BERT 和 LSTM-CRF 在不同数据集上的对比研究

许书闻

2025 年 7 月 20 日

摘要

本报告介绍了基于 BERT 和 LSTM-CRF 模型的中英双语命名实体识别（NER）实验。实验在三个主流数据集（CoNLL-2003、WikiAnn、CoNLL-2012 OntoNotes v5）上进行了全面的性能评估，支持中文和英文。实验结果表明，BERT 模型在英文数据集上表现优异，F1 分数达到 86.01%，但在中文数据集上表现相对较差，F1 分数仅为 52.24%。LSTM-CRF 模型在中文 WikiAnn 数据集上取得了 68.66% 的 F1 分数，显示出在特定语言任务上的优势。

目录

1 引言

命名实体识别（Named Entity Recognition, NER）是自然语言处理中的一项基础任务，旨在从文本中识别和分类命名实体，如人名、地名、组织名等。随着深度学习技术的发展，基于神经网络的 NER 模型取得了显著进展。本实验旨在比较两种主流的 NER 模型架构：基于 Transformer 的 BERT 模型和基于循环神经网络的 LSTM-CRF 模型，并在多语言环境下评估其性能。

2 相关工作

2.1 传统方法

传统的 NER 方法主要基于规则和统计机器学习，如隐马尔可夫模型（HMM）和条件随机场（CRF）。这些方法依赖于手工特征工程，性能有限且更耗费人工。

2.2 深度学习方法

近年来，深度学习方法在 NER 任务上取得了突破性进展：

- **LSTM-CRF**: 结合双向 LSTM 和 CRF 层，能够捕获序列依赖关系和标签转移约束
- **BERT**: 基于 Transformer 架构的预训练语言模型，通过大规模语料预训练获得丰富的语义表示

2.3 多语言 NER

多语言 NER 面临的主要挑战包括：

- 不同语言的分词策略差异（如中文需要特殊的 tokenizer）
- 语言特定的实体类型和标注规范
- 跨语言的知识迁移

3 实验设置

3.1 数据集

本实验使用了三个主流的 NER 数据集：

表 1: 数据集比较

数据集	语言	训练样本	验证样本	实体类别	数据来源
CoNLL-2003	英文	14,987 句	3,466 句	4	路透社新闻
WikiAnn	英文	20,000 句	10,000 句	3	维基百科
WikiAnn	中文	20,000 句	10,000 句	3	维基百科
CoNLL-2012 OntoNotes v5	英文	59,924 句	8,528 句	18	多领域（新闻、对话、网络等）
CoNLL-2012 OntoNotes v5	中文	59,924 句	8,528 句	18	多领域（新闻、对话、网络等）

3.2 模型架构

3.2.1 LSTM-CRF 模型

- **嵌入层**: 100 维词嵌入，支持预训练词向量
- **LSTM 层**: 双向 LSTM，隐层维度 256
- **CRF 层**: 条件随机场，描述标签序列之间的约束关系，排除了一些不可能发生的情况，从而提高了分类精度
- **优化器**: Adam，学习率设置为 0.001

3.2.2 BERT 模型

- **预训练模型**: bert-base-cased（英文）、bert-base-chinese（中文）
- **分类头**: 线性层，输出维度为标签数量
- **优化器**: AdamW，学习率 4e-5（英文）、1e-4（中文）
- **权重衰减**: 0.01

3.3 训练参数

表 2: 训练超参数设置

参数	LSTM-CRF	BERT 英文	BERT 中文
批次大小	256	128	128
最大序列长度	128	128	128
训练轮数	30	10	10
学习率	0.001	4e-5	4e-5
嵌入维度	100	-	-
隐藏维度	256	-	-

参数解读：

- batch_size: GPU 显存充足，选择了较大的 256 和 128，且保证训练稳定性
- max_len: 选择 128 对于英文和中文句子都足够
- num_epochs: LSTM 从头训练，所以训练 30 轮；BERT 加载预训练权重，只在目标数据集上微调，所以只训练 10 轮即收敛
- learning_rate: 采用线性放缩规则，batch_size 较大时取的 learning_rate 近似线性增大。

3.4 评估指标

采用标准的 NER 评估指标：

- **精确率（Precision）**：正确识别的实体占预测实体的比例
- **召回率（Recall）**：正确识别的实体占真实实体的比例
- **F1 分数**: 精确率和召回率的调和平均
- **准确率（Accuracy）**：正确预测的 token 比例

4 实验结果

4.1 整体性能对比

表 3: 不同模型在 CoNLL-2003 数据集上的性能对比

模型	精确率	召回率	F1 分数	准确率
BERT	0.8645	0.8558	0.8601	0.9698
LSTM-CRF	0.7456	0.7151	0.7300	0.9446

表 4: 不同模型在 WikiAnn 英文数据集上的性能对比

模型	精确率	召回率	F1 分数	准确率
BERT	0.7529	0.7801	0.7663	0.9057
LSTM-CRF	0.6037	0.6212	0.6123	0.8289

