基于异构编码和词增强的试验鉴定命名实体识别方法

郑恒毅[1]，文瑞[2](✉)，陈曦[2](✉)，
梁思怡[1]，徐明[1](✉)

[1] 深圳大学
[2] 腾讯科技股份有限公司
{ruiwen,jasonxchen}@tencent.com
xuming@szu.edu.cn

摘要 试验鉴定命名实体识别是对被试对象进行全面考核并作出评价的第一步，是军事大数据建设的重要力量，具有极高的研究价值和应用价值。此次全国知识图谱与语义计算大会（CCKS）针对试验鉴定命名实体识别设立了一项评测任务，要求对试验要素、系统组成、任务场景以及性能指标在内的四种类型的实体进行识别。针对这项任务，本文探索了中文命名实体识别领域前沿的多种编码器结构在该任务上的效果，并针对军事领域数据稀缺的特点提出了一种简单易用的词信息增强方法，充分的利用了现有语料，在最终测试集上取得了 0.72128 的 F1 分数，排名第一。

Keywords: 试验鉴定 · 命名实体识别 · 词增强

1 引言

军事装备试验鉴定是指通过规范化的组织形式和试验活动，对被试对象进行全面考核并作出评价结论的国家最高检验行为，涵盖方法、技术、器件、武器系统、平台系统、体系、训练演习等领域，涉及面广、专业性强。近年来，自然语言理解和人工智能技术飞速发展，日趋成为推动大数据建设的重要力量。试验鉴定由于试验目的的不同、被试对象的特点、武器系统的特性，有着自身较为特殊的语言形式。

命名实体识别通常被建模为序列标注任务，其最常使用的模型可分为两类，即基于统计机器学习的模型和基于神经网络的模型。隐马尔科夫模型（HMM）[6]、条件随机场（CRF）[7] 等基于统计的模型，试图在给定输入序列的条件下以最优联合概率为目标，去推断整个标注序列。Collobert 等 [1] 最先采用基于神经网络的模型解决序列标注问题，用预训练词向量作为特征输入神经网络模型，得到单个字符所属类别的概率分布。随着神经网络

技术的不断发展，预训练语言模型与统计机器学习模型相结合的方式成为了 NER 任务的中较为稳定的基线模型，例如 BERT+CRF 等 [2]。

本文中，我们参与了 CCKS 2020 面向试验鉴定的命名实体识别任务并提出了基于异构编码和词增强的试验鉴定命名实体识别方法。以预训练语言模型 NEZHA[10] 为支撑，与多种异构的下游编码器及词增强方法相结合，得到多个异构模型。通过模型融合方法，在最终测试集取得了 0.72128 的 F1 分数。接下来，本文将对评测中所使用的方法进行介绍。

2 模型结构

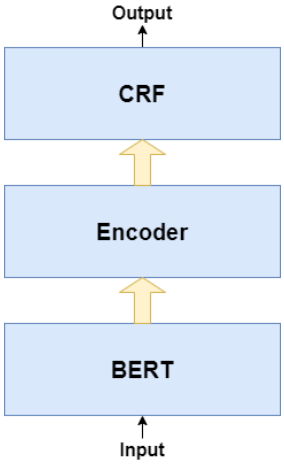

图 1. 模型结构总览

如图 1所示，本文所使用的模型结构为 BERT+Encoder+CRF 的范式。其中，在 Encoder 部分，本文尝试了一种经典的编码结构和两种前沿的编码结构，分别为 BiLSTM[4]、TENER[11] 以及 RTransfomer[9]。

2.1 双向 LSTM

LSTM 的全称为 Long Short-Term Memory，是 RNN 的一种。简单来说，LSTM 是通过对细胞状态中旧信息遗忘和新信息的记忆来传递对后续时刻计算有用的信息，同时丢弃无用的信息，并在每个时间步对隐层状态进行输出。

LSTM 的特性使其可以更好的捕捉到较长距离的依赖关系。但在一些自然语言处理任务中，除了从左到右的时序外也需要逆向的时序信息，所以 BiLSTM 应运而生。BiLSTM，即 Bi-directional Long Short-Term Memory，由前向 LSTM 与后向 LSTM 组合而成。BiLSTM 在自然语言处理任务中常被用来建模上下文信息，在命名实体识别任务中获得了极为广泛的应用 [4]。

2.2 TENER

由于传统 Transfomer 结构 [8] 在 NER 任务上表现欠佳，TENER 在其位置编码上做了一些改进，使得对于 NER 非常重要相对距离信息和方向信息能够被 Transfomer 编码为语义信息。

$$\mathcal{A}_{t,j}^{rel} = Q_t K_j^T + Q_t R_{t-j}^T + \boldsymbol{u} K_j^T + \boldsymbol{v} R_{t-j}^T \tag{1}$$

