计算机科学
COMPUTER SCIENCE

基于样本贡献度对抗迁移的审计领域细粒度实体识别模型

庞博文, 陈一飞, 黄佳

引用本文

庞博文, 陈一飞, 黄佳. 基于样本贡献度对抗迁移的审计领域细粒度实体识别模型[J]. 计算机科学, 2024, 51(11A): 240300197-8.

PANG Bowen, CHEN Yifei, HUANG Jia. Fine-grained Entity Recognition Model in Audit Domain Based on Adversarial Migration ofSample Contributions [J]. Computer Science, 2024, 51(11A): 240300197-8.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240300197

基于样本贡献度对抗迁移的审计领域细粒度实体识别模型

庞博文　陈一飞　黄　佳
南京审计大学计算机学院　南京 211815
(xiaopang6831@163.com)

摘　要　细粒度命名实体识别(Named Entity Recognition,NER)在审计领域扶贫文本中识别实体信息,对优化扶贫政策成效分析与评估至关重要。近年来,深度学习在细粒度NER任务中取得显著成效,但特定领域仍面临语料集匮乏、迁移学习中细粒度特征不兼容性加剧及数据不平衡等问题。针对这些问题,制定了细粒度扶贫审计实体标签体系,并构建了细粒度扶贫审计语料集(FG-PAudit-Corpus)以解决审计领域数据集匮乏的问题。提出了基于样本贡献度对抗迁移的细粒度实体识别模型(FGATSC),该模型做对抗迁移训练,提出将样本贡献度权重纳入迁移特征中以解决细粒度特征的不兼容问题。同时,针对源域高资源与扶贫审计领域低资源样本的不平衡,提出了平衡资源对抗鉴别器(BRAD)以降低这种影响。实验结果表明,FGATSC模型在FG-PAudit-Corpus上F1的值为75.83%,较基线模型提高了9.03%,较其他主流模型提升了4.01%～6.53%;在Resume数据集上进行泛化性验证,F1值较近几年的主流模型提高约0.14%～1.31%,达到了95.77%。综上,验证了FGATSC模型的有效性和泛化性。

关键词:细粒度实体识别;扶贫审计;对抗训练;样本贡献度;平衡资源
中图分类号　TP391

Fine-grained Entity Recognition Model in Audit Domain Based on Adversarial Migration of Sample Contributions

PANG Bowen,CHEN Yifei and HUANG Jia
School of Computer Science,Nanjing Audit University,Nanjing 211815,China

Abstract　Fine-grained named entity recognition(NER) identifies entity information in pro-poor texts in the auditing domain, which is crucial for optimising the analysis and evaluation of pro-poor policy effectiveness. In recent years, deep learning has achieved significant results in fine-grained NER tasks, but the specific domain still faces problems such as the lack of corpus set, the increasing incompatibility of fine-grained features in transfer learning, and data imbalance. To address these issues, we formulate a fine-grained pro-poor audit entity labelling system and construct a fine-grained pro-poor audit corpus(FG-PAudit-Corpus) to address the scarcity of datasets in the audit domain. A fine-grained entity recognition model(FGATSC) based on sample contribution against migration is proposed, which does the training against migration and proposes to incorporate the sample contribution weights into the migrated features to solve the incompatibility problem of fine-grained features. Meanwhile, for the imbalance between high resources in the source domain and low resource samples in the pro-poor audit domain, balanced resource adversarial discriminator(BRAD) is proposed to reduce this effect. Experimental results show that the F1 value of the FGATSC model on FG-PAudit-Corpus is 75.83%, which is improved by 9.03% compared with the baseline model, and 4.01% to 6.53% compared with the other mainstream models. For the generalisation validation on the Resume dataset, the F1 is improved by about 0.14% to 1.31% compared with the mainstream models in recent years, and reaches 95.77%. In summary, the validity and generalizability of the FGATSC model are verified.

Keywords　Fine-grained entity recognition, Pro-poor auditing, Adversarial training, Sample contribution, Balancing resources

1　引言

审计领域的智能化研究正聚焦于文本数据结构化技术,如NER技术可以自动识别和提取文本中的人物、时间、组织机构等关键信息,将非结构化数据转化为结构化数据,为审计分析提供有力支持。随着我国扶贫审计工作的深入开展,扶贫审计文本作为记录扶贫政策实施情况的重要载体,显得愈发重要。这些文本涵盖了扶贫政策、扶贫项目和措施等信息,是评估扶贫工作的重要依据。细粒度NER技术是在NER的基础上进一步细化的技术,能够从审计文本中识别出更具体的实体,如扶贫人口、各级扶贫机构、中央和地方政府制定的扶贫政策、各领域的扶贫项目和措施等。这使得决策者能够更准确、更具体地评估扶贫政策成效,提高决策效率,并制定出更具针对性的决策。

目前主流的深度学习方法如Transformer[1]、Bidirectional Encoder Representations from Transformers(BERT)[2]、长短期记忆网络(Long Short-Term Memory,LSTM)[3]等模型在NER任务上已经展现出良好的效果。Lample等[4]提出双向长短期记忆网络[5](Bi-directional Long Short-Term Memory,BiLSTM)与条件随机场(Conditional Random Fields,CRF)结合的模型已成为解决NER问题的代表模型。但在特定领域中,传统的人名、地名等实体已难以满

通信作者:陈一飞(yifeichen91@nau.edu.cn)

足下游任务对精细实体进行划分的需求，基本的神经网络模型也很难在细粒度上获得更高的性能。因此，面向特定领域的细粒度 NER 技术显得尤为重要。

然而，在特定领域的应用上，这项技术仍然面临着很多问题。其中，特定领域语料库的匮乏成为了主要挑战之一，尤其在扶贫审计等领域，可用于训练的数据往往非常有限。在这种情况下，迁移学习[6]成为一种有效的方法。通过迁移学习，可以利用其他领域的数据或预训练模型的知识来弥补特定领域数据不足的情况，以提升模型在目标领域的表现。然而，在迁移过程中源域和目标域可能存在特征空间的差异，导致迁移时特征不兼容，细粒度特征的迁移更是会加剧这种不兼容性。在这种情况下，直接进行迁移可能会导致特征不匹配或者信息丢失。此外，源域和目标域之间可能存在数据分布不均衡的问题。高资源的源域拥有大量数据、丰富的信息，而低资源的目标域数据量有限、缺乏多样性。在这种情况下，迁移学习无法很好地利用源域的信息来促进目标域任务的性能提升。因此，针对上述问题提出了基于样本贡献度对抗迁移的审计领域细粒度实体识别模型(FGATSC)。本文的主要贡献如下：

