计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(9):272−279 [doi: 10.15888/j.cnki.csa.008750]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于主题提示的电力命名实体识别①

康雨萌, 何　玮, 翟千惠, 程雅梦, 俞　阳

(国网江苏营销服务中心, 南京 210019)
通信作者: 何　玮, E-mail: 562691978@qq.com

摘　要: 传统的命名实体识别方法可以凭借充足的监督数据实现较好的识别效果. 而在针对电力文本的命名实体识别中, 由于对专业知识的依赖, 往往很难获取足够的监督数据, 即存在少样本场景. 同时, 由于电力行业的精确性要求, 相比于一般的开放领域任务, 电力领域的实体类型更多, 因此难度更大. 针对这些挑战, 本文提出了一个基于主题提示的命名实体识别方法. 该方法将每个实体类型视为一个主题, 并使用主题模型从训练语料中获取与类型相关的主题词. 通过枚举实体跨度、实体类型、主题词以填充模板并构建提示句. 使用生成式预训练语言模型对提示句排序, 最终识别出实体与对应类型标签. 实验结果表明, 在中文电力命名实体识别数据集上, 相比于几种传统命名实体方法, 基于主题提示的方法取得了更好的效果.

关键词: 命名实体识别; 预训练模型; 提示模板; 主题模型; 电力语料

引用格式: 康雨萌,何玮,翟千惠,程雅梦,俞阳.基于主题提示的电力命名实体识别.计算机系统应用,2022,31(9):272–279. http://www.c-s-a.org.cn/1003-3254/8750.html

Electric Power Named Entity Recognition Based on Topic Prompt

KANG Yu-Meng, HE Wei, ZHAI Qian-Hui, CHENG Ya-Meng, YU Yang

(State Grid Jiangsu Marketing Service Center, Nanjing 210019, China)

Abstract: Traditional named entity recognition methods can achieve favorable results owing to sufficient supervision data. As far as named entity recognition from electric power texts is concerned, however, the dependence on professional knowledge often makes it difficult to obtain sufficient supervision data, which is also known as a few-shot scenario. In addition, electric power named entity recognition is more challenging than general open domain tasks due to the accuracy requirements of the electric power industry and the more categories of entities in this industry. To overcome these challenges, this study proposes a named entity recognition method based on topic prompts. This method regards each entity category as a topic and uses the topic model to obtain topic words related to the category from the training corpus. Then, it fills in the template and constructs prompt sentences by enumerating entity spans, entity categories, and topic terms. Finally, the generative pre-trained language model is used to rank the prompt sentences and ultimately identify the entity and the corresponding category label. The experimental results show that on the dataset of Chinese electric power named entities to be recognized, the proposed method achieves better results than those offered by several traditional named entity recognition methods.

Key words: named entity recognition; pre-trained model; prompt template; topic model; electric power corpus

随着人工智能技术的迅速发展, 国网集团启动了“互联网+电力营销”的工作模式, 将传统的线下营业厅与人工客服热线升级为自动化的电力客服机器人. 为了支撑智能化的客服问答, 构建知识图谱成为了一个

① 基金项目: 国网江苏省电力有限公司科技项目 (J2021151)
收稿时间: 2021-12-16; 修改时间: 2022-01-13, 2022-01-28; 采用时间: 2022-03-03; csa 在线出版时间: 2022-06-28

主要途径. 而在整个构建流程中, 如何从电力领域文本中进行命名实体识别 (named entity recognition, NER)[1] 是一个重要环节, 它旨在将输入文本中的单词或短语识别为不同类型的实体标签[2], 为后续关系抽取等步骤提供基础.

传统的 NER 方法主要是基于 BiLSTM-CRF 框架. 鉴于预训练语言模型 (pre-trained language model, PLM)[3] 在多项自然语言处理任务上带来的显著提升, 微调 PLM 的参数以编码输入文本, 并利用 Softmax 或条件随机场 (conditional random field, CRF)[4] 分配实体标签, 成为了 NER 领域的普遍做法. 尽管这类方法在一般任务上表现不俗, 但是由于预训练和下游 NER 任务之间存在差距, 且对于新的目标领域, 模型需要足够的训练实例进行微调, 因此在电力场景下, NER[5] 任务仍然面临着以下挑战:

首先, 现有方法大多假定具有充足的标注训练数据, 然而, 提供电力领域的标注往往需要具备领域知识的专业人员. 这使得在实际应用中训练数据不足, 即存在少样本 (few-shot) 问题. 其次, 在传统开放领域 NER 数据集中, 实体类型一般较少且更含义宽泛, 如在广泛使用的英文数据集 CoNLL03[2] 中, 只有 4 种实体类型. 而在中文电力场景中, 由于其行业特殊性, 实体类型高达 14 种, 而且训练数据更少, 这无疑加大了预测实体类型的难度.

