基于 BERT 与二级融合的小样本命名实体识别

弓源 毛璐 汪美玲 李长亮 *

AI Lab, KingSoft Corp, Beijing, China
{gongyuan, maolu, wangmeiling1, lichangliang}@kingsoft.com

Abstract. 目前，面向军事文档的命名实体识别作为军事文档信息抽取的基本任务获得了极大关注。2020 年 CCKS 组委会联合军事科学院系统工程研究院发布的面向试验鉴定的命名实体识别任务，要求对包括试验要素、性能指标、系统组成和任务场景在内的 4 种类型实体进行识别。由于领域的特殊性及保密性，主办方共公布标注数据 400 篇。本文将该评测任务抽象为小样本场景下的命名实体识别问题，并提出基于 BERT 与二级融合的小样本命名实体识别方法。所提方法首先基于 BERT 在评测训练数据上微调得到多个基础模型，之后提出一种应用于多个基础模型预测结果的二级融合策略，以缓解基础模型在小样本训练数据上存在的过拟合问题，最后通过后处理的方式剔除标注错误并生成最终的预测结果。本文方法在评测任务的测试集上达到 0.7203 的 F1 值。

Keywords: 小样本命名实体识别 · BERT · 二级融合.

1 引言

命名实体识别（Named Entity Recognition, NER）[9] 是自然语言处理领域的基础任务之一，其目标是抽取文本中具有基本语义的实体单元，在知识图谱构建、信息抽取、信息检索、机器翻译、智能问答等系统中都有广泛应用。目前，由于基于深度学习端到端的实体识别方法能够避免手工特征工程且能够挖掘深层特征，其效果远优于传统基于规则的方法与基于统计学习的方法，因此已经成为主流解决方案。其中，BERT 预训练模型 [4] 更是因其强大的特征学习能力在 NER 任务中取得了卓越的效果。

近年来，随着信息技术的发展，军事业务数据呈现爆炸式增长，诸如军事装备试验鉴定等方面的军事文档大量存在并被使用，如何从这些军事文档中自动获取有效信息成为亟待解决的问题。面向军事文档的命名实体识别作为军事文档信息抽取的基本任务获得了极大关注。然而由于数据稀缺以及标注困难，面向军事文档的命名实体识别技术仍需要进一步研究改进。

以推动试验鉴定领域命名实体识别技术为目的，2020 年 CCKS 组委会联合军事科学院系统工程研究院发布了面向试验鉴定的命名实体识别任务。在该任务中，由于领域的特殊性及保密性，主办方共公布标注数据 400 篇，将军事装备试验鉴定的命名实体划分为试验要素、性能指标、系统组成和任务场景 4 种类型。本文将该评测任务抽象为小样本场景下的命名实体识别问题，并提出了基于 BERT 与二级融合的小样本命名实体识别方法。所提方法在训练阶段利用基础模型 BERT+CRF[10] 和 BERT+Bi-LSTM+CRF[5] 在评测训练数据集上进行微调，在预测阶段首先利用基础模型的微调训练结果进行预测，之后为了缓解基础模型在小样本训练数据上存在的过拟合问题，提出一种由 Logits 融合

美军正在测试一款新型电磁导轨炮，可以约 7 2 4 0 千米 / 小时的速度发射弹药 。
O O O O O O O O O O B-s I-s I-s I-s I-s O O O O O O O O O O O O O O O O B-x I-x O O B-t I-t O

Fig. 1. 样本数据标签映射示例.