1)为了解决语料库缺乏的问题，在现有审计语料库的基础上选取了部分扶贫审计文本，进行细粒度标注，构建了细粒度扶贫审计语料库(FG-PAudit-Corpus)。

2)为了缓解迁移过程中的细粒度特征不兼容的问题，提出将样本贡献度权重纳入共享特征的方法。通过为不同样本赋予适当的权重，使得模型更加倾向于迁移与目标领域相关的特征。

3)为了解决高低资源数据不平衡的问题，提出了平衡资源对抗鉴别器 BRAD，在训练过程中施加样本贡献度权重，平衡源域和目标域的数据分布，从而提高模型对低资源数据的关注度，进一步改善模型的泛化性和鲁棒性。

2 相关研究

2.1 细粒度 NER 任务

与 NER 相比，细粒度 NER 不仅仅是找出实体的位置和范围，还涉及按照特定领域的规范来细致地划分实体，即识别能够表现出该领域特定含义的内容。目前，大部分研究都集中在通用领域的实体识别上，对于在特定领域内对实体进行更细致识别和提取的相关研究非常有限。Fleischman 等[7]在 2002 年首次将细粒度 NER 描述为“命名实体的细粒度分类”。他们专注于人名的细粒度标签集，将一般的人标签分为 8 个子类别，即运动员、政治家/政府、神职人员、商人、艺人/艺术家、律师、医生/科学家和警察。他们尝试了各种经典的机器学习方法来完成这项任务，并在支持向量机(Support Vector Machine，SVM)、前馈神经网络和 C4.5 决策树的准确率方面分别取得了 68.1%，69.5%和 70.4%的良好结果。Mai 等[8]在英文数据集上进行实验，发现其最佳效果的模型采用了卷积神经网络[9](Convolutional Neural Network，CNN)与 LSTM 和 CRF 的组合。在此基础上通过移除 CNN 层并利用词典和类别嵌入，进一步提升了基于神经网络的日语细粒度实体识别性能，该方法将日语细粒度实体的识别率从 66.76%提高到 75.18%。Dogan 等[10]结合了当时最新的深度学习预训练语言模型 ELMo(Embeddings from Language Models)[11]和广阔的知识库 Wikidata，在 112 个跨越不同领域的细粒度实体类型上进行实验，实验结果表明该方法缓解了大量细粒度实体导致的实体识别率低的问题。Jiao 等[12]建立了反恐领域细粒度数据集标注体系，并将 MacBERT 预训练模型应用到反恐领域细粒度实体识别任务中，在该领域中 F1 值达到了约 78%。Cao 等[13]提出的细粒度 NER 模型融合了字向量信息和外部词典的潜在词向量信息，增强了字向量的表达，在 CLUENER2020 数据集上 F1 值达到了 82.46%。

2.2 对抗迁移

在特定领域的细粒度 NER 任务中，数据的稀缺性和领域特定性常常成为制约模型性能的重要因素。为了解决这些问题，研究者们开始关注对抗迁移技术的应用，通过在不同领域之间共享知识，提升模型在目标领域的性能。

近年来，对抗迁移技术在自然语言处理领域取得了显著进展。Lian 等[14]在军事科技领域 NER 任务中使用对抗自适应的迁移学习方法，同时在 CRF 解码层之后增加了虚拟对抗训练，对预测标签进行约束并正则化，使得 F1 值达到了 81.15%。Qian 等[15]构建粗粒度审计领域数据集，并在该数据集上结合 MacBERT 预训练语言模型进行对抗迁移训练，使用 CWS 任务辅助 NER 任务，F1 值达到了 91.05%。Cao 等[16]提出一种对抗迁移学习框架，结合了自注意力机制，通过整合共享词边界信息到命名实体识别任务中，并且成功防止了特定于分词任务的信息泄露，其在 SIGHAN 2006 数据集上取得了 90.64%的 F1 值。Zhou 等[17]提出了双重对抗迁移网络 DATNet，针对资源数据不平衡的问题提出了平衡资源权重超参数 α 和广义资源对抗鉴别器，在跨语言数据集 CoNLL-2002 上 F1 值达到了 88.16%。虽然上述研究使用对抗迁移技术在领域自适应、跨语言等方面取得了显著成果，但均未在细粒度 NER 领域有深入的研究，其所需的更为精细和具体的实体识别并未得到充分的关注。

3 细粒度扶贫审计语料库构建

由于公开文献中没有审计领域可用的细粒度 NER 数据集，因此在实体标签划分为人物名、地名和机构名 3 类的粗粒度审计领域数据集[15]的基础上选取部分数据，进一步细分并补充了更为详细的扶贫审计实体信息。其中，引用的粗粒度数据集是从中华人民共和国财政部官网及审计署官网上爬取官方新闻报道审计文本相关数据，使用标注软件标注。制定了 FG-PAudit-Corpus 标签体系，并构建了细粒度扶贫审计实体语料集 FG-PAudit-Corpus，对该语料集进行了统计。

3.1 制定细粒度扶贫审计语料库标签体系

结合审计领域的特点及专业审计人员的意见，与具备审计背景知识的人共同制定完善的标注体系。FG-PAudit-Corpus 标签体系涵盖了扶贫项目和措施等 5 个大类、15 个小类，如图 1 所示。根据贫困和非贫困人口，将人物类实体细化为扶贫人口类和普通人口类；地点类实体按照贫困地区和非贫困地区细化为扶贫地区和普通地区；组织机构类实体按照在扶贫审计工作中的不同职能和责任划分为扶贫部门、扶贫基地、扶贫工作小组、国际扶贫机构以及普通机构类实体。此外，为了优化对扶贫政策的评估工作，除了需要具体的扶贫政策，还需要各级政府和组织的具体执行措施。因此，提出了扶贫政策类实体与扶贫项目和措施类实体。扶贫政策对国家出台的政策和地方出台的政策进行了细化。

