BERT-based sequence labeling

Zhijian Zhou
20307140081@fudan.edu.cn
Fudan NLP 实验报告

May 2023

1 背景介绍

顺序标记是自然语言处理中的一项重要任务，其目标标记是为给定的输入顺序中的每个元素分配一个标记。顺序标记在许多应用领域都有广泛的应用，包括命名实体识别、词性标记、语言义角颜色标记等。随着人们对自然语言处理技术的不断深入研究和应用，对中文文本的顺序标记任务也变得越来越多越重要。

中文是一种复杂的语言，包含丰富的词汇、语言法和语言义结构。因此，在对中文本行程序列表标记时临着一些特殊的挑战。首先先，中文单词之间没有明确的空格或分隔符，因此词边界的谜别是一个关键问题。其次，中文存在于大量的多音字、正确词和简化词等现象，导向词性和语义的准确标记更具备挑战性。此外，中文的语言结构和句法规则与其他语言也存在一定的差异，这也增加了顺序标记的复杂性。

针对这些挑战，研究者提出了各种方法和技术来改进中文的顺序标记性能。传统的基于规则的方法和统计机器学习方法已经取得了一定的成果，但随着深度学习技术的发展，基于神经网络的方法在序列标记任务上取了显着的突破。例如，长时记忆网络（LSTM）和转换器（Transformer）等模型在中文序列中标记任务中取得了优异的性能。

2 问题分析

本次实验所使用的数据集包含已标注好的文本及其对应标签。这些文本已经完成了词语的分割，实际上降低了一定的处理难度。因此，我们的主要挑战在于，在这种粗粒度的任务中，如何准确地识别并标注出每个词语的标签。

整个实验主要包含两个部分：一是词向量的编码，使其在高维空间中有更好的语义表达；二是在知道一个句子中词汇编码的基础上，对每个词语进行正确的分类。

在词向量编码环节，最初级的方法可能是使用 one-hot 向量，但这几乎无法实现，因为中文词汇的数量庞大，这将导致维度灾难。于是，我们考虑将其映射到一个 N 维的向量上，利用 torch.nn.embedding 的方法，将离散的标记（例如单词或字符）映射为连续的向量表示。然而，这种方式也存在问题，它不能表达丰富的语义，只能代表一个词汇。因此，我们考虑采用预训练的方式，首先会想到用 word2vec 方式来训练词向量。然而，word2vec 生成的词向量是静态的，而中文的语义却是非常复杂的，比如” 中国” 这个词在” 中国人民银行” 和” 中国上海” 这两个词组中的含义就是不同的，因此他们的标签也应该是不一样的。

近年来，谷歌推出的 BERT 模型在处理 NLP 下游任务中展现出了卓越的性能。BERT 模型的全称为 Bidirectional Encoder Representations from Transformers，它是一种预训练语言模型。这种模型的核心优点在于它能够理解单词在上下文中的含义，因为它采用了全方位的上下文预测机制，这是它的一项重要创新。更重要的是，BERT 在对词向量编码的处理上是动态的，这对我们的任务帮助巨大。

但同样的目前还没有人在词汇级别上对中文文本进行标注，这对于我们也是一个极大的挑战。

3 数据集分析

3.1 标签

数据集中的标签，主要有 17 类，分别是 S-GPE、S-PER、S-ORG、S-LOC、B-GPE、B-PER、B-ORG、B-LOC、E-GPE、E-PER、E-ORG、E-LOC、M-GPE、M-PER、M-ORG、M-LOC,O。

word	expected
菲律宾	S-GPE
总统	O
埃斯特拉达	S-PER
2 号	O
透过	O
马尼拉	S-GPE
当地	O
电台	O
宣布	O
说	O

表 1: 数据集展示，左边是文本，右边是标注

总结下来有五类

1. 人名（Person，PER）：人名或人物的名称。
2. 地理政治实体（Geo-Political Entity，GPE）：包括国家、城市、州或省等地理、政治和行政区划。
3. 组织（Organization，ORG）：政府、公司、机构的名称。
4. 位置（Location，LOC）：地理位置，除政治实体之外的任何地名，如山脉、河流等。
5. ”O” 用来表示一个词不属于任何命名实体类别。