为了克服上述挑战, 本文提出了一种基于主题提示的 NER 模型 (topic prompt NER model, TP-NER). 该模型打破了 BERT-LSTM-CRF 范式, 使用自然语言提示模板挖掘 PLM 的潜在知识, 以提升少样本 NER 的效果. 同时, 该模型利用了电力语料中的主题信息, 使得实体类型预测更加准确.

1 相关工作

近年来, 基于神经网络的方法在 NER 任务中提供了有竞争力的表现. Lewis 等人[5] 和 Chiu 等人[6] 将 NER 视为对输入文本的每个单词的分类问题. Ma 等人[4] 利用 CRF 和“序列-到-序列”框架[7], 从而得到实体跨度与对应类型标签. Zhang 等人[8], Cui 等人[9] 和 Gui 等人[10] 分别使用标签注意网络和贝叶斯神经网络. 随着预训练模型的兴起, Yamada 等人[11] 提出了基于实体感知的预训练, 从而 NER 上获得不错的效果. 这些方法与本文方法的区别是它们是为指定的命名实体类型[12–14] 设计的, 采用了序列标注的框架, 这令它们在少样本场景难以适应新的类型.

目前已经有一些关于少样本场景下 NER 的研究. Wiseman 等人[15] 提出了不同的预训练方法和微调策略. Yang 等人[16] 利用常见的少样本分类方法, 如原型网络和匹配网络, 其中还学习了提高性能的转换分数. 这些方法依赖复杂的训练过程, 但结果并不显著. Chen 等人[17] 的方法不需要元训练, 通过最近邻分类器和结构化解码器, 取得了更好效果.

利用外部知识来提高 PLM 的性能近年来得到了广泛的研究, 通常应用于预训练和微调阶段. 具体来说, 在文本分类任务中, Li 等人[18] 探索了利用知识图谱来增强输入文本. 与这些方法不同, 本文的方法在提示调优结合了主题知识, 因此在少样本 NER 任务中产生了显著的改进.

自从 GPT-3 出现以来, 提示调优受到了相当大的关注. GPT-3 表明, 通过即时调整和上下文学习, 大规模语言模型可以在低资源情况下表现良好. Schick 等人[19] 认为小规模语言模型也可以使用提示调整获得不错的性能. 虽然大多数研究都是针对文本分类任务进行的, 但一些工作将提示调整的影响扩展到其他任务, 例如关系抽取. 除了对各种下游任务使用提示调优, 提示模板还用于从 PLM 中探查知识. 因此, 这为 NER 任务提供了一种前景, 即通过运用提示模板, 模型可能有效利用预训练带来的知识.

2 基于主题提示的电力 NER 模型

2.1 任务定义

给定一条输入电力文本$X=\{x^1,x^2,\cdots,x^n\}$, 其中x^i表示文本中的第i个字, T^+为文本总字数. 命名实体识别任务的目标是输出三元组$Y=(u_s,u_e,l)$, 其中$u_s\in[1,n]$和$u_e\in[u_s,n]$分别表示识别出的实体在X中的起始索引与结束索引, $l\in L$表示实体的类型标签, L为数据集中所有类型的集合. 如果输入文本X中不包含实体, 则输出 (−1, −1, −1). 如下展示了两个电力场景中关于 NER 任务的例子, 其中例 1 中的输入文本包含“业务需求”类型实体“复电”, 例 2 中的输入文本没有包含任何实体.

例 1. 输入文本: 复电手续如何申请?

输出: $(u_s=1,u_e=2,l=\text{business})$

解释: (1,2)表示实体跨度“复电”, business 表示标签“业务需求”.

例 2. 输入文本: 这是怎么回事?

输出: $(u_s=-1,u_e=-1,l=-1)$

解释: 该输入文本中无实体.

2.2 基于提示调优的 NER 框架

PLM 模型蕴含了从海量语料中学习到的丰富知识. 利用这些涵盖各个领域的知识即可在仅有少量训练样本的情况下对电力领域完成快速适配. 在传统 NER 常用的 BERT+LSTM+CRF 模型[20] 中, 尽管预训练的 BERT 被用于编码输入文本, 但最终还是需要通过微调 (fine-tuning) 其参数以适应 NER 任务. 由于预训练的目标 (掩码预测) 与 NER 微调的目标 (序列标注) 不一致, 因此知识无法被有效利用, 使得基于微调的模型在电力 NER 上通常无法取得较好的结果.