与差异性融合组成的二级融合策略以提高模型预测效果，最后通过后处理的方式剔除标注错误并生成最终的预测结果。本文方法在评测任务的测试集上达到了 0.7203 的 F1 值。

2 相关工作

命名实体识别的主要方法包括基于规则的方法、基于统计学习的方法和基于深度学习的方法。

基于规则和字典的命名实体识别方法依赖大量的先验知识，人工成本较高。此外，其还存在时间效率低、可移植性弱等缺点 [8]。

基于统计学习的命名实体识别方法避免了大量规则的制定，常用方法包括最大熵模型 [1]、隐马尔可夫模型 [15]、支持向量机 [7] 和条件随机场 [14] 等。然而，基于统计学习的命名实体识别方法依赖于预先定义的特征，特征工程不仅代价高而且与特定领域相关，因此模型的泛化能力和迁移能力较差 [13]。

基于深度学习的端到端模型能够避免手工特征工程且能够挖掘深层特征，是当前的研究热点。循环神经网络及其变体模型 [6] 和卷积神经网络及其变体模型 [3] 在命名实体识别中得到广泛应用。近年来，预训练词向量技术受到了越来越多的关注 [11]。其中，BERT 预训练语言模型由谷歌 AI 团队于 2018 年发布并开源 [4]。BERT 本质上是在海量的无标签语料上，通过自监督学习的方式训练得到一个泛化能力较强的特征表示，能够更深层次地提取文本的语义信息，目前已在 11 种 NLP 领域测试任务中获得了最佳成绩，其中也包括命名实体识别任务，基于 BERT 预训练模型在命名实体识别任务数据上微调训练可达到卓越的实体标注效果。

3 问题定义

给定一段军事试验鉴定描述文本，评测任务的目标是识别出文本中的命名实体提及并确定其所属预定义类别。官方共提供 400 条可用于训练的文本标注数据，给出描述文本以及文本语料中的实体位置与类别，同一文本语句中出现过一次的实体后续不再进行标注。样本数据中包含 4 类实体：试验要素、性能指标、系统组成、任务场景，每个类别的实体数目分布不平衡。因此，该评测任务可以抽象为小样本场景下的命名实体识别问题。

命名实体识别是一种典型的序列标注任务，本文对上述训练数据进行标签映射，采用 BIO（Begin, Inside, Outside）标注格式标注文本数据。具体是令 s 表示试验要素、x 表示性能指标、t 表示系统组成、r 表示任务场景，则标注标签的总数 $label\_num = 9$，包括 4 类实体的 B 标签和 I 标签以及 1 个其它 O 标签，最终生成字符级别的标注数据。例如 “美军正在测试一款新型电磁导轨炮，可以约 7240 千米/小时的速度发射弹药。” 的标签映射结果如图 1所示。

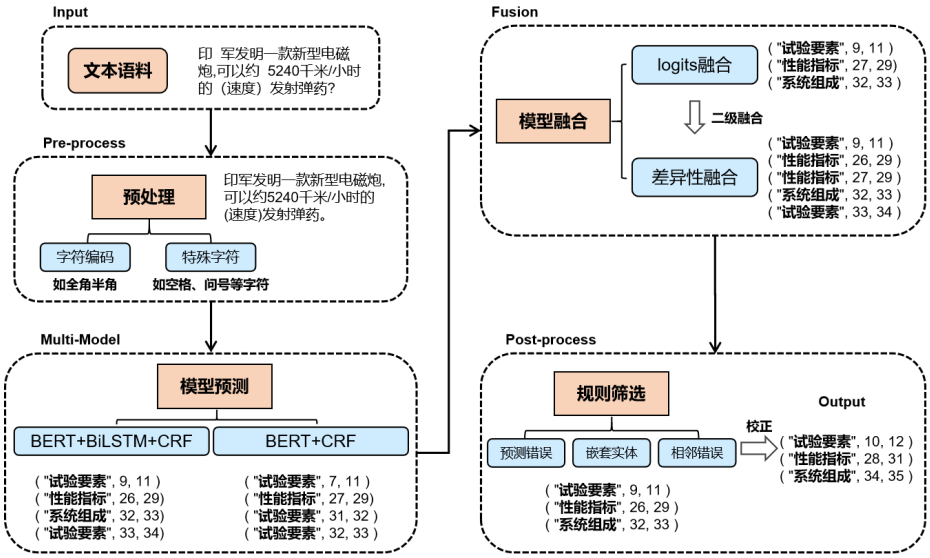

Fig. 2. 方法预测阶段流程图.

