R. Fusion of association information and entity recogni-tion of CNN[J]. Journal of Zhengzhou university(natural science edition), 2023, 55(5): 53-59.

[3] ZHAO Y C, ZHANG B K, GAO D. Construction of petrochemical knowledge graph based on deep learning[J]. Journal of loss prevention in the process industries, 2022, 76: 104736.

[4] HAN X, ZHANG Z Y, DING N, et al. Pre-trained models: past, present and future[J]. AI open, 2021, 2: 225-250.

[5] DING N, CHEN Y L, HAN X, et al. Prompt-learning for fine-grained entity typing[C]//Findings of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2022: 6888-6901.

[6] SCHICK T, SCHÜTZE H. It's not just size that matters: small language models are also few-shot learners[EB/OL]. (2020-09-15)[2023-12-16]. https://doi.org/10.48550/arXiv.2009.07118.

[7] WU Y Q, LIU Y F, LU W M, et al. Towards interactivity and interpretability: a rationale-based legal judgment prediction framework[C]//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2022: 4787-4799.

[8] SHEN Y, TAN Z, WU S, et al. Prompt-NER: prompt locating and typing for named entity recognition[EB/OL]. (2023-05-26)[2023-12-16]. https://doi.org/10.48550/arXiv.2305.17104.

[9] CUI L Y, WU Y, LIU J, et al. Template-based named entity recognition using BART[C]//Findings of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2021: 1835-1845.

[10] MA R T, ZHOU X, GUI T, et al. Template-free prompt tuning for few-shot NER[C]//Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2022: 5721-5732.

[11] LEE D H, KADAKIA A, TAN K M, et al. Good examples make a faster learner: simple demonstration-based learning for low-resource NER[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2022: 2687-2700.

[12] LU Y J, LIU Q, DAI D, et al. Unified structure generation for universal information extraction[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2022: 5755-5772.

[13] LING T T, CHEN L, SHENG H X, et al. Sentence-level event detection without triggers via prompt learning and machine reading comprehension[EB/OL]. (2023-06-25)[2023-12-16]. http://arxiv.org/abs/2306.14176.

[14] CHEN H N, LUO X W. An automatic literature knowledge graph and reasoning network modeling framework based on ontology and natural language processing[J]. Advanced engineering informatics, 2019, 42: 100959.

[15] BIERMANN D, GOODWIN M, GRANMO O C. Knowledge infused representations through combination of expert knowledge and original input[C]//Symposium of the Norwegian AI Society. Cham: Springer International Publishing, 2022: 3-15.

[16] TEIXEIRA J, SARMENTO L, OLIVEIRA E. A bootstrapping approach for training a NER with conditional random fields[C]//Portuguese Conference on Artificial Intelligence. Berlin: Springer Press, 2011: 664-678.

[17] SCHICK T, SCHÜTZE H. Exploiting cloze-questions for few-shot text classification and natural language inference[C]//Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2021: 255-269.

[18] HE K, MAO R, HUANG Y C, et al. Template-free prompting for few-shot named entity recognition via semantic-enhanced contrastive learning[J]. IEEE transactions on neural networks and learning systems, 2023, 99: 1-13.

[19] WANG Y G, HUANG Y C, GONG T L, et al. Enhancing cross-lingual few-shot named entity recognition by prompt-guiding[C]//International Conference on Artificial Neural Networks. Cham: International Springer Publishing, 2023: 159-170.

[20] COMMENTZ-WALTER B. A string matching algorithm fast on the average[M]//Automata, Languages and Programming. Berlin: Springer Press, 1979: 118-132.

[21] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[EB/OL]. (2019-05-24)[2023-12-16]. http://arxiv.org/abs/1810.04805.

[22] HU X S, ZHANG H J, HU S L. Chinese named entity recognition based on BERTbased-BiLSTM-CRF model[C]//IEEE/ACIS 22nd International Conference on Computer and Information Science. Piscataway: IEEE Press, 2022: 100-104.

[23] 谢博. 基于深度学习的中文网络威胁情报信息抽取技术研究[D]. 贵阳: 贵州大学, 2022.
XIE B. Research on information extraction technology of chinese cyber threat intelligence based on deep learning[D]. Guiyang: Guizhou University, 2022.

[24] HE Q, CHEN G W, SONG W C, et al. Prompt-based word-level information injection BERT for Chinese named entity recognition[J]. Applied sciences, 2023, 13(5): 3331.