区别于这些微调模型, 本文提出的 TP-NER 构建了一种基于提示调优 (prompt-tuning) 的框架, 以解决的电力场景的少样本问题. 简单来说, TP-NER 将 NER 的输出包装成自然语言提示模板. 相比于原有的三元组形式, PLM 更适合对自然语言进行语义表示和打分, 这是因为它原本就在自然语言语料上进行预训练. 这种提示模板统一了预训练任务与下游 NER 任务的形式, 使得 PLM 中的知识可以被直接利用. 这样, 仅使用少量的训练样本即可完成对电力领域的适配.

整个方法流程概览如图 1 所示. 在离线阶段, 预先构建 NER 自然语言模板; 在推理阶段, 首先通过枚举候选跨度填充模板, 生成候选提示句, 再利用 PLM 对候选提示句直接打分排序. 得分最高的提示句所对应的实体与类型作为输出被返回.

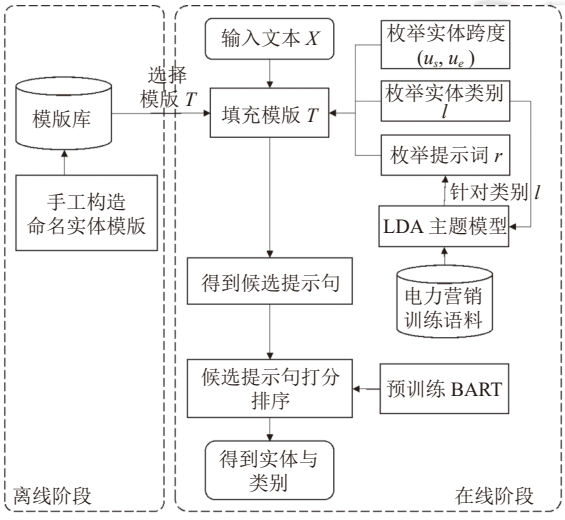

图 1 基于主题提示的 NER 方法流程图

2.2.1 NER 提示模板构建

在本文的定义中, NER 提示模板是一个包含空槽位的自然语言句子. 例如, “[MASK-e] 是一个 [MASK-t] 类型的实体”是一个模板. 其中, [MASK-e] 表示识别出的实体跨度, 如“电能表”; [MASK-t] 表示实体 [MASK-e] 的类型, 如“机器设备”. 这种模板以自然语言的形式对候选的实体与类型进行了重新包装, 以便 PLM 模型可以利用在自然语言语料上学习到的先验知识克服少样本问题.

如引言所提到的, 电力领域中实体类型较多, 包含 14 种, 如“业务需求”“机器设备”. 在少样本场景下, PLM 模型缺少足够的训练数据去理解这些细粒度实体类型的差别. 因此, 对上述 NER 提示模板进行实体类型方面的增强. 具体地, 模板被扩充为“[MASK-e] 是一个 [MASK-t] 类型的实体, 与 [MASK-r] 相关”. 其中, [MASK-r] 表示与实体类型 [MASK-t] 语义关联的提示词. 这些词与实体类型密切相关, 在预训练的语料中往往与对应的类型共同出现, 因此对 PLM 可以起到有效的提示作用, 从而进一步帮助它理解实体类型的语义.

在离线阶段, 为了涵盖不同的自然语言表达方式, 设计了 3 种正样本模板T^+与 1 个负样本模板T^-, 如表 1 所示. T^+表示句子中存在实体, 而T^-表示句子中无实体. 这样, 模板既能利用 [MASK-t] 带有的全局类型信息, 也能利用与 [MASK-r] 获得局部信息.

表 1 命名实体模板

模板类型	模板
正样本	[MASK-e]是一个[MASK-t]类型的实体, 与[MASK-r]相关 [MASK-e]的实体类型是[MASK-t], 与[MASK-r]相关 [MASK-e]属于[MASK-t]类型, 与[MASK-r]相关
负样本	[MASK-e]不是一个实体

2.2.2 模板填充与候选提示句生成

在推理阶段, 首先从正样本模板中随机选择一个模板T^+, 如“[MASK-e] 是一个 [MASK-t] 类型的实体, 与 [MASK-r] 相关”, 作为待填充的模板. 接着, 枚举命名实体跨度(u_s,u_e). 具体做法是, 对于任意一个索引$u\in[1,n]$, 枚举长度从 1 到 m 之间的所有跨度, 即$(u,u),(u,u+1),\cdots,(u,u+k)$. 对跨度$(u_s,u_e)$, 将$X_{u_s:u_e}$填入$T^+$. 如跨度$(1,3)$, 即文本“复电手”, 填入模板后得到提示句“复电手是一个 [MASK-t] 类型的实体, 与 [MASK-r]相关”. 随后, 枚举一个实体类型标签$l\in L$, 如“business”, 将其对应的标签词“业务需求”填入到T^+的 [MASK-t] 中, 模板更新为