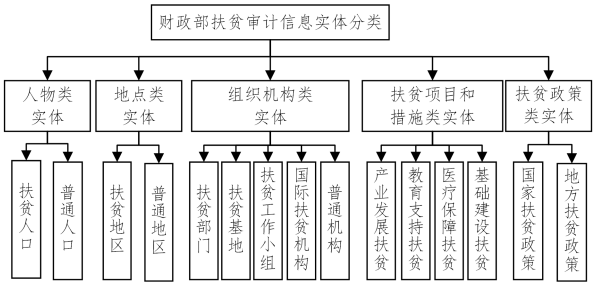

图 1　细粒度扶贫审计实体标签

Fig. 1　Fine-grained entity labels in pro-poor auditing field

根据扶贫工作所涉及的领域,扶贫项目和措施类划分为产业发展、教育支持、医疗保障和基础建设 4 个方面。产业发展具体包括农业、种植业,养殖业等扶贫项目和措施;教育支持包括提供教育资源、兴办学校等;医疗保障主要涉及医疗卫生资源的改善等;基础建设包括道路修建、供水、电力设施建设等方面的扶贫项目。具体细粒度实体类别名称、类别标识符及各个类别说明如表 1 所列。

表 1　FG-PAudit-Corpus 实体类别说明

Table 1　FG-PAudit-Corpus entity category description

细粒度实体类别	标识符	类别说明
扶贫人口	POV	处于贫困状态,需要接受政府或社会帮助的的人群
普通人口	PER	其他人名
扶贫地区	PSA	需要帮扶的地区名
普通地区	LOC	其他地名
扶贫部门	PAO	国家专设的扶贫机构
扶贫基地	POB	开展地区扶贫工作和帮助贫困地经济发展的机构
扶贫工作小组	PWG	致力于减贫的团队
国际扶贫机构	IPO	国际上有关扶贫的机构
普通机构	ORG	机构或者组织名
产业发展扶贫	IDP	从农业、种植业等产业角度实施的减贫项目或措施
教育支持扶贫	ESP	从教育资源的供给、改善教育条件等角度实现的减贫项目或措施
医疗保障扶贫	MSP	从医药资源、医疗保险覆盖等角度实现的减贫项目
基础建设扶贫	IIP	从建设基础设施等角度实现的减贫项目或措施
国家扶贫政策	NP	中央出台的扶贫政策
地方扶贫政策	LP	地方出台的扶贫政策

3.2　FG-PAudit-Corpus 统计

按照上述标签体系进行标注并构造,形成了细粒度扶贫审计语料库 FG-PAudit-Corpus。该语料库共计 2 053 个句子,按照 6∶2∶2 的比例划分为训练集、验证集以及测试集。其中,训练集中实体个数总计 3 760 个,验证集实体个数总计 1 171 个,测试集实体个数总计 1 375 个。FG-PAudit-Corpus 具体的实体类型和实体数量分布如表 2 所列。

表 2　FG-PAudit-Corpus 的划分及实体数量统计

Table 2　Division of FG-PAudit-Corpus and statistics of the number of entities

细粒度实体类型	细粒度实体数量分布		
	Train	Dev	Test
扶贫人口(POV)	148	33	48
普通人口(PER)	186	69	125
扶贫地区(PSA)	541	156	153
普通地区(LOC)	293	108	156
扶贫部门(PAO)	242	87	106
扶贫基地(POB)	355	112	150
扶贫工作小组(PWG)	306	77	90
国际扶贫机构(IPO)	219	72	61
普通机构(ORG)	697	253	267
产业发展扶贫(IDP)	168	44	52
教育支持扶贫(ESP)	20	9	6
医疗保障扶贫(MSP)	27	6	7
基础建设扶贫(IIP)	216	38	39
国家扶贫政策(NP)	45	17	14
地方扶贫政策(LP)	297	90	101

4　基于样本贡献度对抗迁移的细粒度实体识别模型

针对细粒度特征的迁移过程中出现的特征不兼容和资源不平衡的问题,提出了一种基于样本贡献度对抗迁移的细粒度 NER 模型 FGATSC。FGATSC 模型架构如图 2 所示,由两个私有任务和一个共享任务组成:两个私有任务分别是源域中文细粒度 NER(CFG_NER)任务和目标域细粒度 NER(Audit_NER)任务;共享任务是平衡资源对抗任务。接下来将详细介绍所提出模型中的任务。

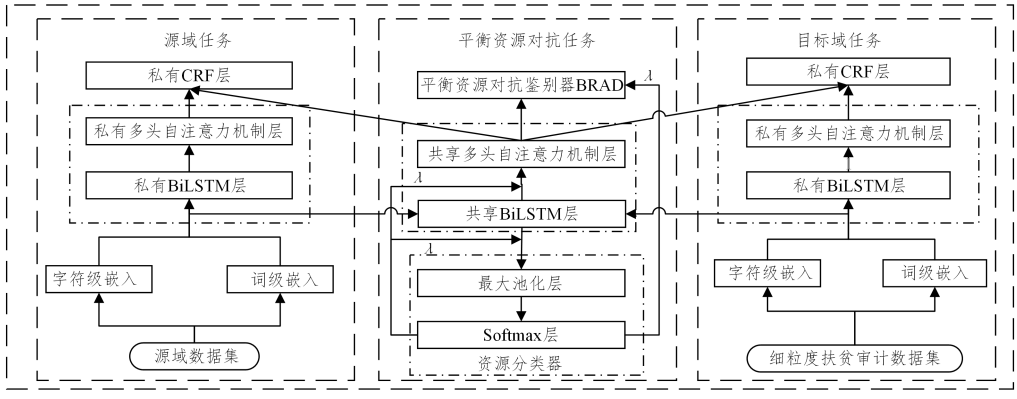

图 2　基于样本贡献度对抗迁移的细粒度实体识别框架

Fig. 2　Fine-grained entity recognition based on adversarial migration of sample contributions

4.1　Audit_NER 任务和 CFG_NER 任务

Audit_NER 任务和 CFG_NER 任务模型均由字词嵌入层、各自任务的私有特征编码层和私有 CRF 解码层构成。其中私有特征编码层由私有 BiLSTM 层和私有多头自注意力机制层组成,用于提取私有任务的特征信息。下面将对模型中的结构进行详细介绍。

4.1.1　字词嵌入层表示

嵌入层包含字符嵌入表示和词嵌入表示两种策略。在处理细粒度的数据集时,从其他领域学习到的字符特征对获取高质量的词特征和提升序列标记性能来说尤为重要。字符级编码器通常使用 BiLSTM 和 CNN 获取。Reimers 等[18]的研究指出,在序列标注任务中,字符级 CNN 的计算效率高、