我们可以明显地观察到，我们的标签数量遵循了一种长尾分布的模式。长尾分布是一种特殊的统计分布，其特点在于某些值的出现频率远高于其他值，因此在图像上呈现出一种高度倾斜，或“长尾”的形状。这种分布的主要特征是数据集的绝大部分都集中在少数几个类别上，而剩下的数据虽然分布在大量其他类别上，但每个类别的数据量相对较少。

在我们的任务中，“O”标签的出现频率远超其他标签，几乎占据了全部的数据量。然而，如果我们进一步将数据细分为 17 个类别，那么数量最少的“M-GPE”标签与“O”标签之间的数量差距就近乎达到了 3 个数量级。这种巨大的差异为我们的训练带来了显著的挑战，我们需要找到有效的策略来处理这些小样本标签的识别问题。

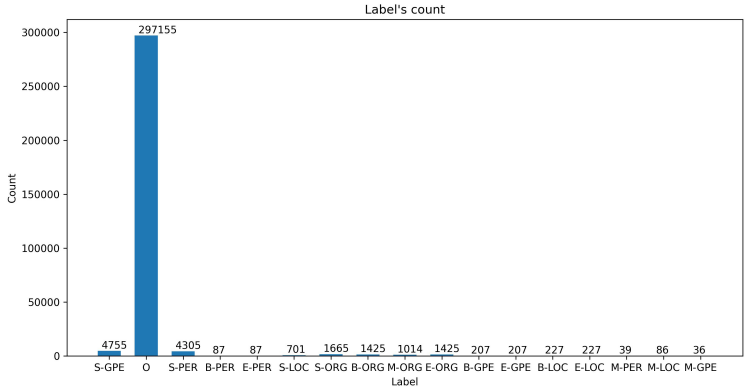

图 1: 17 类标签分布

为了简化我们的识别问题，我们决定从原来的识别十七个标签转变为仅识别五个标签，并利用规则进行进一步的划分：

- 如果是单独的一个非“O”标签, 将被添加前缀“S-”。
- 如果是连续的两个非“O”标签, 第一个词汇将被添加前缀“B-”，第二个词汇将被添加前缀“E-”。
- 如果是连续的多个非“O”标签, 第一个词汇将被添加前缀“B-”，最后一个词汇将被添加前缀“E-”，其他词汇将被添加前缀“M-”。

3.2 文本

在我们的语料库中，文本含有标点符号，这对我们进行序列标注时带来了一定的便利性。我们需要对标注进行优化，从句子级别进行处理。为了实现这个目标，我们决定对文本进行划分，将其分割为句子级别的文本。

我们主要依据的划分标准是标点符号，包括“。”、“.”和“!”。我们将根据这些符号对文本进行分割，并且为了更好地理解我们的数据，我们会检查每个句子的长度。这样的处理方式可以使我们更精确地进行序列标注，提高我们任务的准确性和有效性。