“复电手是一个业务需求类型的实体, 与 [MASK-r] 相关”. 最后, 利用 LDA 模型从电力训练语料中获取b个提示词, 记为$R=\{r_1,r_2,\cdots,r_m\}$. 这些提示词与实体类型在语义上密切相关, 其获取过程将在第 2.3 节中详细阐述. 枚举每个提示词$r_i\in R$, 如“申请”, 填入到 [MASK-r] 中, 最终得到完整的提示句“复电手是一个业务需求类型的实体, 与申请相关”, 记为$T_{u_s,u_e,l}$. 对输入文本X完成所有枚举后, 一共得到$n\times m\times b\times|L|$个提示句. 这里, n为X的长度, $|L|$表示类型标签集合大小. 此外, 考虑文本X中无实体的情况, 此时仅枚举实体跨度填充负样本模板T^-, 而不需要枚举实体类型, 得到$n\times m$个负样本提示句. 综上, 一共得到$n\times m\times(b\times|L|+1)$个候选提示句.

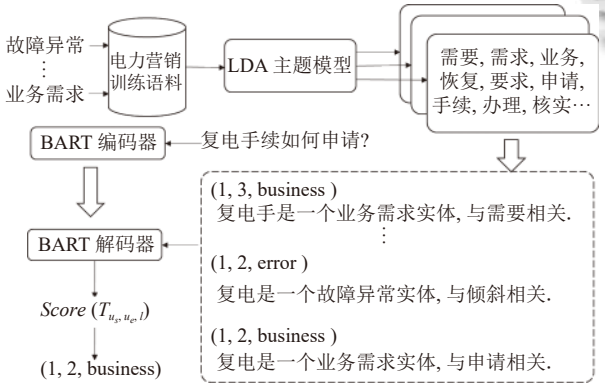

图 2 TP-NER 框架

2.2.3 候选提示句打分排序

此阶段的目标是计算每个候选提示句的分数. 为了克服电力领域的少样本问题, 使用生成式 PLM 模型 BART[5], 以其蕴含的丰富知识弥补训练样本的缺失. BART 是一种基于编码器-解码器框架的 PLM 模型, 集成了 BERT 双向编码和 GPT 自左向右解码的特点, 这使得它比 BERT 更适合文本生成的场景. 在本文中, 将文本X输入到 BART 编码器中, 通过自注意力得到的上下文表示. 接着使用 BART 解码器进行自回归解码, 在每个解码时刻得到单词的输出概率.

具体来说, 对输入文本$X=\{x^1,x^2,\cdots,x^n\}$, 设候选提示句$T_{u_s,u_e,l}=\{t^1,t^2,\cdots,t^{n'}\}$, 其中$t^i$表示该句的第$i$个字, n'表示文本X总字数. 则$T_{u_s,u_e,l}$的语义分数$score(T_{u_s,u_e,l})$由每个解码时间步生成字t^i的概率乘积计算得到, 如式 (1) 所示.

$$score(T_{u_s,u_e,l})=\sum_{i=1}^{n'}\log P(\hat{t}^i=t^i|t^{1:i-1},X) \tag{1}$$

$$\boldsymbol{c}^{1:n}=encoder(x^{1:n}) \tag{2}$$

$$\boldsymbol{h}^i=decoder(t^{1:i-1},\boldsymbol{c}^{1:n}) \tag{3}$$

$$P(\hat{t}^i|t^{1:i-1},X)=Softmax(\boldsymbol{W}\boldsymbol{h}^i+\boldsymbol{b}) \tag{4}$$

其中, *encoder* 和 *decoder* 分别表示 BART 的编码器与解码器, $\boldsymbol{c}^{1:n}\in\mathbb{R}^{d\times n}$表示对$X$进行自注意力机制编码后得到的上下文语义向量, $\boldsymbol{h}^i\in\mathbb{R}^d$表示在解码时间步$i$时, 结合$\boldsymbol{c}^{1:n}$与之前$i-1$步结果$t^{1:i-1}$, 得到的隐藏向量. d表示向量维数, $\boldsymbol{W}\in\mathbb{R}^{|V|\times d}$, $\boldsymbol{b}\in\mathbb{R}^{|V|}$为可训练的参数矩阵与向量, 用于将$\boldsymbol{h}^i$投影到 BART 词表$V$上, $|V|$表示字典大小. $P(\hat{t}^i|t^{1:i-1},X)$表示在解码时间步i时, 模型生成字$\hat{t}^i$的概率.

最终, TP-NER 选择$score(T_{u_s,u_e,l})$最高的(u_s,u_e,l)作为输出返回, 如图 2 所示, 返回 (1, 2, business), 识别出实体“复电”与实体类型“业务需求”.

