

# 基于 Transformer 的命名实体识别

## 1. 实验简介

### 1.1 实验目的

本实验旨在基于 Transformer 构建命名实体识别（NER）模型，通过对给定的训练语料进行训练，实现对测试语料的命名实体识别。

### 1.2 实验背景

命名实体识别是自然语言处理中的重要任务之一，涉及识别文本中的特定实体，如人名、地名、时间等。基于 Transformer 的模型，特别是 BERT、GPT 和 T5 等，已在该领域取得显著成效。

## 2. 数据说明

### 2.1 训练数据

- **train.txt**: 包含字（符号）序列的训练数据。
- **train\_TAG.txt**: 与 train.txt 对应的实体标签。

#### 示例

- train.txt: 人民网 1 月 1 日讯 据《纽约时报》报道，
- train\_TAG.txt: O O O B\_T I\_T I\_T I\_T O O O B\_LOC I\_LOC O O O O O O

### 2.2 验证数据

- **dev.txt**: 验证集数据。
- **dev\_TAG.txt**: 与 dev.txt 对应的实体标签。

### 2.3 测试数据

- **test.txt**: 用于测试模型。

## 3. 算法说明

### 3.1 Transformer

Transformer 是一种基于注意力机制的神经网络架构，广泛应用于自然语言处理任务。Transformer 通过并行处理输入序列，克服了 RNN 序列处理的局限性，能够更有效地捕捉长距离依赖关系。

#### 核心组件：

- **多头自注意力机制 (Multi-Head Self-Attention)**：通过多个注意力头 (head) 对输入序列进行不同子空间的投影和处理，从而增强模型捕捉不同特征的能力。
- **位置编码 (Positional Encoding)**：因为 Transformer 不具备处理序列位置信息的能力，通过加入位置编码，使模型能够识别输入序列中每个元素的位置。
- **前馈神经网络 (Feed-Forward Neural Network)**：位置编码后，通过一系列前馈神经网络进行非线性变换。
- **层归一化 (Layer Normalization) 和残差连接 (Residual Connection)**：每个子层后添加层归一化和残差连接，有助于加速训练收敛并稳定模型性能。

Transformer 架构由编码器 (Encoder) 和解码器 (Decoder) 组成，每个部分由若干个相同的层堆叠而成。编码器处理输入序列，生成隐层表示；解码器根据编码器输出和目标序列生成最终输出。

### 3.2 BERT (Bidirectional Encoder Representations from Transformers)

BERT 是基于 Transformer 编码器的预训练语言模型，专为理解双向上下文信息而设计。

#### 关键特点：

- **双向编码 (Bidirectional Encoding)**：BERT 通过同时考虑序列的左侧和右侧上下文信息，提高了对语义的理解能力。
- **预训练和微调 (Pre-training and Fine-tuning)**：BERT 通过大规模无监督预训练，然后在特定任务上进行微调，适应多种下游任务。
- **掩码语言模型 (Masked Language Model, MLM)**：在预训练过程中，随机掩盖输入序列中的部分词汇，训练模型预测被掩盖的词汇，从而捕捉上下文信息。
- **下一句预测 (Next Sentence Prediction, NSP)**：通过预测两个句子是否相邻，增强模型对句子关系的理解。

#### BERT 结构：

- **输入层**：输入由 token、segment 和 position 三部分编码组成。

- **多层 Transformer 编码器：** BERT-base 由 12 层 Transformer 编码器组成，每层包含 12 个自注意力头，隐藏层维度为 768。
- **输出层：** 经过编码器处理后，输出用于特定任务的预测。

在命名实体识别任务中，BERT 利用其强大的双向上下文编码能力，对输入序列进行编码，并预测每个 token 的标签，精确识别文本中的命名实体。

## 4. 模型设计与训练

### 4.1 标注集

基于训练数据统计得到的标签集合：

```
{'B_LOC': 0, 'B_ORG': 1, 'B_PER': 2, 'B_T': 3, 'I_LOC': 4, 'I_ORG': 5, 'I_PER': 6, 'I_T': 7, 'O': 8}
```

### 4.2 模型结构

本次实验采用基于 Transformer 的命名实体识别模型。模型设计如下：

#### 模型类型

- BERT

#### 超参数

- 层数：12
- 多头注意力数量：12
- 输入 token 数：512
- 初始字向量来源：BERT预训练模型 (bert-base-chinese)
- 向量维度：768

