DOI:10.19651/j.cnki.emt.2210566

基于 BERT 和知识蒸馏的航空维修领域命名实体识别

顾佼佼[1] 翟一琛[1] 姬嗣愚[2] 宗富强[1]
(1.海军航空大学 烟台 264001; 2.91475 部队 葫芦岛 125001)

摘 要: 针对军事航空维修领域命名实体识别训练数据少,标注成本高的问题,改进提出一种基于预训练 BERT 的命名实体识别方法,借鉴远程监督思想,对字符融合远程标签词边界特征得到特征融合向量,送入 BERT 生成动态字向量表示,连接 CRF 模型得到序列的全局最优结果,在自建数据集上进行实验,F1 值达到 0.861。为压缩模型参数,使用训练好的 BERT-CRF 模型生成伪标签数据,结合知识蒸馏技术指导参数量较少的学生模型 BiGRU-CRF 进行训练。实验结果表明,与教师模型相比,学生模型以损失 2%的 F1 值为代价,参数量减少了 95.2%,运算推理时间缩短了 47%。

关键词: 航空维修文本;命名实体识别;BERT;知识蒸馏;伪标签增强;词向量增强
中图分类号: TP391.1 **文献标识码:** A **国家标准学科分类代码:** 520.6010

Aviation maintenance text named entity recognition based on BERT and knowledge distillation

Gu Jiaojiao[1] Zhai Yichen[1] Ji Siyu[2] Zong Fuqiang[1]
(1. Naval Aviation University, Yantai 264001, China; 2. The No. 91475th Troop of PLA, Huludao 125001, China)

Abstract: Aiming at the problems of less training data and high labeling cost of named entity recognition in the military aircraft maintenance field. The paper proposed an improved named entity recognition method based on pre-training BERT. Firstly, learn from the idea of remote supervision, we fuse the boundary features of remote Tag word on token to get the feature fusion vector. Then the vector is sent to BERT to generate a dynamic word vector representation. Finally, the CRF model is connected to get the global optimal result of the sequence. Experiments are carried out on the self-built dataset, and the F1 value reaches 0.861. In order to compress the model parameters, the trained BERT-CRF model is used to generate pseudo label data, and the student model BiGRU-CRF with less parameters is trained in combination with knowledge distillation technology. The experimental results show that compared with the teacher model, the student model reduces 95.2% of the parameters and 47% of the reasoning time at the cost of losing 2% of the F1 value.

Keywords: aviation maintenance text; named entity identification; BERT; knowledge distillation; pseudo label enhancement; word vector enhancement

0 引 言

随着军事航空装备信息化的高速发展,军事航空维修领域积攒了海量的非结构化装备维修文本,利用命名实体识别技术可以对丰富而杂乱的信息进行处理和识别,从中提取出与保障维修、作战指挥等有关的重要信息,及时准确地为人员决策提供重要依据[1]。

命名实体识别任务(named entity recognition, NER)[2]是信息抽取领域内的一个子任务,其任务目标是识别非结构化文本中的实体并将其分类为预定的类别,通常包括人名、地名、机构名等[3],是自然语言处理领域的基础性工作之一,可为信息过滤、知识图谱、语义检索等任务提供支撑[4]。相比于英文文本,中文文本没有表示词语边界的符号,每一个字符一般也不是单独的词语,相关实体边界更难确定、语法结构更加复杂,所以中文 NER 任务更困难。中文命名实体识别模型一般分为基于字符的模型和基于词语的模型,基于词语的模型效果建立在分词的基础上,分词过程的错误会直接降低 NER 任务的识别效果,通常采用基于字符的模型效果更好[5]。

近年来,得益于 LSTM(long short term memory)、

收稿日期:2022-07-04

ELMo (embeddings from language models)[6]、BERT (bidirectional encoder representation from transformers)[7]等深度学习模型的提出，命名实体识别技术实现了新的突破，并由通用领域迅速扩展到医疗、农业、电商等垂直领域，基于深度学习的方法已成为研究的主流，并且可以满足一定的应用需求[8]，文献[9]利用命名实体识别方法提取关键词，用于提升文本匹配算法性能。文献[10]提出使用预训练 BERT 模型并融合词性信息进行军事命名实体识别，识别效果较 BiLSTM-CRF 框架有较大提升，验证了在军事命名实体识别领域应用 BERT 模型的有效性。模型性能提升的同时，模型参数呈几何倍数增加，实时性也无法得到保证，限制了深度学习技术的应用。文献[11]提出知识蒸馏理论，使用教师模型预测的标签概率分布指导学生模型的训练，通过这种方式指导简单模型训练，可以显著提升简单模型的性能。