2.3 主题模型生成提示词

上文已提到, NER 提示模板中的槽位 [MASK-r] 用于补充与类型 [MASK-t] 相关的语义信息, 以帮助 PLM 在少样本电力场景下区分实体类型. 由于行业文本的特殊性, 电力领域中的一种实体类型标签 (如“机器设备”“财务票据”“业务需求”) 往往可以看作一个主题, 而相关主题词可以视为对主题的进一步描述, 用于提示 PLM. 例如, 对于“故障异常”类型, 常见主题词有“掉落、停电、故障、破坏、波动、倾斜、失败、没电、欠费”等; 对于“业务需求”类型, 常见主题词有“需要、需求、业务、恢复、要求、申请、手续、办理、核实”等. 这些主题词在预训练的语料中就常常伴随着类型 (主题) 共同出现, 有利于为预训练语言模型提供语义提示, 从而帮助确定实体类型. 基于此动机, 本文使用经典的主题模型 LDA[20] 从训练语料中抽取主题词加入到提示模板中, 以增强 PLM 处理电力领域数量较多实体类型的能力.

2.3.1 文档构成

对于电力训练集中每个实体类型标签$l\in L$, 将其视为一个主题, 并收集包含l类型实体的所有训练文本, 例如, 当l为“业务需求”时, “复电手续如何办理”即为一个被收集的文本. 将所有收集到的文本拼接成一整篇电力文档, 记为D, 以便后续抽取与l相关的提示词.

2.3.2 文档建模过程

参考 LDA 模型[20], 对电力文档D进行基于实体类型 (即主题) 的建模. 整个过程包含单词、实体类型和

文档 3 层结构, 如图 3 所示. 在此设定中, D被视为一个词袋 (bag-of-words) 模型, 忽略其中单词的先后顺序.

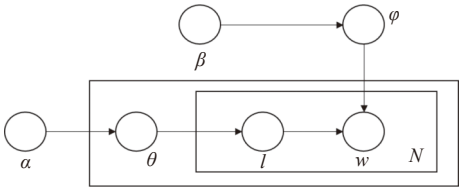

图 3 LDA 模型示意图

具体来说, 设θ表示实体类型在电力文档D上的概率分布, φ表示特定类型l上的单词概率分布, 则生成D的过程由参数α和β控制, 步骤如下:

(1) 根据泊松分布, 得到电力文档的词数N.

(2) 根据狄利克雷分布$Dir(\alpha)$, 得到电力文档的实体类型概率分布θ.

(3) 对于隐含实体类型l, 根据狄利克雷分布$Dir(\beta)$, 得到实体类型l下的单词概率分布φ.

(4) 对于D中的N个单词中的每个单词w_i, 首先根据θ的多项式分布$M(\theta)$, 随机选择一个实体类型l; 再根据l的多项式分布$M(\varphi)$, 随机选择一个单词作为w_i.

基于此过程, 在参数α,β条件下, 当所有单词都确定后, 得到电力文档D, 而生成D的条件概率$P(D|\alpha,\beta)$通过式 (5) 计算:

$$P(D|\alpha,\beta)=\int P(\theta|\alpha)\left(\prod_{i=1}^{N}\sum_{l}P(l|\theta)P(w_i|l,\beta)\right)d\theta \tag{5}$$

其中, $P(\theta|\alpha)$表示在参数α条件下实体类型θ的概率, $P(l|\theta)$表示选择类型l的概率, $P(w_i|l,\beta)$表示在已选择l的条件下选择单词w_i的概率.

2.3.3 生成提示词

为了对θ和φ进行估计, TP-NER 采用 LDA 中常用的 Gibbs 采样算法. 其过程可以看成上述文档生成过程的逆向过程, 即对于第 2.3.1 节中得到的电力文档D, 通过以下步骤进行参数估计:

(1) 为每个单词w_i随机分配一个实体类型l_i.

(2) 对于任意w_i, 设l_{-i}表示除w_i以外的其他单词的实体类型分布. 在已经得到l_{-i}的情况下, 计算w_i与实体类型为j的后验概率$P(l_i=j|l_{-i},w_i)$, 并将最可能的实体类型分配给w_i.

(3) 重复迭代步骤 (2), 直到每个单词w_i相关的实体类型分布收敛到稳定状态.