根据图 3 的数据，我们可以观察到单句文本长度的分布主要集中在 100 个词汇以内。这个长度足以覆盖我们所需的上下文文本信息。同时，我们也

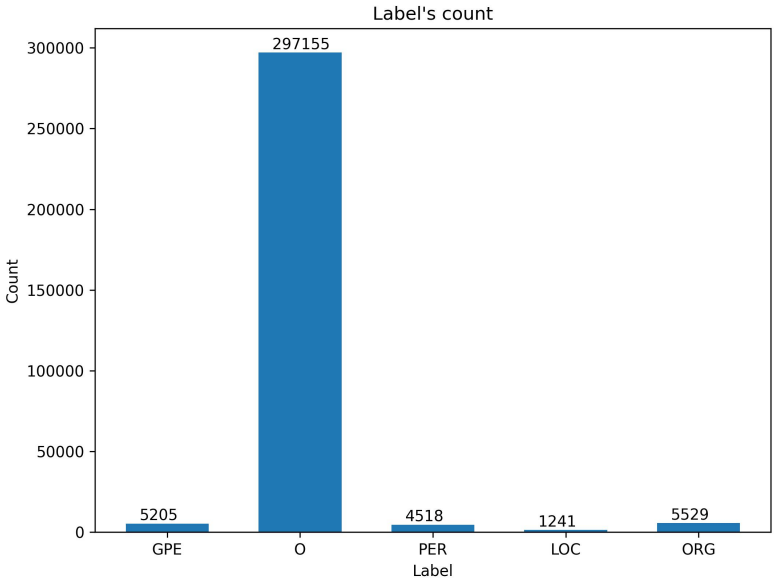

图 2: 5 类标签分布

发现文本的长度并不是越长越好。实验测试表明，依然是自然地按照标点符号来划分句子，这样的效果最佳。

4 模型

4.1 标签对齐

如图 4，我们收集的数据主要是由各个句子（Sentence）组成，而标签（Initial Label）则是对句子中每个词汇的标注。然而，Bert 的中文分词策略是基于字符级别的分词的，它将每个句子拆分为单个的中文字，而英文则是基于 Word Piece, 句子被拆分为子词。词表中不存在的词会被标记为 [UNK]，句首会被标记为 [CLS]，句尾则会被标记为 [SEP]。为了统一输入的长度，长度不足 512 的句子会被 [PAD] 填充至 512。因此，我们需要将分词后的结果与 Initial Label 进行对齐。Label 是我们的对齐结果，并且图 4 只是为了方理解，用了字符进行表达，实际操作中，Label 也会被映射到一个 0-4 作为标签，ID 中的字符实际上也会唯一对应字表里面一个编码。

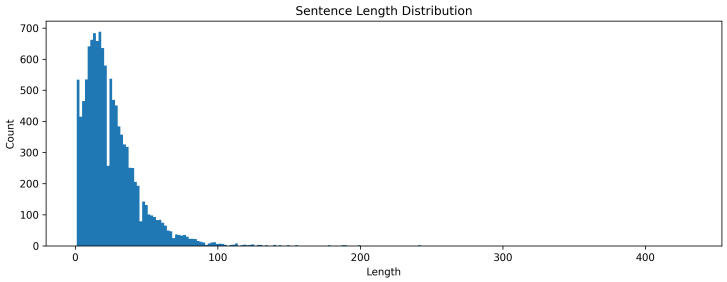

图 3: 单句长度分布

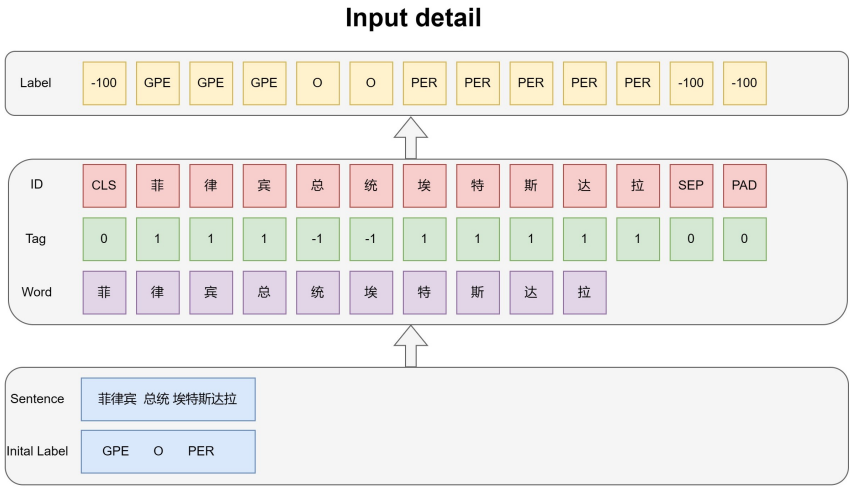

图 4: 标签对齐细节展示

其中 Tag 的目的是为了在推理过程中复原标签，将逐字标注，转化为词汇级别的标注。

4.2 模型结构

我们的模型主要由四个部分构成：输入的句子首先会经过 Tokenizer 层，然后是 Bert 层，接着是一个全连接层，最后是 Softmax 归一化层。在经过 Tokenizer 层处理后，句子会生成 Bert 所需的 Mask 和 Input ID。通过 Bert 层的处理，我们可以得到句子在字级别的编码。这些编码随后通过一个全连接层，将每个字的高维表达映射到对应五个标签的五个维度。最后，