非结构化装备维修文本中包含了大量的航空装备、机务维修等相关专业术语，实体密度大，标注成本高，为了在少量标注数据下达到较好的命名实体识别效果，本文选择基于预训练 BERT 模型的命名实体识别方法。针对缺乏领域词典的问题，借鉴远程监督思想，将数据集进行 10 折划分，构造每折数据的远程标签，加强词典的领域属性。为压缩模型参数，提升模型的推理速度，使用知识蒸馏结合伪标签[12]数据增强的方法，首先使用教师模型预测无标签数据得到伪标签数据，之后结合真实标签以 BERT-CRF 为教师模型指导学生模型 BiGRU-CRF 的训练。实验结果表明，经过此方法进行知识蒸馏的学生模型在命名实体识别准确度上较传统方法蒸馏的模型效果更好。与教师模型相比，极大降低了模型的参数量，减少了模型的预测时间。

1　命名实体识别模型

1.1　BERT 模型

BERT 是一种使用 Transformer Encoder[13]结构的双向编码语言模型，其结构如图 1 所示，本文使用 BERT 完成对文本序列的特征抽取。BERT 模型采用两阶段的训练模式，首先在大规模无监督数据上进行掩码预训练(masked language model, MLM)和下半句分类预训练(next sentence prediction, NSP)，接着再针对具体的下游任务进行参数微调，大量实验证明，这种训练方式使得 BERT 模型针对小样本数据也具有很强的泛化能力。

将 BERT 模型应用于军事航空维修领域的命名实体识别任务中，能够更加细粒度的考虑句子中字符之间的关系，更好地获取字符的向量表示，在少量标注数据上实现较高的命名实体识别准确度。

1.2　词边界特征

字符级别的句子解码方式对词语边界信息考虑不充分。本文借鉴 LEBERT[14]和 Simple-Lexicon[15]增强文本词边界表示的方式，分别在词嵌入层和词嵌入层后进行两次词信息增强。命名实体识别模型结构如图 2 所示。

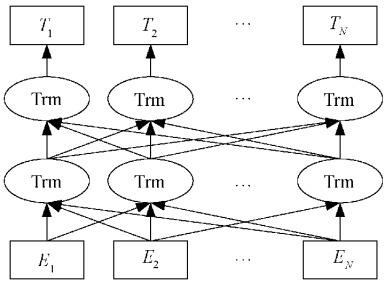

图 1　BERT 模型结构

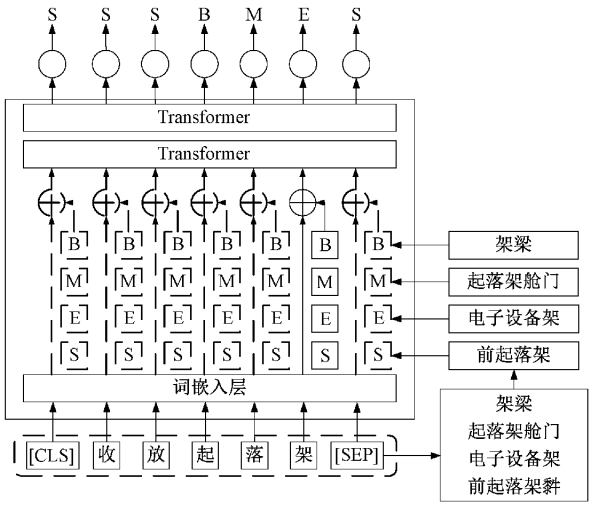

图 2　命名实体识别模型结构

首先如果句子中连续 2～5 个字符出现在词表中，则将词向量与该字的字向量进行拼接，一并送入词嵌入层。然后考虑字符在词中出现的位置，以“架”字为例，使用 BMES 标注法，如果这个字在词的开始位置，则将该词统计到这个字 B 对应的集合中，如“架梁”中“架”字位于词的开始位置；如果这个字在词的中间位置，则将该词统计到这个字 M 对应的集合中，如“起落架舱门”中“架”字位于词的中间位置；如果这个字在词的结束位置，则将该词统计到这个字 E 对应的集合中，如“电子设备架”中“架”字位于词的结束位置；如果这个字单独成词，则将该词统计到这个字 S 对应的集合中。