其 Attention 计算方式如公式 1所示，其中 Q_t 和 K_j 是 t 和 j 位置 token 的 query 向量和 key 向量，R_{t-j} 是正弦相对位置编码，$\boldsymbol{u}$ 和 $\boldsymbol{v}$ 是可学习参数。通过对传统 Transformer 结构中 Attention 计算方式的改进，TENER 相比于 BiLSTM 结构能够更好的编码长距离依赖关系，使得其对于长实体的捕获能力较强，在模型融合阶段丰富模型的输出结果。

2.3 RTransformer

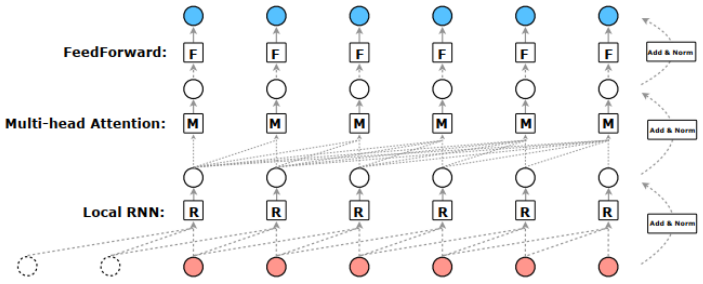

图 2. RTransformer 结构

RTransfomer 结构同样是在传统 Transfomer 结构上的改进。如图 2所示，模型在传统 Transformer 的 self-attention 计算之前加入了 Local RNN

层，相比于传统 Transformer 结构，将更多的文本局部信息编码到神经网络层中。同时，受益于 RNN 结构天然的时序信息，RTransformer 结构编码了更多的相对位置信息。对于命名实体识别任务来说，实体本身是文本中的局部内容，而文本中实体的相对位置对于类别的判断也至关重要，这体现了 RTransformer 结构的可用性。

3 词增强

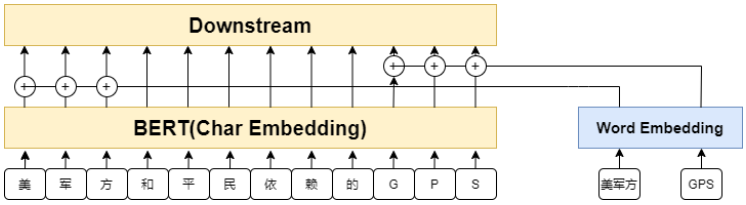

图 3. 词向量融合方法

考虑到军事试验鉴定领域语料稀少的特点，为了充分利用主办方提供的数据，本文使用词增强的方法将语料中的词汇信息融入到预训练语言模型的结果中 [3]，进一步提升下游命名实体识别任务的表现。融合词向量的方法如图 3所示。具体步骤为：

1. 使用 jieba[3] 对全部语料进行分词，载入训练集实体标注作为分词词典。
2. 利用第一步得到的词表，使用贪婪搜索的方式从前至后对文本进行匹配。
3. 采用 embedding 方式得到对应的词向量，与 BERT 输出的对应位置字向量求和。

如式 2所示，词向量与字向量求和方式为：

$$\mathcal{V}^* = \mathcal{V} + \mathcal{M} \times \mathcal{U} \tag{2}$$

其中 $\mathcal{V} \in \mathbb{R}^{l_{char} \times h}$ 为 BERT 输出的字向量，$\mathcal{U} \in \mathbb{R}^{l_{word} \times h}$ 为 Word Embedding 输出的词向量，$\mathcal{M} \in \mathbb{R}^{l_{char} \times l_{word}}$ 为字与词的位置对应矩阵。

[3] https://github.com/fxsjy/jieba

4 实验方法

4.1 数据相关

CCKS 2020 面向试验鉴定的命名实体识别任务共提供 400 例语料作为训练数据集，语料中共有包括试验要素、性能指标、系统组成以及任务场景在内的四种类型实体的标注。另外，CCKS 2020 组委会还提供了 400 例未标记的语料作为测试数据集对评估模型进行评估。出于对模型效果进行对比的需求，我们使用了 K 折交叉验证的方法，将数据集划分为训练集 320 例，验证集 80 例以及测试集 400 例。

4.2 数据分析与预处理

在制定模型之前，我们对原始语料进行了定量的分析。其中包括文本长度分布、分类别实体数量与长度分布、训练集实体出现频率等，如表 1所示。

表 1. 对原始语料的分析

文本长度分布						
Mean	Std.	Min.	25%	50%	75%	Max.
149	44	2	120	145	175	358
实体类别分布						
Abbr.	试验要素	系统组成	任务场景	性能指标	总计	
Count	1537	396	516	712	3167	

基于这些分析，做了一些预处理工作，例如：

1. 发现其中包含了一些无意义的特殊字符，例如 “\n”，“ ” 等，在对原始语料分词之前，我们将这些字符替换为 “[UNK]”。

2. 文本平均长度 149，最大长度 358，因此在输入模型前需要分割。本文设置最大输入文本长度为 256，对于长度超过 256 的文本，采用了一种动态规划的文本分割方法 [4]，使得保留最多原始文本信息的前提下冗余数据最少，见图 4。