参数更少。因此,使用 CNN 做字符嵌入以增强文本的特征表示。

词嵌入表示使用 BERT 预训练语言模型获取,并拼接字符级 CNN 获得的字符嵌入表示。给定输入的句子 $c=\{c_1,c_2,\cdots,c_n\}$。拼接的公式如式(1)所示,表示句子序列中第 i 个字符的最终嵌入表示。

$$x_i=\boldsymbol{x}_i^c\oplus\boldsymbol{x}_i^w \tag{1}$$

其中,$\oplus$表示连接操作,$\boldsymbol{x}^w$ 表示经过 BERT 处理之后得到词向量;$\boldsymbol{x}^c$ 表示字符级 CNN 处理之后得到字符向量。经过字词嵌入层之后,得到的最终向量序列为 $\boldsymbol{x}=\{x_1,x_2,\cdots,x_n\}$,将作为 BiLSTM 层的输入。

4.1.2 BiLSTM 层

长短期记忆网络(LSTM)引入门控单元来解决循环神经网络(Recurrent Neural Network,RNN)在处理长序列信息时可能出现的梯度消失或梯度爆炸问题。然而,单向 LSTM 学习的能力有限,无法学习到更完整的前后文信息,因此采用双向长短期记忆网络 BiLSTM 以充分利用输入的上下文信息,如图 3 所示。BiLSTM 由两个 LSTM 网络组成,一个 LSTM 从输入序列的开始到结束进行前向计算(前向 LSTM 层),另一个 LSTM 从输入序列的结束到开始进行反向计算(后向 LSTM 层)。

将 $\boldsymbol{x}=\{x_1,x_2,\cdots,x_n\}$ 作为语义编码层的输入,前向 LSTM 和后向 LSTM 都会产生一个输出。这两个输出被拼接在一起,形成 BiLSTM 在该时间的最终输出,隐藏层状态的计算式如式(2)和式(3)所示:

$$\overrightarrow{h_t}=\overrightarrow{LSTM}(\overrightarrow{h_{t-1}},x_t) \tag{2}$$

$$\overleftarrow{h_t}=\overleftarrow{LSTM}(\overleftarrow{h_{t+1}},x_t) \tag{3}$$

$$h_t=\overrightarrow{h_t}\oplus\overleftarrow{h_t} \tag{4}$$

其中,$\oplus$表示连接操作,$\overrightarrow{h_t}$和$\overleftarrow{h_t}$表示 t 时刻下的前后隐藏状态,$H=\{h_1,h_2,\cdots,h_n\}$表示私有 BiLSTM 层的输出。

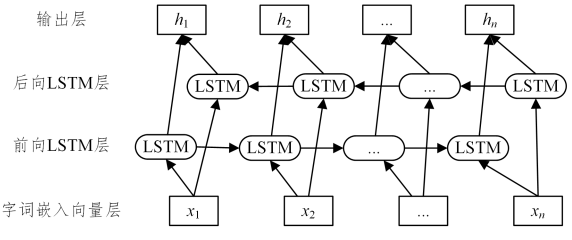

图 3 BiLSTM 结构图

Fig. 3 BiLSTM structure diagram

4.1.3 多头自注意力机制层

运用多头自注意力机制来明确学习句子中两个字符之间的依赖关系,以及捕获句子的内部结构信息。单头自注意机制的计算方式如式(5)所示:

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d}}\right)\boldsymbol{V} \tag{5}$$

其中,$\boldsymbol{Q}$ 是查询矩阵,$\boldsymbol{K}$ 是键矩阵,$\boldsymbol{V}$ 是值矩阵。

将 $\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}$ 的值均设置为 H。对输入的 $\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}$ 矩阵分别进行 h 次线性投影,计算如式(7)所示。第 i 次的线性投影计算式如式(6)所示。

$$head_i=Attention(\boldsymbol{Q}\boldsymbol{W}_i^Q,\boldsymbol{K}\boldsymbol{W}_i^K,\boldsymbol{V}\boldsymbol{W}_i^V) \tag{6}$$

$$H'=(head_i\oplus\cdots\oplus head_h)\boldsymbol{W}_o \tag{7}$$

其中,$\boldsymbol{W}_i^Q,\boldsymbol{W}_i^K,\boldsymbol{W}_i^V$ 是线性投影所需要的投影参数,$\boldsymbol{W}_o$ 是可训练的参数。

4.1.4 CRF 层

CRF 模型通常涉及对序列中的每个位置进行标签预测,提高模型对整体序列结构的建模能力。预测标签的计算式如式(8)、式(9)所示:

$$T_{i,y_i}=W_s h_i'+b_s \tag{8}$$

$$Score(X,Y)=\sum_{i=1}^{n}T_{y_i-1,y_i}+\sum_{i=1}^{n}T_{i,y_i} \tag{9}$$

其中,W_s 和 b_s 是可训练参数,h_i'是多头自注意力机制层的输出,X 为输入序列,Y 为预测标签序列,T_{y_i-1,y_i} 为从标签 y_i-1 转移到标签 y_i 的概率得分,T_{i,y_i} 表示为 c_i 被确认为 y_i 标签的得分。

采用极大似然估计来求得损失,计算式如下:

$$L=-\log\frac{\exp[Score(X,Y)]}{\sum\limits_{\widetilde{Y}\in Y_c}\exp[Score(X,\widetilde{Y})]} \tag{10}$$

4.2 平衡资源对抗任务

为了缓解迁移中细粒度特征不兼容以及资源不平衡的问题,提出了样本贡献度权重计算并将其纳入共享特征和平衡资源对抗鉴别器 BRAD 两种方法。平衡资源对抗任务由 3 个部分组成:资源分类器、共享特征编码层、平衡资源对抗鉴别器 BRAD。资源分类器由最大池化层和 Softmax 层组成,用于生成样本贡献度权重以兼容迁移的特征;共享特征编码层由共享 BiLSTM 层和共享多头自注意力机制层组成,用于提取两个任务的共享特征信息;平衡资源对抗鉴别器 BRAD 将提取的共享信息和样本贡献度权重相结合以平衡对抗迁移中的高低资源数据。下面将作具体介绍。

4.2.1 样本贡献度权重

样本贡献度权重是为了量化样本对源领域和目标领域的影响程度。考虑到源领域内不同样本对目标领域的影响程度不同,提出了一种样本贡献度加权的方法,即将得到的样本贡献度权重引入共享 BiLSTM 层中,使得模型在训练源域任务样本时能够根据该样本对目标领域样本的影响程度动态调整权重。