其中, $P(l_i=j|l_{-i},w_i)$通过式 (6) 计算得到:

$$P(l_i|l_{-i},w_i)=(n_{l_i}^{w_i}+\beta)/(n_{l_i}+N\beta)\cdot(n_D^{l_i}+\alpha)/(n_D+L\alpha) \tag{6}$$

其中, N和L分别表示电力文档中的单词总数和实体类型总数. $n_{l_i}^{w_i}$表示单词w_i与实体类型l_i相关的频数, n_{l_i}为所有单词都与l_i相关的总频数; $n_D^{l_i}$表示文档D中与类型l_i相关的单词频数, n_D表示D中的总单词数. 根据每个单词分配的相关实体类型时, 通过如式 (7) 和式 (8) 计算参数:

$$\varphi_{l,w}=(n_l^w+\beta)/(n_l+N\beta) \tag{7}$$

$$\theta_{D,l}=(n_D^l+\alpha)/(n_D+L\alpha) \tag{8}$$

其中, $\varphi_{l,w}$表示单词w与实体类型l相关的概率, $\theta_{D,l}$表示电力文档D出现实体类型l的概率. 结合$\varphi_{l,w}$与$\theta_{D,l}$, 则单词w在文档D中出现的概率$P_{D,w}=\varphi_{l,w}\times\theta_{D,l}$. 此概率反映了单词与文档主题的相关性, 是判断主题词的依据.

最后, 将电力文档D中所有的词按$P_{D,w}$排序, 选取前b个词作为与实体类型l的提示词, 填入模板T^+的 [MASK-r] 槽位, 完成最终提示句, 用于 BART 打分排序 (第 2.2.2 节). 这些提示句引导 BART 利用其蕴含的知识, 在少样本场景下, 完成对电力领域实体与细粒度类型的识别.

2.4 模型训练

为了训练 TP-NER 中的打分排序模型, 需要使用训练数据中提供的正确实体构建提示句. 假设实体跨度(u_s,u_e)的类型是l, 则将(u_s,u_e)与l填入正样本模板T^+, 得到目标提示句$T_{u_s,u_e,l}$. 若(u_s,u_e)不是一个实体跨度, 则将(u_s,u_e)填入负样本模板T^-, 作为目标提示句T_{u_s,u_e}. 根据训练集中所有正确实体, 可以构造出正样本对集合$\{(X,T^+)\}$; 通过随机枚举非实体的跨度(u_s,u_e), 可以构造负样本对集合$\{(X,T^-)\}$. 在实验中, 负样本对集合的大小为正样本集合大小的 1.5 倍. 对于每个样本对(X,T), 计算模型解码器输出的交叉熵损失, 以反向更新模型参数.

$$loss=-\sum_{i=1}^{n'}\log P(\tilde{t}^i|t^{1:i-1},X) \tag{9}$$

其中, $P(\tilde{t}^i|t^{1:i-1},X)$表示在解码时间步$i$生成模板中的字$\tilde{t}^i$的概率, 由式 (4) 得到, n'为X的字数.

3 实验与分析

3.1 实验环境

实验的硬件环境: Intel® Core 7700, 内存 8 GB. 软

件环境: Ubuntu 16.04, Python 3.6.8, GPU 采用 Nvidia RTX-2080ti 11 GB, 深度学习框架采用 PyTorch 1.4.0. 代码开发环境选择 PyCharm 2019.3.4.

3.2 数据集

本文重点关注中文电力领域, 采用国家电网真实工单数据与用户互动数据, 构建了电力领域命名实体识别数据集. 该数据集定义包括以下 14 种类型的实体: “机器设备、电价电费、业务需求、故障异常、财务票据、电子渠道、用户信息、文件法规、营销活动、身份、公司、违法行为、专业词汇”. 训练集, 验证集, 测试集, 分别包含 10 244、1 059、2 032 条电力文本与对应的实体、类型标注.

3.3 评价指标

本文采用准确率 (P)、召回率 (R) 以及 $F1$ 值 ($F1$) 作为模型性能的评价指标, 对测试集上的实体识别结果进行评估, 计算方式如下:

$$P = T_{TP}/(T_{TP}+F_{TP})\times 100\% \tag{10}$$

$$R = T_{TP}/(T_{TP}+F_{FN})\times 100\% \tag{11}$$

$$F1 = 2PR/(P+R)\times 100\% \tag{12}$$

其中, T_{TP}表示模型正确识别出的实体个数; F_{TP}表示模型识别出的不相关实体个数; F_{FN}表示实际为相关实体但模型识别错误的实体个数.

3.4 总体实验结果

本文对比的方法包括几种常用的 NER 模型: BiGRU、BiLSTM-CNN、BiLSTM-CRF、BiLSTM-CNN. 同时, 本文也与开放领域上表现优异的预训练模型进行对比: BERT 和 BERT-BiLSTM-CRF.

实验结果如表 2 所示. 从中可以看出, 本文提出 TP-NER 模型在中文电力 NER 数据集上击败了所有的对比模型, 取得了最好的结果. 对比开放领域中表现优异的 BERT-BiLSTM-CRF 模型, TP-NER 在 $F1$ 指标上提升了 2.17%, 这证明了本文提出的主题提示调优方法相比于传统序列标注方式, 在处理多实体类型的 NER 任务时更加有效.

