

A 任务简介

本次作业实现的任务为命名实体识别 (Named Entity Recognition, NER), 其目的是从给定的文本中识别出具有特定意义的实体, 如人名、地名、组织机构名等。

本次作业 NER 的实现所依托的数据为网络公开的若干罪名起诉意见书, 需要识别人 (犯罪嫌疑人、受害人)、货币 (被盗货币、物品价值、盗窃获利)、物 (被盗物品、作案工具)、时间、地点、组织机构等六大类十小类的实体。

B 代码说明

B.1 数据处理

Step 1. 读入数据后, 将句子分成字, 并为每个字做标记。如”被害人严某某”对应的标签为”[O, O, O, B-NHVI, I-NHVI, I-NHVI]”, ”O”表示该字不属于所需识别的实体分类, ”B-某类”表示这个字是某个类别的词的首字, ”I-某类”表示这个字是某个类别的非首字。

Step 2. 由于不同句子有不同长度, 故将较短的句子做 padding 补齐到与最长的句子等长, 并为每个句子使用一个等长的 mask 标记该句子的相应位置是属于原句子的部分还是 padding 补上的部分。

B.2 存储、读取模型

使用 `torch.save()` 和 `torch.load()` 实现模型存储和读取功能。

B.3 模型实现

使用预训练的 BERT 为基础, 将 BERT 所得的结果通过线性的 classifier 分类 (10 个类别的 B-和 I-, 以及 O, 共 21 类)。总参数量约 110M。

损失函数为交叉熵损失。

B.4 模型训练

训练集和验证集分割比为 8:2.

使用 Adam 优化器和学习率指数衰减的策略进行训练.

Batch Size 设为 16, 初始 Learning Rate 为 1E-5, 训练 epoch 为 20.

B.5 预测和输出评价

在 eval 模式下对测试集进行预测.

验证集的评价指标为 Accuracy 和 Micro F1-score, 取训练过程中 F1-score 最高的参数作为最终的模型参数.

C 模型评价

验证集上的表现为 Acc=97.96%, F1-score=87.08%.

以下是一些识别结果与 Ground Truth 不匹配的例子.

Text	GT	Pred
..., 窃得戚某某电动自行车上的电瓶 4 只, 计价值人民币 352 元。	戚某某 NHVI 电瓶 4 只 NASI	戚某某电动自行车上的电瓶 4 只 NASI
..., 所盗财物合计人民币 800.00 元。	所盗财物 NASI	财物 NASI
2018 年 5 月 27 日凌晨, ..., 将被害人杨某某身上的一部苹果 6S 手机 (被害人自述于 2017 年 9 月份以人民币 1300 元购买)...	2018 年 5 月 27 日凌晨 NT	2018 年 5 月 27 日凌晨 NT 2017 年 NT

表 1: 与 GT 不匹配的例子

Text	GT	Pred
..., 被告人黄某某驾驶银色助力车, 随车携带拆电瓶的扳手、启子、螺丝刀等工具, ...	扳手 NATS 启子 NATS 螺丝刀等工具 NATS	扳手、启子、螺丝刀等工具 NATS
..., 被告人皮某某携带起子、剪刀来到石门县 ** 街道老西门社区九澧怡园小区门口, ..., 其先将把摩托车推至旁边铁门楼梯间内, ..., 打火骑至临澧县 ** 镇, ...	石门县 ** 街道老西门社区九澧怡园小区门口 NS 旁边铁门楼梯间内 NS 临澧县 ** 镇 NS	石门县 ** 街道老西门社区九澧怡园小区门口 NS 铁门楼梯间内 NS 临澧县 ** 镇 NS

表 1: 与 GT 不匹配的例子 (续)

从表中例子可以看出, 模型识别的结果基本准确, 大部分无法与 GT 对应的结果, 仅在部分字词的取舍或是否分割上有所区别, 如表示地点的”旁边”二字、表示被盗物品的”所盗”二字等.

另外, 模型也存在一定的不足之处, 如当句子中出现多个时间点时, 可能无法准确识别案发时间和其他时间.

总体而言, 模型在粗粒度上表现比较良好, 在细粒度上仍有改进空间