4 模型与方法

本文方法在训练阶段，首先对输入文本进行清洗和预处理，修正存在的错误标注及标注不统一数据，之后利用基础模型 BERT+CRF 和 BERT+Bi-LSTM+CRF 在评测训练数据集上进行微调训练；在预测阶段，如图 2所示，首先对输入文本进行预处理，之后利用基础模型的训练结果进行预测，为了缓解基础模型在小样本训练数据上存在的过拟合问题，分别通过 Logits 融合提高预测结果质量、通过差异性融合提高预测能力，最后进行规则筛选，通过后处理的方式剔除错误实体、嵌套实体以及相邻实体类型错误等问题，生成最终的预测结果。

4.1 数据预处理

通过统计分析发现，原始训练数据集中存在大量噪声，例如图 3(a) 所示的空格、问号等字符，且语料中存在标注错误和标注不统一问题。本文通过预处理清洗训练数据和测试数据的文本，具体包括统一字符编码、全角转半角、去除空格等噪声字符、校正实体位置，针对图 3(a) 的预处理结果如图 3(b) 所示。

4.2 基础模型

在基础模型选择阶段，我们尝试了 BERT+CRF、BERT+Bi-LSTM+CRF、BERT+Bi-GRU+CRF[2]、BERT+IDCNN+CRF[12] 等模型，最后基于预测能力选择 BERT+CRF 和 BERT+Bi-LSTM+CRF 作为最终的基础模型。

BERT+CRF. 以 BERT 作为编码层输出深层特征向量表示、以 CRF 作为下游任务层生成文本序列的实体标注结果。通过 BERT 预训练模型在 NER 训练数据上的微调训练，特征向量表示中融合了 BERT 预训练模型中包含的语言学知识和 NER 训练数据包含的任务知识。通过训练，CRF 网络层可捕捉不同标签之间的条件转移概率，以在预测过程中缓解实体标签序列中例如以 I 标签开始的逻辑错误。

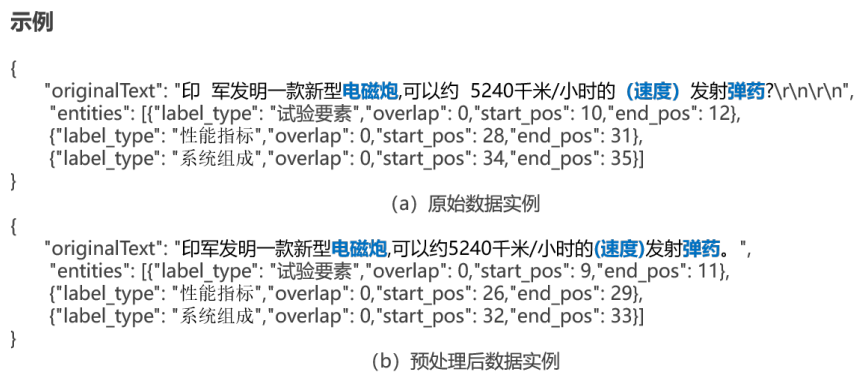

Fig. 3. 数据预处理结果示例.

BERT+Bi-LSTM+CRF. 在 BERT+CRF 模型的基础上，在 BERT 层之后、CRF 层之前增加 Bi-LSTM 层到编码层中，该 Bi-LSTM 层对 BERT 层输出的特征向量做进一步转换映射，以提取更加多样化的上下文特征。

4.3 集成融合

BERT+CRF 模型和 BERT+Bi-LSTM+CRF 模型在小样本训练数据上存在过拟合问题，针对此问题，本文提出了一种应用于预测阶段的二级融合策略以提高模型预测效果。

Logits 加权融合. 针对特定预测输入文本，基础模型的编码层 Logits 输出矩阵为 $\mathbf{M}$，其维度为 $max\_seq\_length * label\_num$，其中 max_seq_length 为模型允许输入的预测文本的最大长度、$label\_num$ 为 NER 标签数量。