表 5: 不同模型在 WikiAnn 中文数据集上的性能对比

模型	精确率	召回率	F1 分数	准确率
BERT	0.7816	0.8291	0.8047	0.9395
LSTM-CRF	0.7013	0.6726	0.6866	0.9058

表 6: 不同模型在 CoNLL-2012 OntoNotes v5 英文数据集上的性能对比

模型	精确率	召回率	F1 分数	准确率
BERT	0.8109	0.8331	0.8218	0.9737
LSTM-CRF	0.7529	0.7264	0.7394	0.9598

表 7: 不同模型在 CoNLL-2012 OntoNotes v5 中文数据集上的性能对比

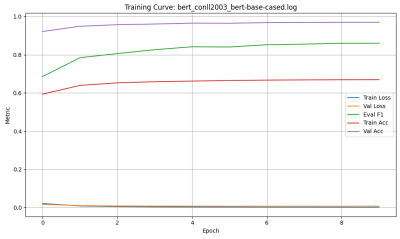
模型	精确率	召回率	F1 分数	准确率
BERT	0.5329	0.5123	0.5224	0.9199
LSTM-CRF	0.6450	0.5676	0.6038	0.9385

4.2 训练过程分析

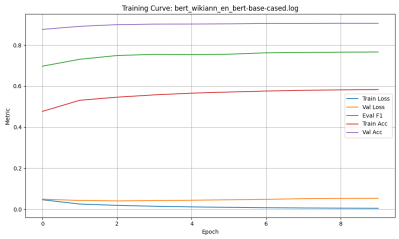
4.2.1 损失函数收敛

- **BERT 模型**: 损失函数快速收敛，在 10 个 epoch 内达到稳定状态，训练损失从 0.0210 快速下降到 0.0014
- **LSTM-CRF 模型**: 需要更多训练轮数（30 个 epoch）才能达到最佳性能，训练损失从 0.9014 逐渐下降到 0.0263
- **过拟合现象**: LSTM-CRF 在后期出现轻微过拟合，验证损失略有上升，但 F1 分数保持稳定