在垂直领域中，完善的领域专业词典往往难以获得，且输入端额外加入信息通常需要人工标注。采取十折划分数据集的方式远程构建每折数据的候选词典，即利用其他九折数据中的标注标签词语和通用领域词典作为每一折数据的词典，使每折数据都利用到了所有数据的标注信息。

在融合词汇信息时，将每个集合中的词转换为词向量并进行加权求和，再将所有集合的词向量表示与字向量进行拼接。时间步 t 的字符向量表示记作：

$$\boldsymbol{x}_t \leftarrow [\boldsymbol{e}^c(c_t);\boldsymbol{e}^w(w_t),\boldsymbol{e}^s(\mathrm{B,M,E,S})] \tag{1}$$

其中，$\boldsymbol{e}(w_t)=\sum_{i-1}^{m}a_{ti}\boldsymbol{e}(w_i)$，$Z=\sum_{w\in \mathrm{B}\cup\mathrm{M}\cup\mathrm{E}\cup\mathrm{S}}z(w_t)$，

$$v^s(S)=\frac{4}{Z}\sum_{w\in S}z(w_t)\boldsymbol{e}^w(w_t),$$

$$\boldsymbol{e}^s(B,M,E,S)=[v^s(B);v^s(M);v^s(E);v^s(S)]$$

c_t 表示当前时间步的字符，w_i 为 c_t 对应的词集合中的词，$\boldsymbol{e}(w_t)$ 为 c_t 对应的词集合中所有词嵌入加权，a_{ti} 为该词在语料库中出现的频率，S 为词的集合，$\boldsymbol{e}^w$ 为在语料库上训练的 Word2Vec[16] 模型的词嵌入层，e^c 为字符表示嵌入层，$z(w_t)$ 为每个集合的权重，本文使用词嵌入对字符嵌入的注意力作为权重，$\forall w\in B\cup M\cup E\cup S$ 公式定义如下：

$$z(w)=\mathrm{softmax}(s(\boldsymbol{e}^c(c_t),z(w)))=\frac{\exp(s(\boldsymbol{e}^c(c_t),z(w)))}{\sum_{w\in S}\exp(s(\boldsymbol{e}^c(c_t),z(w)))} \tag{2}$$

$$s(\boldsymbol{e}^c(c_t),z(w))=\tanh(\boldsymbol{e}^c(c_t))\odot\tanh(z(w)) \tag{3}$$

1.3 CRF 层

在特征提取层 BERT 的输出后接 Softmax 层可以得到模型对每个输入时间步的预测结果，但模型缺少对标签间转移规律的约束，使用条件随机场(conditional random field，CRF)可以得到整个预测序列的全局最优结果。

首先特征提取部分的输出经过多层感知器(multilayer perceptron，MLP)映射到标签维度，再经过 Softmax 层归一化得到 CRF 层输入记作 $\boldsymbol{X}=\mathrm{softmax}(\mathrm{MLP}(\boldsymbol{h}_1,\cdots,\boldsymbol{h}_t))$，假设对应的标签序列为 $\boldsymbol{Y}_{\mathrm{RealPath}}=\{y_1,y_2,\cdots,y_t\}$，则对应的真实路径得分可表示为：

$$s(\boldsymbol{X},\boldsymbol{Y}_{\mathrm{RealPath}})=\sum_{i=1}^{t}(\boldsymbol{Z}_{y_{i-1},y_i}+\boldsymbol{P}_{i,y_i}) \tag{4}$$

其中，$\boldsymbol{Z}$ 为转移特征，Z_{y_{i-1},y_i} 为标签从 y_{i-1} 转移到 y_i 的分值，P 为状态特征，P_{i,y_i} 为输入序列第 i 个字被预测为标签 y_i 的概率。为最大化真实路径的得分，建立损失函数表示如下：

$$loss_{crf}=-\log\frac{s(\boldsymbol{X},\boldsymbol{Y}_{\mathrm{RealPath}})}{\sum_i s(\boldsymbol{X},\boldsymbol{Y}_i)} \tag{5}$$

1.4 知识蒸馏

BERT-CRF 模型的命名实体识别性能优越，但同时存在参数量大，推理时间长等缺点。本文使用知识蒸馏结合伪标签数据对 BERT-CRF 模型进行蒸馏，以参数量较多的 BERT-CRF 模型为教师模型，以参数量较少的 BiGRU-CRF 模型为学生模型，进行知识蒸馏的流程如图 3 所示。