设两个模型的 Logits 输出矩阵分别为 $\mathbf{M}_1$ 和 $\mathbf{M}_2$，在此基础上 Logits 加权融合结果为：

$$\mathbf{M} = \alpha\mathbf{M}_1 + \beta\mathbf{M}_2$$

其中 α 和 β 为实数，分别表示赋予 $\mathbf{M}_1$ 和 $\mathbf{M}_2$ 的加权权重。α 和 β 的选取采用经验赋值的方式。具体地，针对单模型在预测数据集上表现更好的基础模型将赋予更高的权重，提升其在融合结果中的影响。

上述两个基础模型的 Logits 加权融合过程可扩展至多个基础模型的 Logits 加权融合，而且在具体应用时可遍历所有基础模型的组合来计算 Logits 加权融合结果，并从中选出预测表现较好的组合实施融合。

差异性融合. 差异性融合针对多组预测结果，通过结果取交集、取并集或投票等方式融合。本文针对评测任务最终选择取并集作为第二级融合策略，以使多组预测结果能够相互补充，提升预测结果的召回率，但同时可能会造成融合之后的预测结果中包含文本中特定字符的多组标注结果，即嵌套实体问题。

Table 1. 基础模型参数设置

模型参数	BERT+CRF-1	BERT+CRF-2	BERT+Bi-LSTM+CRF
max_seq_length	320	300	300
$dropout\_rate$	0.4	0.5	0.3
$bi-lstm\ units$	/	/	128
$label\_num$	9		
$batch\_size$	4		
$hidden\_size$ (BERT)	1024		
$learning\_rate$	动态调整，10epoch 更新一次，序列为 5e-5、3e-5、2e-5、1e-5、5e-6 和 1e-6		
$crf\_lr\_multiplier$	CRF 层学习率为 BERT 层学习率的 100 倍		
$optimization$	Adam		
$epoch$	60		

4.4 后处理校正

如上所述，差异性融合可能引发嵌套实体问题，同时基础模型的预测结果中可能包含预测错误。本文方法通过规则校正进行后处理，以提升预测结果的精确率。

(1) 针对预测结果中存在的嵌套实体问题，本文方法保留了更长的实体，删除了被嵌套的实体。
(2) 针对预测结果中存在的相邻实体问题，本文方法考虑嵌套实体的类别，若类别相同则将其合并为一个长实体，否则保留全部实体。
(3) 删除预测结果中具有明显错误的实体，例如实体内部括号不完全、以“、”、“,”结尾的实体。

5 实验

5.1 数据集

CCKS2020 面向试验鉴定的命名实体识别评测任务要求标注试验要素、性能指标、系统组成、任务场景共计 4 类实体。官方提供 400 条训练数据。在模型训练过程中，出于模型调优以及超参数选择的需求，我们从 400 条训练数据中随机抽取 90% 样本作为训练集、10% 样本作为验证集。

5.2 实验设置

针对基础模型，本文主要训练了 2 个版本的 BERT+CRF，即 BERT+CRF-1 和 BERT+CRF-2，和 1 个版本的 BERT+Bi-LSTM+CRF。从理论上讲，引入更多新的模型有利于学习到更多样化的特征表示，从而提升模型集成融合的表达效果。各模型基本参数设置如表 1所示。

本文方法选用中文 large 版本的 roberta_wwm[1]作为基础模型的 BERT 预训练语言模型，该模型共包含 24 个 block 层，16 个多头 attention，输出 1024 维

[1] https://github.com/ymcui/Chinese-BERT-wwm

Table 2. 最终线上测试集实验结果

模型与方法	F1 score
BERT+CRF-1（model 1）	0.6860
BERT+CRF-2（model 2）	0.6833
BERT+Bi-LSTM+CRF（model 3）	0.6993
Logits-2（model 1 和 model 3）	0.7051
Logits-3（model 1、model 2、model 3）	0.7076
差异性并集融合（Logits-2+Logits-3）	0.7145
后处理规则校正（二级融合结果）	0.7203