使用资源分类器来预测样本贡献度权重。该分类器的输出是预测某个样本属于目标领域的概率值,计算式如式(11)—式(13)所示:

$$L\_ad^m=MaxPooling(L^m) \tag{11}$$

$$D(l\_ad^m;\eta_d)=\mathrm{softmax}(W_d l\_ad^m+b_d) \tag{12}$$

$$[P^{\mathrm{target}},P^{\mathrm{source}}]=D(l\_ad^m;\eta_d) \tag{13}$$

其中,L^m 表示任务为 m 时共享 BiLSTM 层的输出,$m\in\{\mathrm{Audit\_NER},\mathrm{CFG\_NER}\}$;$W_d$ 和 b_d 表示可训练参数,η_d 表示鉴别器参数;P^{source} 表示样本为源域样本的概率,P^{target} 表示该样本为目标领域样本的概率,即样本贡献度权重。

样本贡献度权重的目的是通过衡量样本对目标域的影响程度来分配样本权重,因此需要归一化处理,如式(14)所示:

$$\lambda=\frac{\exp P^{\mathrm{target}}}{\sum\limits_i\exp P_i^{\mathrm{target}}} \tag{14}$$

其中,P_i^{target} 表示每个样本的贡献度权重,λ 为归一化后加入到共享 BiLSTM 层中的样本贡献度权重系数。

用 P^{target} 来衡量样本对于目标领域的重要程度。如果 P^{target} 的值较大,则表示该样本对于目标领域的重要程度更高,反之则重要程度较低。

当 m 为任务 CFG_NER 时,将样本贡献度权重系数加权到共享 BiLSTM 层中,如式(15)所示;再将共享 BiLSTM 层的输出结果输入到最大池化层简化特征,如式(16)所示;最后使用 softmax 函数得到新的样本贡献度权重。计算式如下:

$$L\_w^m = \lambda \cdot L^m \tag{15}$$

$$L\_mp^m = MaxPooling(L\_w^m) \tag{16}$$

$$D(l\_w'^m;\eta_d) = \text{softmax}(W_d L\_mp^m + b_d) \tag{17}$$

4.2.2 平衡资源对抗鉴别器

高资源的源域与低资源目标域的数据不平衡,导致模型在训练时随机梯度下降优化可能会更偏向于高资源领域,使得模型的识别效果变差,泛化能力减弱。为此,在将样本贡献度纳入共享特征的基础上,进一步提出了一种 Focal Loss[19] 损失的变体,称为平衡资源对抗鉴别器 BRAD。

FocalLoss 最初用于图像领域解决数据不平衡造成的模型性能下降问题,该损失函数通过降低简单负样本的权重,实现了困难样本挖掘,使模型更关注于分类难度较大的样本。而用于文本领域中,从交叉熵损失函数出发,将样本贡献度权重加入其中进行资源不平衡问题分析。通过在源域高资源上施加样本贡献度权重来平衡高低资源的影响。针对源域数据集中的每个样本对目标域的重要程度动态调整权重,使得模型更加平衡地利用高资源和低资源领域的样本信息,关注低资源样本的学习,防止高资源领域的样本主导梯度,从而更好地适应目标域的特点。BRAD 的损失函数如式(18)所示,将共享 BiLSTM 层的结果输入到共享多头自注意力机制中编码为单个向量,再通过线性变换投影到标量 r。

$$l_{\text{BRAD}} = -\sum_i Z_{i\in \text{CFG\_NER}} \lambda_i (1-r_i)^{\rho} \log r_i - \sum_i Z_{i\in \text{Audit\_NER}} r_i^{\rho} \log(1-r_i) \tag{18}$$

其中,Z 表示是 Audit_NER 任务和 CFG_NER 任务的判断函数。r_i^{ρ} 和 $(1-r_i)^{\rho}$ 是通过测量预测值与真实标签之间的差异控制各个样本的损失贡献。超参数 ρ 是用来衡量在训练过程中困难样本和简单样本的损失贡献对比,在实验中通常设置为 2。引入样本贡献度权重系数 λ 和 $(1-r_i)\rho$ 可以分别降低高资源样本和简单样本对总损失的贡献。对于高资源样本,权重系数会相应减小,以减轻其在模型训练中的主导地位;而对于简单样本,权重系数同样会减少,以避免这些样本过度影响训练过程,从而促使模型更加关注难以预测的样本。

4.3 迁移训练

4.3.1 迁移

迁移通过将平衡资源对抗任务中通过共享特征编码层的输出与两个私有任务的私有特征编码层实现拼接,计算式如式(19)所示。其中 $L\_w'^m$ 表示将样本贡献度加权的结果 L_w^m 输入到共享多头自注意力机制中,通过计算得到的输出。再将拼接后的结果作为输入分别传至 Audit_NER 任务和 CFG_NER 任务的私有 CRF 层计算损失值,如式(20)—式(23)所示:

$$H'' = H' \oplus L\_w'^m \tag{19}$$

$$T'_{i,y_i} = W_j h_i'' + b_j \tag{20}$$

$$Score'(X,Y) = \sum_{i=1}^{n} T'_{y_{i-1},y_i} + \sum_{i=1}^{n} T'_{i,y_i} \tag{21}$$

$$L_{\text{Audit\_NER}} = -\log \frac{\exp[Score(X,Y)]}{\sum_{\widetilde{Y}\in Y_c} \exp[Score(X,\widetilde{Y})]} \tag{22}$$

$$L_{\text{CFG\_NER}} = -\log \frac{\exp[Score(X,Y)]}{\sum_{\widetilde{Y}\in Y_c} \exp[Score(X,\widetilde{Y})]} \tag{23}$$

其中,W_j 和 b_j 是可训练参数。

4.3.2 对抗迁移训练

受到对抗网络和迁移学习的启发,使用源域任务来辅助目标域任务训练。将通过共享 BiLSTM 获得的两个任务的的共享信息输入最大池化层,进行特征压缩。将池化后的特征向量进行 sofmax 分类,判断特征归属,得到样本贡献度度量值 λ。将 λ 纳入共享特征中进行训练以减少细粒度迁移特征的不兼容性,然后再使用多头自注意力机制增强编码特征。此外,使用平衡资源对抗鉴别器 *BRAD* 帮助模型更好地平衡资源。对抗迁移训练的伪代码如算法 1 所示。