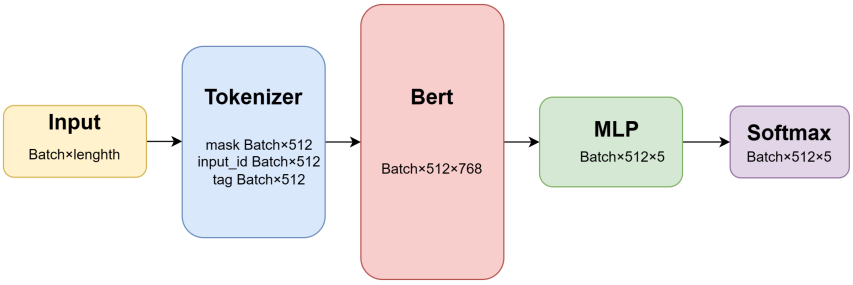

图 5: 模型结构

经过 Softmax 层，我们可以得到每个字对应每一类标签的概率。

我们也尝试过在 Bert 输出后添加更多的层数或者添加 CRF 层，但实验结果显示，这些改动并没有显著提高模型的效果，反而增加了训练的耗时和参数量。因此，根据奥卡姆剃刀原则，我们最终选择了表现良好且更为简洁的全连接层作为模型的一部分。

而相较于 BiLSTM，Bert 具有以下的优势：

- 预训练性：BERT 是一个预训练模型，这意味着它在大量无标签数据上进行预训练，学习语言的一般特性，然后在特定任务上进行微调。这使得 BERT 能够利用大量的无标签数据，而 BiLSTM 通常需要大量的标签数据。
- Transformer 结构：BERT 基于 Transformer 结构，这使得它能够更好地处理长距离依赖问题。相比之下，BiLSTM 虽然设计用来处理长距离依赖，但在实践中，BiLSTM 往往难以处理非常长的序列。
- 并行计算：由于 Transformer 的自注意力机制，BERT 可以在处理序列时进行并行计算，而 BiLSTM 由于其递归性质，处理序列时必须按顺序进行，这在处理长序列时可能会导致计算效率低下。

因此，我们最终采用图 5 的模型结构训练模型。

4.3 训练策略

4.3.1 层次化学习率

在训练模型的时候，其实我们是微调模型，因此不同层模型的学习率应该设置的不一样。

- 模型层次的不同表达能力：在 BERT 模型中，不同的层捕捉了不同级别的语义信息。一般来说，底层（靠近输入的层）更多地捕捉了局部和语法信息，而顶层（靠近输出的层）则更多地捕捉了全局和语义信息。因此，底层和顶层的学习需求可能会有所不同。

- 防止过拟合：微调时，如果所有层的学习率都设置得过高，可能会导致模型在训练数据上过拟合，尤其是当训练数据较少时。通过对底层设置较低的学习率，可以在一定程度上防止过拟合。

- 稳定预训练的知识：BERT 模型在大量无标签数据上进行了预训练，已经学习到了丰富的语言知识。如果在微调时所有层的学习率都过高，可能会破坏这些预训练的知识。通过对底层设置较低的学习率，可以保持预训练的知识更为稳定。

4.3.2 学习率指数衰减

$$Lr = Init\_lr \cdot (decay)^{\frac{global\_step}{decay\_step}}$$

由于训练语料比较多，训练轮数较多，模型结构大，因此我们采用学习率指数衰减的方式，具有以下几点优势：

- 加快训练速度：一开始用较大的学习率可以快速接近最优解，然后逐渐减小学习率，可以更快地收敛。

- 提高模型性能：逐渐减小的学习率可以在训练后期让模型在损失函数的最小值处更稳定，这可以帮助模型找到更优的解，从而提高模型的性能。

- 避免过拟合：通过降低学习率，可以降低模型在训练过程中对训练数据的过度拟合。

- 防止震荡：在训练的初期，较大的学习率可以帮助模型快速逃离不良局部最优解，而在训练的后期，较小的学习率可以减小震荡，帮助模型更稳定地收敛。

4.3.3 Focal **损失函数**

Focal Loss 函数是一种专门为解决分类问题中的类别不平衡（class imbalance）提出的损失函数，它是在交叉熵损失函数的基础上进行改进的。传统的分类损失函数，如交叉熵，对于所有的样本都给予同等的重要性。这在类别平衡的情况下是没有问题的，但在类别不平衡的情况下，这可能会导致模型对于多数类的样本过拟合，而忽视了少数类的样本。Focal Loss 的设计目标就是减少那些容易分类（被正确分类的）样本在损失函数中的权重，使模型更加关注那些难以分类的样本。Focal Loss 函数的定义为：

$$FL(p_t) = -\alpha_t * (1 - p_t)^{\gamma} * log(p_t)$$

其中，p_t 是模型预测的概率，α_t 是平衡因子，γ 是调控因子。

α_t 是类别平衡因子，可以用于调整不同类别的重要性，解决类别不平衡问题。γ 参数控制了易分类样本权重的下降速度，增大 γ 会使得模型更加关注难分类的样本。使用 Focal Loss 函数，可以有效地处理类别不平衡问题，提高模型对少数类的识别性能。对于 α_t 的设置，我们采用五个标签的分布倒数的归一化对其赋值，而 γ 的设置遵循论文里面的设置为 2。