的特征向量。模型训练学习率设置动态调整，每 10 个 *epoch* 调整一次学习率，模型共训练 60 个 *epoch*。CRF 层与 BERT 层设定不同学习率进行训练，本文方法中统一设为 BERT 层学习率的 100 倍，模型采用 Adam 优化算法进行迭代训练。Logits 融合阶段，主要采用了两个基础模型融合和三个基础模型融合两种方式，设定的权重参数分别为（1.1, 0.9）和（0.4, 0.3, 0.3）。

5.3 实验结果

为了进一步分析本文方法策略在实际应用中的有效性，我们比较了基础模型添加 Logits 融合策略、差异性并集融合策略以及后处理校正策略的线上测试集实验结果，如表 2所示。

从表 2实验结果中可以看出，在小样本命名实体识别场景下，本文提出的二级融合策略相较于基于 BERT 的基础模型有明显提升。具体相较于基础模型识别结果的 F1 评分，经过一级 Logits 融合后提高约 0.83%，经过二级差异性并集融合后提升约 1.52%，之后再经过后处理规则校正，本文方法最终线上结果的 F1 评分为 0.7203。

6 总结

本文提出了一种基于 BERT 与二级融合的小样本命名实体识别方法，可以有效缓解小样本场景下深度模型训练过程中遇到的过拟合问题，提升算法模型的实体识别效果，最终在 CCKS2020 面向试验鉴定的命名实体识别评测任务中 F1 评分为 0.7203。未来的工作中，我们将更多侧重于如何更好解决小样本场景下的实体识别问题，注重提升算法模型的准确性和泛化性。

References

1. Borthwick, A., Grishman, R.: A maximum entropy approach to named entity recognition. Ph.D. thesis, Citeseer (1999)
2. Cai, Q.: Research on chinese naming recognition model based on bert embedding. In: 2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS). pp. 1–4. IEEE (2019)
3. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. Journal of machine learning research **12**(ARTICLE), 2493–2537 (2011)

4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
5. Guan, G., Zhu, M.: New research on transfer learning model of named entity recognition. In: Journal of Physics: Conference Series. vol. 1267, p. 012017. IOP Publishing (2019)
6. Huang, Z., Xu, W., Yu, K.: Bidirectional lstm-crf models for sequence tagging. arXiv preprint arXiv:1508.01991 (2015)
7. Ju, Z., Wang, J., Zhu, F.: Named entity recognition from biomedical text using svm. In: 2011 5th international conference on bioinformatics and biomedical engineering. pp. 1–4. IEEE (2011)
8. Lei, Z., Yi, Z.: Big data analysis by infinite deep neural networks. Journal of computer research and development **53**(1), 68 (2016)
9. Li, J., Sun, A., Han, J., Li, C.: A survey on deep learning for named entity recognition. IEEE Transactions on Knowledge and Data Engineering (2020)
10. Pang, N., Qian, L., Lyu, W., Yang, J.D.: Transfer learning for scientific data chain extraction in small chemical corpus with joint bert-crf model. In: BIRNDL SIGIR. pp. 28–41 (2019)
11. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)
12. Wang, Z., Wu, Y., Lei, P., Peng, C.: Named entity recognition method of brazilian legal text based on pre-training model. In: Journal of Physics: Conference Series. vol. 1550, p. 032149. IOP Publishing (2020)
13. Yadav, V., Bethard, S.: A survey on recent advances in named entity recognition from deep learning models. arXiv preprint arXiv:1910.11470 (2019)
14. Zhang, S., Zhang, S., Wang, X.: Automatic recognition of chinese organization name based on conditional random fields. In: 2007 International Conference on Natural Language Processing and Knowledge Engineering. pp. 229–233. IEEE (2007)
15. Zhou, G., Su, J.: Named entity recognition using an hmm-based chunk tagger. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. pp. 473–480 (2002)