算法 1 对抗迁移训练

初始化:学习率 learning_rate,批次 batch_size,贡献对比参数 ρ 等超参数

输入:细粒度扶贫审计语料集,CLUENER2020 数据集

输出:LAudit_NER,LCFG_NER,BRAD

1. 输入数据,字词嵌入向量表示
2. 将 Audit_NER 任务和 CFG_NER 任务的嵌入向量表示输入共享 BiLSTM 层
3. for 循环,使用平衡资源对抗任务提取两个私有任务中的共享信息
4. 私有 BiLSTM 和共享 BiLSTM 提取特征
5. 分别输入到私有和共享多头自注意力机制
6. 两个私有任务由 CRF 层计算损失

$$L_{\text{Audit\_NER}} = -\log \frac{\exp[\text{Score}(X,Y)]}{\sum_{\widetilde{Y}\in Y_c} \exp[\text{Score}(X,\widetilde{Y})]}$$

$$L_{\text{CFG\_NER}} = -\log \frac{\exp[\text{Score}(X,Y)]}{\sum_{\widetilde{Y}\in Y_c} \exp[\text{Score}(X,\widetilde{Y})]}$$

7. 将共享 BiLSTM 层提取到的特征输入到由 MaxPooling 和 softmax 构成的资源分类器进行判别,得到样本贡献度并归一化后得到权重系数 λ
8. 计算平衡资源对抗任务的损失:

$$l_{\text{BRAD}} = -\sum_i Z_{i\in \text{CFG\_NER}} \lambda_i (1-r_i)^{\rho} \log r_i - \sum_i Z_{i\in \text{Audit\_NER}} r_i^{\rho} \log(1-r_i)$$

9. 优化参数 W_d,b_d 直到结果最佳
10. end for

4.3.3 模型训练

通过对 Audit_NER 任务损失函数 $L_{\text{Audit\_NER}}$、CFG_NER 任务的损失函数 $L_{\text{CFG\_NER}}$ 以及平衡资源对抗任务损失函数 l_{BRAD} 的计算,构建了总损失函数 L,计算式如式(24)所示:

$$L = I \cdot L_{\text{Audit\_NER}} + (1-I) \cdot L_{\text{CFG\_NER}} + \psi \cdot l_{BRAD} \tag{24}$$

其中,I 表示判断输入来自 Audit_NER 任务还是 CFG_NER 任务,ψ 为损失权重系数。

5 实验结果与分析

5.1 实验数据

使用 FG-PAudit-Corpus 作为目标域 Audit_NER 任务的

1) https://github.com/CLUEbenchmark/CLUENER2020

实验数据集，CLUENER2020 数据集[1)] 作为源域 CFG_NER 任务的实验数据集。在 Resume 数据集[20]上进行实验，验证模型的泛化性和有效性。Resume 数据集使用 4700 多条公司人员的简历信息作为语料，其中涵盖了 8 种实体类型，如种族(RACE)、国籍(CONT)、职称(TITLE)、专业(PRO)、人名(NAME)等。各个数据集的信息如表 3 所列。

表 3 各数据集实体数和句子数统计

Table 3 Statistics of the number of entities and sentences in each dataset

数据集	类型	训练集	验证集	测试集
细粒度扶贫审计数据集	句子	1262	406	385
	实体	3760	1171	1375
Resume 数据集	句子	3821	463	477
	实体	1340	160	150
CLUE2020 细粒度数据集	句子	10748	—	1343
	实体	23338	—	2982

5.2 评价指标

命名实体识别任务通常使用 $F1$ 值，准确率 P、召回率 R 来评估模型的性能，计算式如下：

$$p=\frac{TP}{TP+FP}\times 100\% \tag{25}$$

$$R=\frac{TP}{TP+FN}\times 100\% \tag{26}$$

$$F1=\frac{2PR}{P+R}\times 100\% \tag{27}$$

其中，TP 表示正确的样本被预测为正确的数量，FP 表示错误的样本被预测为正确的数量，FN 表示正确的样本被预测为错误的数量。

5.3 实验环境和模型参数

本次实验运行的环境配置为：操作系统是 Linux 内核的 Ubuntu 系统 20.04，python 的版本号为 3.8.0，GPU 为 RTX3090，深度学习框架是 Pytorch1.9.1+cu102。本次实验的超参数配置如表 4 所列。

表 4 超参数配置

Table 4 hyperparametric configurations

名称	数值
embedding_word	768
embedding_char	200
dropout	0.5
batch_size	32
epoch	100
ψ	0.06
LSTM_dim	256
ρ	0.2
optimizer	Adam
learning_rate	5×10^{-5}

5.4 实验设计与分析

5.4.1 嵌入式对比实验

本文的基线模型为 BERT+BiLSTM+CRF，验证在嵌入层使用 BERT 预训练语言模型作为词级嵌入与使用字符级 CNN 作字符嵌入拼接结果在细粒度扶贫审计数据集上的有效性。结果如表 5 所列。

由表 5 中的实验结果可知，基线模型对比第三组实验，使用 BERT 和 CNN 拼接的字词结合嵌入方式的 F1 值为 69.30%，相比基线模型，P 和 $F1$ 的值分别提升了 8.8%，2.5%，说明了字符级特征能帮助模型更好地理解词内部的结构和语法特征。而基线的 BERT 捕获的上下文信息，是以词级别的 token 为单位处理文本，该方法并不完善，字符级特征的引入填补了这方面的不足，提升了序列标记性能。

表 5 嵌入方式对比

Table 5 Comparison of embedding methods

(%)

模型	P	R	$F1$
BERT+BiLSTM+CRF	61.68	72.84	66.80
BERT+ CNN +BiLSTM+CRF	70.48	68.16	69.30

5.4.2 平衡资源对抗训练对比实验

为了验证本文提出的平衡资源对抗训练任务的有效性，进行平衡资源对抗训练的对比实验，实验结果如表 6 所列。

表 6 平衡资源对抗训练对比

Table 6 Comparison of adversarial training with balanced resources

(%)

模型	P	R	$F1$
BERT+CNN +BiLSTM+CRF (不加迁移对抗)	70.48	68.16	69.30
BERT+CNN +BiLSTM+CRF (加迁移对抗)	72.82	70.85	71.82
BERT+CNN +BiLSTM+CRF (加迁移对抗，加样本贡献度)	73.15	74.32	73.73
FGATSC (加迁移对抗，加样本贡献度，BRAD)	75.30	76.37	75.83