为充分利用未标注数据，使用训练完成的 BERT-CRF 模型预测未标注数据得到伪标签数据，与真实标签一并指导学生模型的训练。在损失函数的计算中赋予伪标签数据动态权重 λ 作为模型参数进行训练。总的蒸馏损失表示如下：

$$L_{KD}=\lambda L_{pseudo}+(1-\lambda)L_{train} \tag{6}$$

使用带温度的 Softmax 函数软化教师模型逻辑层的输出，使用 KL 散度衡量教师模型输出的概率分布与学生模型的差异，同时学生模型的输出概率与训练集的真实标签构造交叉熵损失。伪标签数据与真实标签数据损失计算方式相同，以真实标签数据损失 L_{train} 的计算为例：

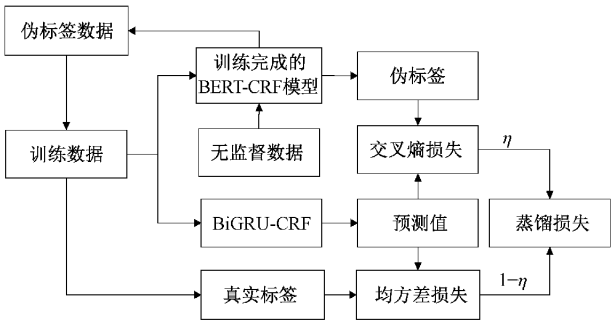

图 3 知识蒸馏流程

$$L_{train}=\eta L_{soft}+(1-\eta)L_{hard} \tag{7}$$

其中，$L_{soft}=-\sum_i^N p_i^T\log(q_i^T)$，$L_{hard}=-\sum_i^N c_i\log(q_i)$ $p_i^T=\frac{\exp(v_i/T)}{\sum_j\exp(v_j/T)}$，$q_i^T=\frac{\exp(z_i/T)}{\sum_j\exp(z_j/T)}$，$q_i=\frac{\exp(z_i)}{\sum_j\exp(z_j)}$，$T$ 表示温度，T 值越大，模型输出的概率分布会越平滑；p_i^T 表示教师模型在温度 T 下第 i 类标签的输出概率，q_i^T 表示学生模型在温度 T 下第 i 类标签的输出概率，c_i 表示第 i 类的真实标签值；v_i 表示教师模型输出的各类别分布，z_i 表示学生模型输出的各类别分布，η 为超参数。

2 实验验证

2.1 数据集构建

本文数据来源于航空行业 IETM 相关技术手册 PDF 文档，首先对文本内容进行抽取、清洗，按句号进行分割，然后对其中长度大于 128 的长句再次进行分割，划分标注实体“装备系统”、“装备部件”、“维修保障”、“人员”和“故障”共 5 类，采用 BMES 序列标注格式进行人工标注数据共 5 030 条。按照 7∶3 的比例划分为训练集和测试集，各实体数量与实体示例如表 1 所示。

表 1 各实体类别数量及示例

实体类别	实体数量		实体示例
	训练集	测试集	
装备系统	358	171	液压系统
装备部件	10 756	4 713	襟翼、天线
维修保障	2 212	921	工作梯、扳手
故障	691	316	卡滞、裂纹
人员	305	117	用户、地勤

对训练集数据进行十折交叉验证规范数据集标注形

式，具体的标注策略为对训练集进行十折划分，分别进行训练得到 10 个模型，再用这 10 个模型对训练集进行预测，如果 10 个预测结果相同而标注数据中不存在该结果，则认定为漏标，补充该标注；如果 10 个预测结果中均不存在标注数据中的某个标注，则认定为错标，删除该标注。

2.2　实验配置

1)实验设置

实验环境为处理器 Inter(R) Xeon(R) Gold 5218R、操作系统 Ubuntu 18.04.2LTS、显卡 RTX 3090，Python3.7、Pytorch1.8.0。模型使用 Adam 优化器，BERT 学习率设置为 2×10^{-5}，CRF 层学习率设置为 2×10^{-3}。

2)评价指标

本文使用精确率(precision，P)，召回率(recall，R)和加权 F1 值(weighted-F1 score)作为评价指标。加权 F1 值是根据每一类的样本数目加权平均每一类的 F1 值，通常用在类别不均衡的 NER 任务中，表示形式如下：

$$P=\frac{T_p}{T_p+F_p}\times100\% \tag{8}$$

$$R=\frac{T_p}{T_p+F_n}\times100\% \tag{9}$$

$$F1=\sum_{i=0}^{n}\alpha_i\frac{2P_i\cdot R_i}{P_i+R_i}\times100\% \tag{10}$$

其中，n 为标签总数，α_i 为第 i 类标签的数量占所有标签数量的比例。

2.3　实验结果及分析

在自建数据集上，选取 BiGRU-CRF、BiLSTM-CRF、RoBERTa-CRF、BERT-CRF 作为基准模型进行比较如表 2 所示，实验结果表明，使用预训练 BERT 模型可以大幅提升命名实体识别准确度，但同时参数量也大幅增加。