表 2 电力命名实体识别总体结果 (%)

模型	P	R	$F1$
BiGRU	72.12	78.45	75.37
BiLSTM-CNN	73.41	79.84	76.56
BiLSTM-CRF	73.92	80.04	76.94
BERT	75.37	82.61	79.05
BERT-BiLSTM-CRF	76.72	83.18	79.88
TP-NER	**78.73**	**85.54**	**82.05**

3.5 消融实验

为了检验 TP-NER 模型中的两个主要改进: 提示模板与 LDA 主题提示词各自的贡献, 我们针对如下两种模型设置进行了消融实验:

1) 移除提示模板: 将提示模板移除, 仅使用 BART 对输入文本执行一般的序列标注以识别实体.

2) 移除 LDA 提示词: 不使用 LDA 模型对提示模板进行扩充, 仅依赖实体跨度与实体类型构建候选提示句并进行排序.

消融实验结果如表 3 所示, 移除提示模板后, 模型 $F1$ 下降了约 3.5%, 而移除 LDA 提示词后, $F1$ 下降约 1%, 这证明了这两个组件对模型均有贡献. 相比之下, 提示模板带来的提升比 LDA 带来的提升更大, 因为它从根本上改变了 NER 的任务形式.

表 3 消融实验结果 (%)

模型	P	R	$F1$
TP-NER	**78.73**	**85.54**	**82.05**
移除提示模板	75.46	81.19	78.57
移除LDA提示词	75.85	83.82	81.03

3.6 少样本场景实验结果

为了探究 TP-NER 在少样本场景下的表现, 本文设计如下少样本场景, 对于每个实体类型, 分别从训练集中随机抽取{10, 20, 50, 100}个样本组成小样本训练集训练模型, 再统计模型在测试集上的 $F1$ 分数.

实验结果如表 4 所示. 从中可以看出, 与使用全部训练集时相比, TP-NER 在少样本场景下相对对比模型的优势更大. 并且, 训练样本越少, TP-NER 的优势越明显. 同时可以发现, 在不同数量的训练样本, TP-NER 整体模型始终比移除提示模板后效果更好. 这充分说明了提示模板对于整个模型的贡献. 值得注意的是, LDA 提示词在样本数较少时提升更大.

表 4 少样本命名实体识别 $F1$ 分数 (%)

模型	10	20	50	100
BiGRU	5.34	8.25	13.55	20.41
BiLSTM-CNN	7.29	10.48	16.23	25.07
BiLSTM-CRF	12.41	16.65	21.31	29.93
BERT	21.32	29.99	40.06	51.42
BERT-BiLSTM-CRF	22.54	30.59	39.98	52.28
TP-NER	**40.25**	**48.20**	**58.12**	**71.24**
移除提示模板	18.45	27.31	39.08	50.47
移除LDA提示词	30.82	41.32	53.29	66.01

3.7 不同类型实验结果

为了探究 TP-NER 在不同实体类型下的表现, 本

文对测试集上每个类型都统计了模型的 *F*1 分数.

实验结果如表 5 所示. 从中可以看出, TP-NER 模型在大部分任务上都能取得比 BERT-LSTM-CRF 模型更好的效果, 尤其是在“文件法规”“公司”“营销活动”和“违法行为”这 4 种类型上提升最大. 造成这种显著提升的原因主要是, 在数据集中这些类型的标注样本较少, 平均不足 100 条, BERT-LSTM-CRF 模型没有足够的训练数据对其参数进行微调, 以至于利用 BERT 的先验知识完成识别. 相反地, TP-NER 将三元组输出的形式包装成自然语言形式, 使得 PLM 可以快速适配到电力 NER 任务上, 从而有效利用其知识. 在如“财务票据”“电子渠道”等类型上, TP-NER 稍稍落后于 BERT-LSTM-CRF, 这是因为这些类型的训练样本较多, 在这种训练资源丰富的场景中预训练模型微调与提示调优的差距不足以体现. 此外, 与“电价电费”“专业词汇”相比, “文件法规”“公司”“违法行为”这些类型的粒度更细, LDA 收集到的主题词与类型有着紧密的语义联系. 例如“违法行为”包含“民事责任”“举报”等密切相关的主题词, 因此可以精准地提示 PLM, 从而取得显著的提升.