由表 6 的实验结果可得，第二组实验与第一组实验对比，表明加入对抗训练之后，P，R，$F1$ 的值分别提升了 2.34%，2.69%，2.52%。这是因为加入对抗训练之后，使得目标域任务在训练的同时学习了来自源域任务的特征信息，使模型更好地适应各种数据的扰动和噪声，有效地减少了模型的过拟合，从而提高了模型的泛化能力。

第三组实验在对抗训练的基础上增加了样本贡献度权重；对比第二组实验，P，R，$F1$ 的值分别提升了 0.33%，3.47%，1.91%，表明模型缓解了细粒度特征迁移带来的特征差异问题，降低了源域样本特征对目标域任务的影响。这种调整使模型更加关注那些对于目标域任务而言更为关键的源域样本特征，通过对这些关键样本特征的强化学习，模型更能够理解并利用源域数据中的重要特征，提高模型在目标任务上的泛化能力和性能。

第四组实验，在样本贡献度的基础上，使用了 BRAD，$F1$ 的值较第一组实验结果提升了 6.53%，表明 FGATSC 在提升模型整体性能方面的有效性。实验结果较第三组实验均有提升，P，R，$F1$ 的值分别提升了 2.15%，2.05%，2.1%，表明了在 BRAD 的帮助下，提高了模型对数据分布的敏感性，使其在兼容迁移特征的同时，也平衡了高低资源数据之间的差异。通过 BRAD，使得模型更多地关注低资源和困难样本，加强了对关键样本的学习能力，进一步优化了模型在低资源目标域上的性能表现。因此，所有的实验结果表明，FGATSC 模型的实验效果是最好的。

5.4.3 细粒度实体识别结果对比

为了能够全面评估模型在不同细粒度实体类别上的表现，以针对性地优化模型，对所有的细粒度实体的识别效果进行统计和对比，如表 7 所列。

表 7　FGATSC 模型和基线模型的实体识别结果对比

Table 7　Comparison of entity identification results between FGATSC model and baseline model

(%)

实体类别	基线模型 *F*1	FGATSC 模型 *F*1
扶贫人口(POV)	11.11	87.23
普通人口(PER)	75.86	90.04
扶贫地区(PSA)	80.70	83.98
普通地区(LOC)	3.74	37.99
扶贫部门(PAO)	51.61	81.31
扶贫基地(POB)	66.24	67.20
扶贫工作小组(PWG)	52.34	71.23
国际扶贫机构(IPO)	70.24	74.67
普通机构(ORG)	73.68	82.30
产业发展扶贫(IDP)	31.50	32.20
教育支持扶贫(ESP)	0.00	50.11
医疗保障扶贫(MSP)	0.00	50.00
基础建设扶贫(IIP)	18.26	30.43
国家扶贫政策(NP)	6.25	51.61
地方扶贫政策(LP)	76.68	80.00

对比两个模型所识别的细粒度实体类别结果来看,FGATSC 模型的细粒度实体识别结果较基线模型都有提高。ESP 和 MSP 类实体因为训练的数据量不足 30,在基线模型上的识别效果为 0,但在 FGATSC 模型上识别的效果却有 50%左右;NP 类实体数量不足 50,基线模型的识别率只有 6.25%,但 FGATSC 模型的识别达到 51.61%,远远高出基线模型。由此可以看出,FGATSC 模型对提高识别低资源的细粒度实体具有很好的效果。在 FGATSC 模型的识别结果中,IDP 类的实体识别效果是最低的,相比基线模型的识别效果提高也是最少的。可能的原因是:一方面,在训练数据中该类别的标注量不够充分,导致模型在学习和识别时存在困难;另一方面,该类别涉及的方面比较多,如农业、种植、养殖、光伏、经济、资产项目等,导致这类实体的特征在文本中不够突出和明确,缺乏独特的词语特征,使得 FGATSC 模型难以区分和准确识别。

在 FGATSC 模型的识别结果中,对比 3 个基础的大类,即人物类、地点类和组织机构类中,PER 实体和 ORG 实体的识别效果都是最好的,但是在地点类中 LOC 实体的识别效果却远低于 PSA 实体。除了数据量的原因,PSA 实体通过上下文可以使模型更容易准确识别,相比之下,LOC 实体的表现形式可能更为独立和固定,对上下文的依赖性不大,导致模型的识别效果差于 PSA 实体。综合全部的细粒度小类别来看,PER 类和 POV 类的实体识别效果是最好和次好的,这是因为这类实体在语料库中出现频率较高且具有一定的规律性,例如固定的姓氏和名字构成模式,这使得模型更容易学习和识别人名的特定模式和规律。此外,这类实体的标注一致性较高,标注者通常能够准确辨识人名,减少了标注的歧义性。

5.4.4　泛化性验证实验

为了验证 FGATSC 模型的泛化能力,将其与近几年在 Resume 数据集上实验效果较好的其他模型做对比实验,实验结果如表 8 所列。

由表 8 可知,与经典的 Lattice-LSTM[19] 模型和 LR-CNN[21] 模型相比,FGATSC 模型分别使用 BiLSTM 网络和 CNN 网络结合词典的方式来融合所有与词典内容相关的词向量表示,其 F1 值提高分别提高了 0.66%和 1.31%。LLL-WCM+avg[22] 模型采用平均策略并结合自注意力机制提取字词融合的低频词汇信;MM-SLLattice[25] 模型在字级别信息融合开放词典和领域词典的基础上,引入深度相互学习的方法来提高模型的识别能力;ProConBERT[23] 采用提示学习与对比学习的 BERT 预训练策略,结合各种预训练模型在中文小样本 NER 任务上训练。对比上述几个模型策略,FGATSC 模型在 *F*1 上有约 0.14%～0.62%的提升,表明了使用对抗迁移对于提升模型的性能是有效的。此外,TES-NER[24] 模型使用迁移学习的方法将跨领域的实体特征信息从源域迁移到目标域来提高目标任务的性能,但 FGATSC 模型的 *F*1 比其提高了 0.28%,表现出样本贡献度和平衡资源鉴别器对于迁移之后模型效果的提升是有帮助的。综上所述,FGATSC 模型有较好的泛化性。

表 8　在 Resume 数据集上的对比实验

Table 8　Comparison experiments on Resume dataset

(%)