对 BERT-CRF 及 BiGRU-CRF 模型的输入端融入词边界特征进行实验如表 3 所示。维修手册内包含大量的装备、工具型号，且同一实体往往有多种表征形式，实验结果表明，将 BERT 或 BiGRU 作为特征抽取模型时，使用远程标签结合通用词典作为领域词典融入词边界特征的方式，与仅使用通用词典相比，可以在不增加外部知识和模型复杂度的情况下，提升垂直领域命名实体识别准确度。

表 2　基准模型对比实验

模型	精确率	召回率	F1	参数量
BiGRU-CRF	0.829	0.772	0.800	6.67×10^6
BiLSTM-CRF	0.820	0.763	0.790	6.78×10^6
RoBERTa-CRF	0.858	0.853	0.855	1.37×10^8
BERT-CRF	0.861	0.852	0.857	1.37×10^8

表 3　词边界特征对比实验

模型	精确率	召回率	F1	参数量
BERT-CRF	0.861	0.852	0.857	1.37×10^8
BERT-CRF+通用词典特征	0.862	0.854	0.858	1.39×10^8
BERT-CRF+领域词典特征	0.867	0.857	0.861	1.39×10^8
BiGRU-CRF	0.829	0.772	0.800	6.67×10^6
BiGRU-CRF+通用词典特征	0.832	0.772	0.802	1.01×10^7
BiGRU-CRF+领域词典特征	0.829	0.785	0.807	1.01×10^7

为充分利用维修文本中的未标注数据，本文使用知识蒸馏结合伪标签数据增强的方式指导学生模型训练，使用经过词边界特征加强的 BERT-CRF 作为教师模型，参数量较少的 BiGRU-CRF 作为学生模型，实验结果如表 4 所示。在不使用伪标签数据增强的情况下，模型蒸馏在自建数据集下的 F1 值提升不大，这表明数据量不足时，传统的蒸馏方式无法很好地将大模型知识迁移至小模型。

表 4　知识蒸馏实验

模型	知识蒸馏	数据集	精确率	召回率	F1	参数量
BERT-CRF		训练集	0.867	0.857	0.861	1.39×10^8
BiGRU-CRF		训练集	0.829	0.772	0.800	6.67×10^6
BiGRU-CRF	√	训练集	0.825	0.788	0.807	6.67×10^6
BiGRU-CRF	√	训练集+伪标签数据增强	0.844	0.839	0.841	6.67×10^6

在使用伪标签数据增强策略时，对使用伪标签的数量进行实验。结果如图 4 所示，随着伪标签数据的逐步增加，在伪标签数量小于 9 000 条时，学生模型性能快速提升，在伪标签数量大于 9 000 条后，模型性能趋于稳定，在使用 18 000 条伪标签时，模型的 F1 值达到了 0.841。与单独训练学生模型相比，F1 值提升了 0.041，与传统知识蒸馏方式相比，F1 值提升了 0.034，与教师模型相比，F1 值仅相差 0.02。

图 5 是知识蒸馏过程中不同超参数 T 和 η 对实验结果的影响，$T=5$，$\eta=0.3$ 时学生模型的蒸馏效果最好。

另外，本文使用教师模型和学生模型预测相同的 1 000 条测试数据，二者耗时分别为 17.33 s 和 9.28 s，学生模型的推理时间缩短至教师模型的 47%。综上，使用此方法进行模型蒸馏，在模型损失 2%的 F1 值的情况下，模型参数量和推理时间都得到了很大的改善。