### 4.3 训练细节

- 训练算法：AdamW 优化器
- 学习率：5e-5
- 训练批次大小：32
- 训练轮数：100
- 损失函数：交叉熵损失

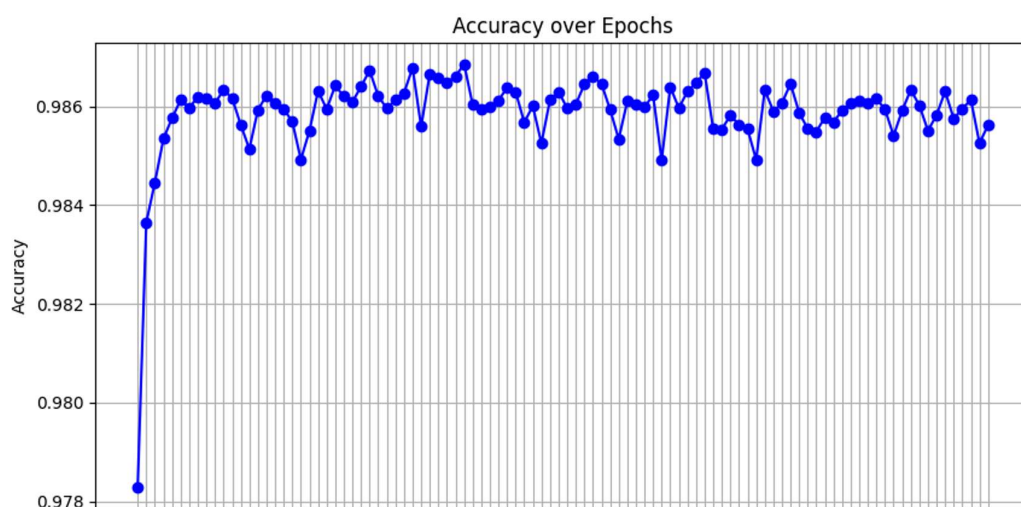
## 5. 实验结果

### 5.1 训练损失和验证集性能

由于训练集过大，每轮训练随机选择2000条数据进行。

在训练过程中，每轮记录训练损失，同时在验证集上进行测试，获得标注性能（准确率）。

#### 训练损失和准确率随时间变化的曲线图



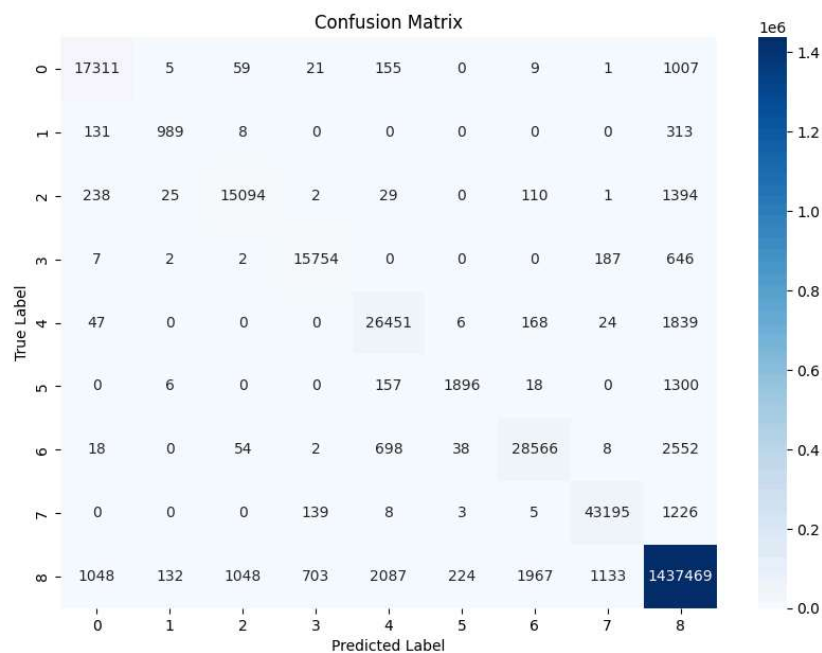
### 5.2 最佳训练轮次

选择第 38 轮作为最佳训练轮次进行测试，得到的结果为：

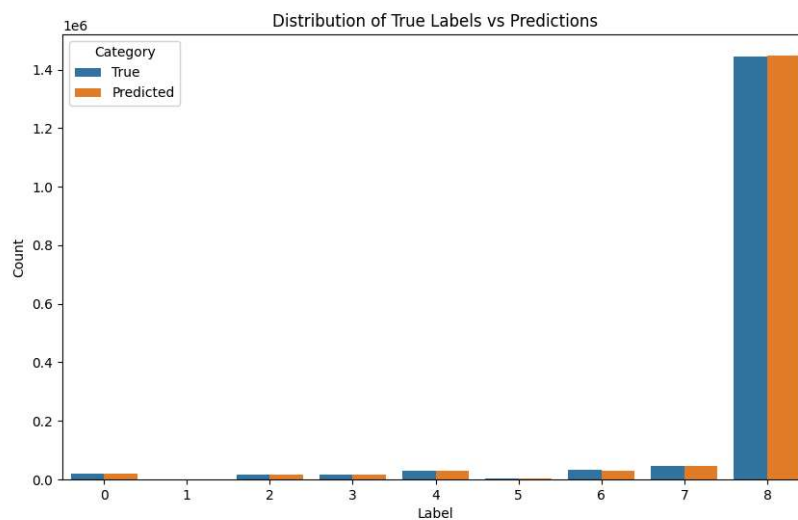
Test Loss: 0.0960, Precision: 0.9868, Recall: 0.9869, F1 Score: 0.9868

判定方式按照字符级评估而不是实体级评估。

- 结果混淆矩阵



- 整体标签分布



## 5.3 实验结论

通过实验，我们证明了基于 BERT 的命名实体识别模型在识别中文文本中的命名实体时有较好的性能。模型在测试集上取得了较高的精确率、召回率和 F1 值。

由于预训练的BERT模型在语义理解上已经非常强大，对于此类非常关注语义信息的任务来说，BERT预训练模型只需要在小数据上进行微调就可以达到很好的效果，实验中也表明在一轮训练后模型就已经在验证集上有96%以上的准确率。

本次实验使用了字符级评估，后续可以继续使用实体级评估来综合评判一下模型性能。

## 6. 参考文献

- [1] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.
- [2] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.