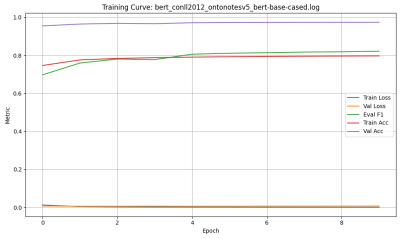
4.2.2 训练损失曲线



(a) BERT 英文 CoNLL-2003 训练曲线

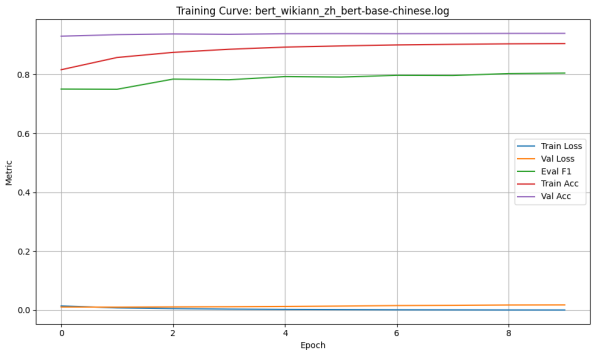


(b) BERT 英文 WikiAnn 训练曲线

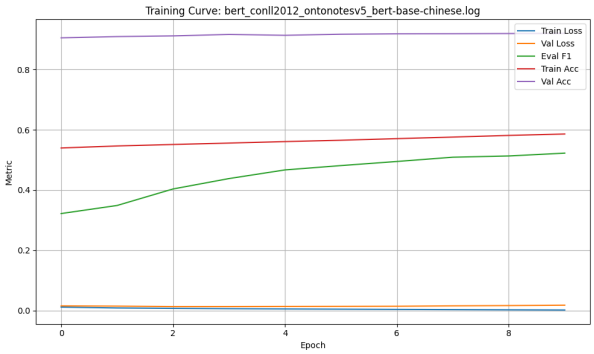


(c) BERT 英文 CoNLL-2012 训练曲线

图 1: BERT 英文模型训练损失曲线

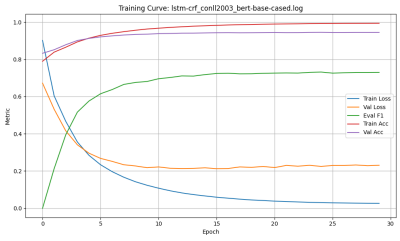


(a) BERT 中文 WikiAnn 训练曲线

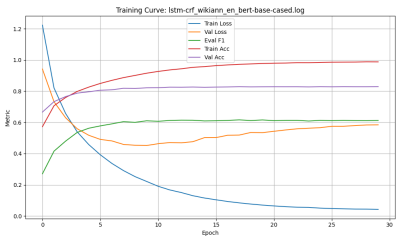


(b) BERT 中文 CoNLL-2012 训练曲线

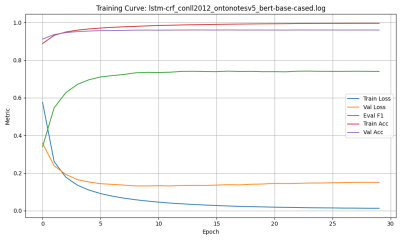
图 2: BERT 中文模型训练损失曲线



(a) LSTM 英文 CoNLL-2003 训练曲线

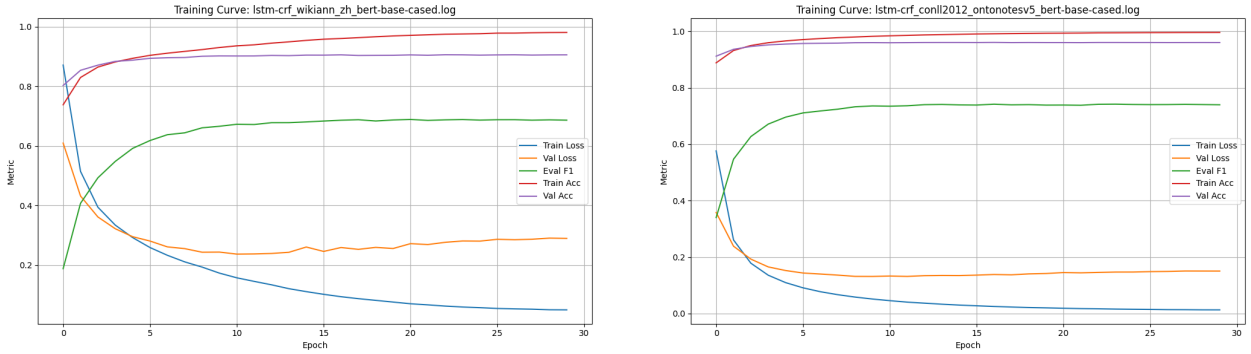


(b) LSTM 英文 WikiAnn 训练曲线



(c) LSTM 英文 CoNLL-2012 训练曲线

图 3: LSTM-CRF 英文模型训练损失曲线



(a) LSTM 中文 WikiAnn 训练曲线

(b) LSTM 中文 CoNLL-2012 训练曲线

图 4: LSTM-CRF 中文模型训练损失曲线

4.3 中英文数据集分析

4.3.1 英文数据集

- **CoNLL-2003**: BERT 模型表现最佳，F1 分数达到 86.01%，相比 LSTM-CRF 提升了 13.01 个百分点
- **WikiAnn 英文**: BERT 模型同样优于 LSTM-CRF，F1 分数为 76.63%，提升了 15.4 个百分点
- **CoNLL-2012 英文**: BERT 模型 F1 分数为 82.18%，相比 LSTM-CRF 提升了 8.24 个百分点
- **原因分析**: BERT 的预训练知识在英文任务上发挥重要作用，特别是在标准化的新闻文本上

4.3.2 中文数据集

- **WikiAnn 中文**: BERT 模型 F1 分数为 80.47%，LSTM-CRF 为 68.66%，BERT 表现更好
- **CoNLL-2012 中文**: 令人意外的是，LSTM-CRF 模型 F1 分数为 60.38%，反而优于 BERT 的 52.24%
- **挑战**: 中文分词和实体边界的复杂性，以及中文预训练模型的质量问题
- **改进空间**: 需要更好的中文预训练模型和分词策略

5 总结

5.1 BERT 模型特点

1. **预训练知识**: BERT 在大规模语料上预训练，获得了丰富的语义表示
2. **上下文理解**: Transformer 架构能够捕获长距离依赖关系
3. **迁移学习**: 预训练模型能够快速适应下游任务
4. **英文表现优异**: 在英文数据集上表现显著优于 LSTM-CRF

5.2 LSTM-CRF 模型特点

1. **序列建模**: 双向 LSTM 能够捕获序列的上下文信息
2. **标签约束**: CRF 层学习标签转移概率，提高预测一致性
3. **计算效率**: 相比 BERT，训练和推理速度更快
4. **可解释性**: 模型结构相对简单，便于理解和调试

5.3 性能差异原因

5.3.1 模型容量

- BERT 模型参数量大（110M），表达能力更强
- LSTM-CRF 参数量相对较小，在复杂任务上可能欠拟合

5.3.2 预训练策略

- BERT 通过掩码语言模型和下一句预测任务预训练
- LSTM-CRF 需要从头训练，缺乏先验知识

5.3.3 语言特定因素

- 英文 BERT 预训练模型质量较高，在英文任务上表现优异
- 中文 BERT 模型可能存在预训练质量问题，导致在复杂中文数据集上表现不佳
- 中文分词和实体边界的复杂性对模型性能影响较大