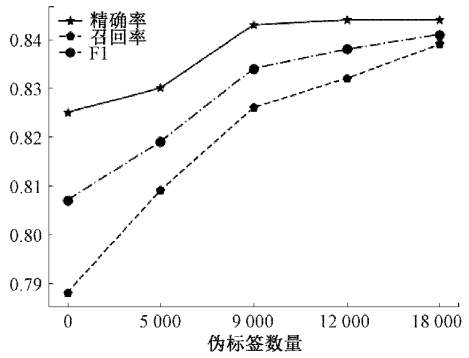

图 4 伪标签数量对学生模型性能的影响

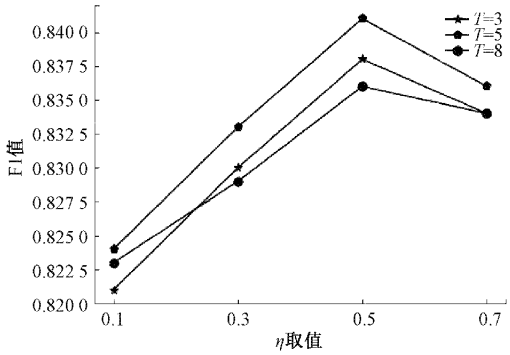

图 5 不同 T、η 对学生模型蒸馏效果的影响

3 结 论

本文提出一种应用于军事航空维修领域的命名实体识别方法,基于 BERT-CRF 命名实体识别模型,借鉴远程监督思想,充分利用全体标签信息增强词典的领域属性,之后对模型融入词边界特征提升模型性能,实验结果验证了所提方法的有效性。此外,引入知识蒸馏和伪标签数据增强技术,训练学生模型 BiGRU-CRF,较传统知识蒸馏方式大幅了提升学生模型的命名实体识别准确度,同时模型参数、推理时间均大幅减少,便于模型的部署与在线应用。

参考文献

[1] 姜文志,顾佼佼,胡文萱,等. 基于多模型结合的军事命名实体识别[J]. 兵工自动化,2011,30(10):90-93.

[2] ZHAO S, CAI Z P, CHEN H W, et al. Adversarial training based lattice LSTM for Chinese clinical named entity recognition [J]. Journal of Biomedical Informatics, 2019, 99(14):103290.

[3] 姜文志,顾佼佼,丛林虎,等. CRF 与规则相结合的军事命名实体识别研究[J]. 指挥控制与仿真,2011,33(4):13-15.

[4] LI Y, LIU L, SHI S. Empirical analysis of unlabeled entity problem in named entity recognition[J]. ArXiv Preprint,2020, ArXiv:2012.05426.

[5] PENG N, DREDZE M. Named entity recognition for Chinese social media with jointly trained embeddings[C]. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, 548-554.

[6] PETERS M E, NEUMANN M, IYYER M, et al. Deep contextualized word representations [J]. The North American Chapter of the Association for Computational Linguistics, 2018, 2227-2237.

[7] LEE J D M C K, TOUTANOVA K. Pre-training of deep bidirectional transformers for language understanding [J]. The North American Chapter of the Association for Computational Linguistics, 2019: 4171-4186.

[8] 刘浏,王东波. 命名实体识别研究综述[J]. 情报学报,2018,37(3):329-340.

[9] 王逸凡,李国平. 基于语义相似度及命名实体识别的主观题自动评分方法[J]. 电子测量技术,2019,42(2):84-87.

[10] 张乐,李健,唐亮,等. 基于预训练 BERT 的军事领域目标实体深度学习识别方法[J]. 信息工程大学学报,2021,22(3):331-337.

[11] HINTON G, VINTALS O, DEAN J. Distilling the knowledge in a neural network[J]. ArXiv Preprint, 2015, ArXiv:1503.02531.

[12] LEE D H. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks [C]. Workshop on challenges in representation learning, International Conference on Machine Learning, 2013, 3(2): 896.

[13] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [J]. Advances in neural information processing systems, 2017, DOI: 10.48550/arXiv.1706.03762.

[14] LIU W, FU X, ZHANG Y, et al. Lexicon enhanced chinese sequence labeling using BERT adapter [J]. Annual Meeting of the Association for Computational Linguistics, 2021: 5847-5858.

[15] MA R, PENG M, ZHANG Q, et al. Simplify the usage of lexicon in Chinese NER[J]. Annual Meeting of the Association for Computational Linguistics. 2020: 5951-5960.

[16] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space [J]. International Conference on Learning

Representations, 2013: 1-12.

作者简介

顾佼佼，博士，讲师，主要研究方向为深度学习技术。
E-mail:542939566@qq.com

翟一琛(通信作者)，硕士研究生，主要研究方向为自然语言处理。
E-mail:912781740@qq.com

姬嗣愚，硕士，主要研究方向为武器系统设计与模拟技术。
E-mail:1694259706@qq.com

宗富强，硕士，主要研究方向为武器工业与军事技术。
E-mail:641940546@qq.com