[4] https://github.com/caishiqing/joint-mrc

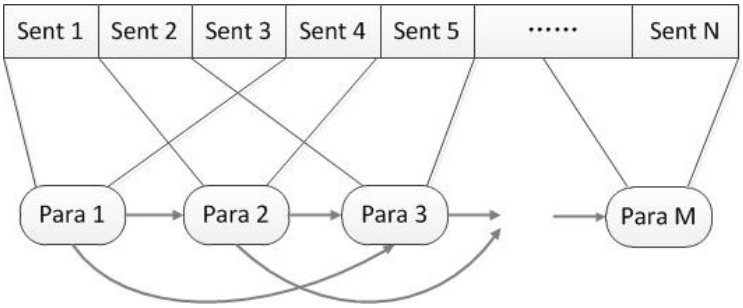

图 4. 文本分割

在得到模型结果后，对比验证集模型结果与真实标签差异，寻找模型结果存在的问题。对比不同模型结构实体抽取结果的异同，分析各模型结构的优缺点。

4.3 模型融合与后处理规则

采用节 2, 3所述模型结构预测得到结果，通过节 4.2结果分析，发现基于 BiLSTM 编码器的模型结构在各类别实体的识别上表现比较稳定，TENER 编码器对于长实体的识别效果较好，而 RTransfomer 结构能识别出一些其余两编码器不能识别出的实体。在加入词增强方法后，模型的识别效果也获得了一定的提升。根据这些特点，将以上模型的结果进行投票融合后，得到最终的识别结果。

针对融合后的结果中存在的一些较为明显的问题，我们制定了一些简单的规则进行处理：

1. 融合后去重叠：投票融合后的最终结果会有实体重叠的情况出现，重叠可分为两类，其一为边界相同，类别不同，其二为边界重叠。对于本次比赛来说，训练集的标注中是没有重叠实体出现的，而融合后的模型结果中两种重叠情况均有出现。保留重叠的一组实体中出现次数最多的那一个，将其余的舍弃，从而消除模型结果中实体重叠的现象。

2. 按照训练集数据的标注习惯，多次出现的实体，只标注一次。

4.4 实验结果

表 2为实验参数设置，优化算法为 Adam 算法 [5]。

表 2. 参数设置

Parameter	**Value**
batch_size	24
max_sequence_len	256
learning_rate	2e-5
warm_up	0.1
weight_decay	0.01

表 3. 模型实验结果对比（F1-score）

模型	综合
NEZHA+BiLSTM+CRF	0.675
NEZHA+BiLSTM+CRF+Word Augment	0.679
NEZHA+TENER+CRF+Word Augment	0.674
NEZHA+RTransfomer+CRF+Word Augment	0.677
Ensemble	**0.721**

NEZHA[10] 是一种基于 BERT 改进的中文预训练语言模型。

表 3列举出了各方法在当前数据集上的 F1-score。可以看出，BiLSTM 编码器的表现较为稳定，词增强方法对于模型性能有一定的提升。由于本次任务数据集较小，通过第 4.3节提到的模型融合方法，模型精度得到了显著的提升，这证明我们的融合策略是有效的。

5 结论

本文提出了一种简单易用的命名实体识别词增强方法，最大程度的利用了稀缺的标注数据，同时探索了多种异构编码器在试验鉴定任务上的效果，在 CCKS 2020 面向试验鉴定的命名实体识别任务中以 0.72128 的成绩排名第一。通过对识别结果的分析，我们发现由于语料规模等限制，很多实体的类别之间存在混淆，我们未来的工作将侧重如何让模型在小规模数据集中更精确的判定实体的类别，同时也将尝试一些模型蒸馏与剪枝方法，使模型能够适应实际的应用场景。

参考文献

1. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. J. Mach. Learn. Res. **12**(null), 2493 – 2537 (Nov 2011)
2. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1423, https://www.aclweb.org/anthology/N19-1423
3. Diao, S., Bai, J., Song, Y., Zhang, T., Wang, Y.: Zen: Pre-training chinese text encoder enhanced by n-gram representations (2019)
4. Huang, Z., Xu, W., Yu, K.: Bidirectional LSTM-CRF models for sequence tagging. CoRR **abs/1508.01991** (2015), http://arxiv.org/abs/1508.01991
5. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
6. McCallum, A., Freitag, D., Pereira, F.C.N.: Maximum entropy markov models for information extraction and segmentation. In: Proceedings of the Seventeenth International Conference on Machine Learning. p. 591 – 598. ICML '00, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2000)
7. McCallum, A., Li, W.: Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003. pp. 188–191 (2003), https://www.aclweb.org/anthology/W03-0430
8. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need (2017)
9. Wang, Z., Ma, Y., Liu, Z., Tang, J.: R-transformer: Recurrent neural network enhanced transformer (2019)
10. Wei, J., Ren, X., Li, X., Huang, W., Liao, Y., Wang, Y., Lin, J., Jiang, X., Chen, X., Liu, Q.: Nezha: Neural contextualized representation for chinese language understanding (2019)
11. Yan, H., Deng, B., Li, X., Qiu, X.: Tener: Adapting transformer encoder for named entity recognition (2019)