4.3.4 **文本增强**

针对我们面临的类别严重不平衡问题，我们采取了针对性的策略：对少数类别的样本进行过采样。具体地说，我们对含有较少类别的句子实施了重复采样，以此来增强我们的文本数据集。这种方式有助于改善数据的分布，更公正地反映各个类别，从而提升模型的性能。

5 模型评估

5.1 **评估指标**-F1score

F1 得分（F1 Score）是一种用于评估分类模型准确性的指标，特别适用于处理数据不均衡的情况。它是精确率（Precision）和召回率（Recall）的调和平均数。F1 Score 的计算公式如下：

$$F1Score = 2\frac{(Precision \cdot Recall)}{(Precision + Recall)}$$

精确率（Precision）：在所有被模型预测为正类的样本中，实际为正类的样本比例。召回率（Recall）：在所有实际为正类的样本中，被模型预测为正类的样本比例。F1 Score 的值介于 0 和 1 之间。值为 1 时，表示模型的精确率和召回率都是完美的；值为 0 时，表示模型的精确率或召回率（或两者都）是零。

F1 Score 对于解决数据不均衡问题很有用，因为它同时考虑了模型的精确率和召回率。当正负样本比例严重不均衡时，模型可能会倾向于预测为数量多的类别，此时精确率可能很高，但召回率却很低。而 F1 Score 正好能够平衡这两个指标，使得模型不能偏向于预测某一类别。

5.2 试验结果

在我们的实验中，我们设定了基准设置 (Baseline)，该设置包括全局学习率设为 2e-5，批量大小（batchsize）设为 2，优化器选用 Adam，而损失函数则选择了交叉熵损失函数。此处，"BL" 代表基准设置。另外，"LRD" 代表学习率指数衰减策略，即每处理 3000 个句子后，学习率进行一次衰减，初始学习率设定为 2e-5。"FL" 代表采用 Focal Loss 损失函数，其参数设置如前文所述。"LLR" 则代表分层学习率策略，在这种策略下，全连接层的初始学习率设置为 Bert 层学习率的 5 倍。最后，"TA" 代表文本增强策略，即对类别数量偏少的句子进行重采样，以解决类别不平衡问题。

Experience	f1-score
Baseline	0.42589
BL+LRD	0.45254
BL+LRD+FL	0.48996
BL+LRD+FL+TA	0.50752
BL+LRD+FL+TA+LLR	0.54395

表 2: 试验对比结果

6 反思与不足

我们并未进行消融实验以探索每一种策略对于模型性能的贡献。消融实验通常涉及逐一移除模型中的各个组成部分，观察这对模型性能的影响，

从而更好地理解每一部分对最终结果的贡献。由于时间的限制，这一关键步骤在本次试验中被忽略，可能使我们错过了更深入理解模型性能的机会。

在模型选择上，我们采用的是基础版的 Bert 模型（bert-base-chinese）进行实验。虽然这种模型在许多任务中都已经证明了其有效性，但仍存在改进空间。许多更先进的 Bert 版本，如 Bert-large、RoBERTa、ALBERT 等，具有更大的模型规模和更复杂的架构，可能会在处理这类任务上表现得更好。未能使用这些更先进的模型是我们实验的一个局限。

此外，我们使用的数据集本身存在问题。一些句子的标注可能并不准确，这无疑会对我们的模型性能产生影响。一个高质量的、准确标注的数据集对于模型的训练和验证至关重要。我们未能在实验初期充分检查和处理这些不准确的标注，可能对我们的结果产生了不利影响。在未来的工作中，我们需要对数据集进行更严格的审查，并尝试使用更高质量的数据集。

以上就是本次实验的反思和不足，尽管我们取得了一些进步，但还有很多可以改进和探索的地方。