表 5 各领域命名实体识别 *F*1 分数 (%)

领域	TP-NER	BERT-LSTM-CRF	领域	TP-NER	BERT-LSTM-CRF
机器设备	**75.35**	74.52	文件法规	**77.23**	72.17
电价电费	**82.56**	81.92	营销活动	**78.20**	71.32
业务需求	**72.19**	69.97	身份	**84.24**	83.18
故障异常	**69.97**	68.72	公司	**70.16**	60.16
财务票据	78.47	**79.32**	违法行为	**73.45**	62.12
电子渠道	88.56	**89.26**	专业词汇	**78.96**	78.45

4 结束语

本文针对中文电力领域场景的少样本问题和多类型问题, 提出一种基于主题提示的中文电力领域命名实体识别方法. 与传统的 BERT-LSTM-CRF 框架不同, 该方法提出了一种新的 NER 方式: 通过枚举实体跨度, 实体类型, 主题词从而构造候选提示句. 这种方式可以有效利用预训练模型中潜在的知识, 从而克服少样本 NER 的挑战.

此外, 该模型还提出使用 LDA 模型从语料中抽取主题词, 作为提示以增强模型对于实体类型的感知, 从而缓解实体类型较多带来的挑战.

实验结果表明, 本文的方法在电力场景中取得了比传统方法更好的结果. 尤其在“营销活动”“公司”“业务需求”等类型的实体识别上, 本文方法的优势更为显著.

在未来工作中, 尝试使用基于神经网络的方法替代主题模型, 引入外部知识以尝试解决更加困难的零样本 NER 任务.

参考文献

1 刘浏, 王东波. 命名实体识别研究综述. 情报学报, 2018, 37(3): 329–340. [doi: 10.3772/j.issn.1000-0135.2018.03.010]

2 Sang EFTK, De Meulder F. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. Proceedings of the 7th Conference on Natural Language Learning. Edmonton: ACL, 2003. 142–147.

3 Devlin J, Chang MW, Lee K, *et al*. BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: ACL, 2019. 4171–4186.

4 Ma XZ, Hovy E. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin: ACL, 2016. 1064–1074.

5 Lewis M, Liu YH, Goyal N, *et al*. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2020. 7871–7880.

6 Chiu JPC, Nichols E. Named entity recognition with bidirectional LSTM-CNNs. Transactions of the Association for Computational Linguistics, 2016, 4: 357–370. [doi: 10.1162/tacl_a_00104]

7 Strubell E, Verga P, Belanger D, *et al*. Fast and accurate entity recognition with iterated dilated convolutions. Proceedings of 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen: ACL, 2017. 2670–2680.

8 Zhang Y, Chen HS, Zhao YH, *et al*. Learning tag dependencies for sequence tagging. Proceedings of the 27th International Joint Conference on Artificial Intelligence. Stockholm: IJCAI, 2018. 4581–4587.

9 Cui LY, Zhang Y. Hierarchically-refined label attention

network for sequence labeling. Proceedings of 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong: ACL, 2019. 4115–4128.

10 Gui T, Ye JC, Zhang Q, *et al*. Uncertainty-aware label refinement for sequence labeling. Proceedings of 2020 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2020. 2316–2326.

11 Yamada I, Asai A, Shindo H, *et al*. LUKE: Deep contextualized entity representations with entity-aware self-attention. Proceedings of 2020 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2020. 6442–6454.

12 买买提阿依甫, 吾守尔·斯拉木, 帕丽旦·木合塔尔, 等. 基于 BiLSTM-CNN-CRF 模型的维吾尔文命名实体识别. 计算机工程, 2018, 44(8): 230–236.

13 李健龙, 王盼卿, 韩琪羽. 基于双向 LSTM 的军事命名实体识别. 计算机工程与科学, 2019, 41(4): 713–718. [doi: 10.3969/j.issn.1007-130X.2019.04.019]

14 顾亦然, 霍建霖, 杨海根, 等. 基于 BERT 的电机领域中文命名实体识别方法. 计算机工程, 2021, 47(8): 78–83, 92.

15 Wiseman S, Stratos K. Label-agnostic sequence labeling by copying nearest neighbors. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: ACL, 2019. 5363–5369.

16 Yang Y, Katiyar A. Simple and effective few-shot named entity recognition with structured nearest neighbor learning. Proceedings of 2020 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2020. 6365–6375.

17 Chen JD, Hu YZ, Liu JP, *et al*. Deep short text classification with knowledge powered attention. Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Honolulu: AAAI Press, 2019. 6252–6259.

18 Li WB, Sun L, Zhang DK. Text classification based on labeled-LDA model. Chinese Journal of Computers, 2008, 31(4): 620–627.

19 Schick T, Schmid H, Schütze H. Automatically identifying words that can serve as labels for few-shot text classification. Proceedings of the 28th International Conference on Computational Linguistics. Barcelona: ACL, 2020. 5569–5578.

20 Jelodar H, Wang YL, Yuan C, *et al*. Latent Dirichlet allocation (LDA) and topic modeling: Models, applications, a survey. Multimedia Tools and Applications, 2019, 78(11): 15169–15211. [doi: 10.1007/s11042-018-6894-4]

(校对责编: 孙君艳)