模型	*P*	*R*	*F*1
Lattice-LSTM[20]	94.81	94.11	94.46
LR-CNN[21]	95.37	94.84	95.11
LLL-WCM+avg[22]	95.12	95.09	95.15
ProConBERT[23]	94.20	94.96	95.42
TES-NER[24]	95.57	95.4	95.49
MM-SLLattice[25]	95.43	95.83	95.63
FGATSC	95.21	96.34	95.77

结束语　针对扶贫审计领域语料集缺乏,以及对抗迁移带来的资源表征差异问题和训练过程中源域与目标域的资源数据不平衡的问题,构建了细粒度扶贫审计语料集并提出了 FGATSC 模型。该模型提出了将样本贡献度权重纳入共享特征的计算方法和平衡资源鉴别器 BRAD。首先将样本贡献度权重作为共享特征的系数加入平衡资源对抗训练中,一定程度使得源域任务和目标域任务的特征表示更加兼容;然后将样本贡献权重系数作为平衡资源鉴别器的参数,缓解了目标模型在训练时优化性能向源域高资源倾斜的问题,致使高资源的源域数据特征难以对目标域的低资源实体数据起到辅助作用。本文的实验验证了该模型的有效性,并且具备良好的泛化能力。

在未来的研究中,将对低资源的扶贫审计领域数据集进行补充,考虑采用远程监督的方法,继续完善细粒度扶贫审计数据集。将主动学习算法与对抗迁移相结合,在降低语料需求的前提下,进一步提高对扶贫审计实体的识别效果。

参 考 文 献

[1] VASWANI A,SHAZEER N,PARMAR N,et al. Attention is all you need[C]// Proceedings of the Annual Conference on Neural Information Processing Systems. 2017:6000-6010.

[2] YUAN L C. Joint Method for Chinese Word Segmentation and Part-of-speech Tagging Based on BERT-BiLSTM-CRF[J]. Journal of Chinese Computer Systems,2023,44(9):1906-1911.

[3] HUANG Z,XU W,YU K. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv:1508.01991,2015.

[4] LAMPLE G,BALLESTORS M,SUBRAMANIAN S,et al. Neural architectures for named entity recognition[J]. arXiv:1603.01360,2016.

[5] JIANG T Q,WAN Z H,ZHANG Q C. Text classification of

food safety judgment document based on BiLSTM and self-attention[J]. Science Technology and Engineering, 2019, 19(29): 191-195.
[6] LI M, LI Y L, LIN M. Review of Transfer Learning for Named Entity Recognition [J]. Journal of Frontiers of Computer Science and Technology, 2021, 15(2): 206-218.
[7] FLEISCHMAN M, HOVY E. Fine grained classification of named entities[C]// Proceedings of the 19th International Conference on Computational Linguistics. 2002: 1-7.
[8] MAI K, PHAM T H, NGUYEN M T, et al. An empirical study on fine-grained named entity recognition[C] // Proceedings of the 27th International Conference on Computational Linguistics. 2018: 711-722.
[9] CHIU J, NICHOLS E. Named entity recognition with bidirectional lstm-cnns[J]. Transactions of the Association for Computational Linguistics, 2016: 4: 357-370.
[10] DOGAN C, DUTRA A, GARA A, et al. Fine-grained named entity recognition using elmo and wikidata[J]. arXiv: 1904. 10503, 2019.
[11] PETERS, MATTEHW E, et al. Deep contextualized word representations[C] // Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics. 2018: 2227-2237.
[12] JIAO K N, LI X, YE H, et al. Fine-grained Entity Recognition Based on MacBERT-BiLSTM-CRF in Anti-terrorism Field [J]. Science Technology and Engineering, 2021, 21 (29): 12638-12648.
[13] CAO H, XU Y. Fine-grained Named Entity Recognition Based on Words Information [J]. Computer Applications and Software, 2023, 40(3): 235-240.
[14] LIAN Y, FENG J C, DING H. Named Entity Recognition In Military Technology Field Based On Adversarial Transfer Learning [J]. Electronic Design Engineering, 2022, 30(20): 121-127.
[15] QIAN T Y, CHEN Y F, PANG B W. Audit Text Named Entity Recognition Based on MacBERT and Adversarial Training [J]. Computer Science, 2023, 50(S2): 93-98.
[16] CAO P, CHEN Y, LIU K, et al. Adversarial transfer learning for Chinese named entity recognition with self-attention mechanism [C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018: 182-192.
[17] ZHOU J T, ZHANG H, JIN D, et al. Dual adversarial neural transfer for low- resource named entity recognition[C] // Proceedings of the 57th Conference of the Association for Computational Linguistics. 2019: 3461-3471.
[18] REIMERS N, GUREVYCH I. Reportingscore distributions makes a difference: Performance study of lstm-networks for sequence tagging[C]// EMNLP. 2017: 338-348.
[19] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection [C] // Proceedings of the IEEE international Conference on Computer Vision, 2017: 2980-2988.
[20] ZHANG Y, YANG J. Chinese NER using lattice LSTM [C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018: 1554-1564.
[21] GUI T, MA R, ZHANG Q, et al. CNN-based Chinese NER with lexicon rethinking [C] // Proceedings of the 28h International Joint Conference on Artificial Intelligence. 2019: 4982-4988.
[22] GUO Z Q, GUAN D H, YUAN W W. Word-Character Model with Low Lexical Information Loss for Chinese NER [J]. Computer Science, 2024, 51(8): 272-280.
[23] YANG S H, LAI P C, FU Y G, et al. Optimization Method of BERT for Chinese Few-shot Named Entity Recognition[J/OL]. Journal of Chinese Computer Systems, 2024: 1-12. http://kns.cnki.net/kcms/detail/21.1106.TP.20240202.0926.002.html.
[24] WU B C, DENG C L, GUAN B, et al. Dynamically Transfer Entity Span Information for Cross-domain Chinese Named Entity Recognition [J]. Journal of Software, 2022, 33(10): 3776-3792.
[24] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection [C] // Proceedings of the IEEE International Conference Cn Vomputer Vision. 2017: 2980-2988.
[25] CHENG T, HONG H Y, YANG D S, et al. Chinese named entity recognition model based on mutual learning and SoftLexicon [J]. Computer Application, 2023, 43(S1): 61-66.

PANG Bowen, born in 1999, postgraduate, is a member of CCF (No. T8233G). His main research interest is text mining.

CHEN Yifei, born in 1977, Ph.D, associate professor. Her main research interests include text mining and intelligent information extraction.