法律咨询智能问答

TCCI

时间：2021年8月8日

赛题分析

- **任务：**给定用户问题，根据多个候选答案生成回复，属于文本生成任务。

问题	信用逾期了，银行打电话骚扰我父母，改如何处理
候选答案	1. 按照约定还款 2.报警
标准回复	你好，这种情况只能按照约定还款，如果构成骚扰可以去报警处理。

- **评价指标：**使用jieba工具分词；采用ROUGE指标（N是n-gram中的n，取值1，2）和ROUGE-L作为评价指标。

$$f-score=0.2*f\text{-}score(ROUGE\text{-}1)+0.3*f\text{-}score(ROUGE\text{-}2)+0.5*f\text{-}score(ROUGE\text{-}L)$$

赛题分析

- **赛题难点**

➢评价指标采用jieba分词，以字为粒度的模型效果不佳。

➢数据集和通用领域存在一定差距。

➢存在部分对抗样本，会对模型造成一定干扰。

对抗样本：信用逾期了，银行打电话骚扰我父母，改如何处理

方案设计

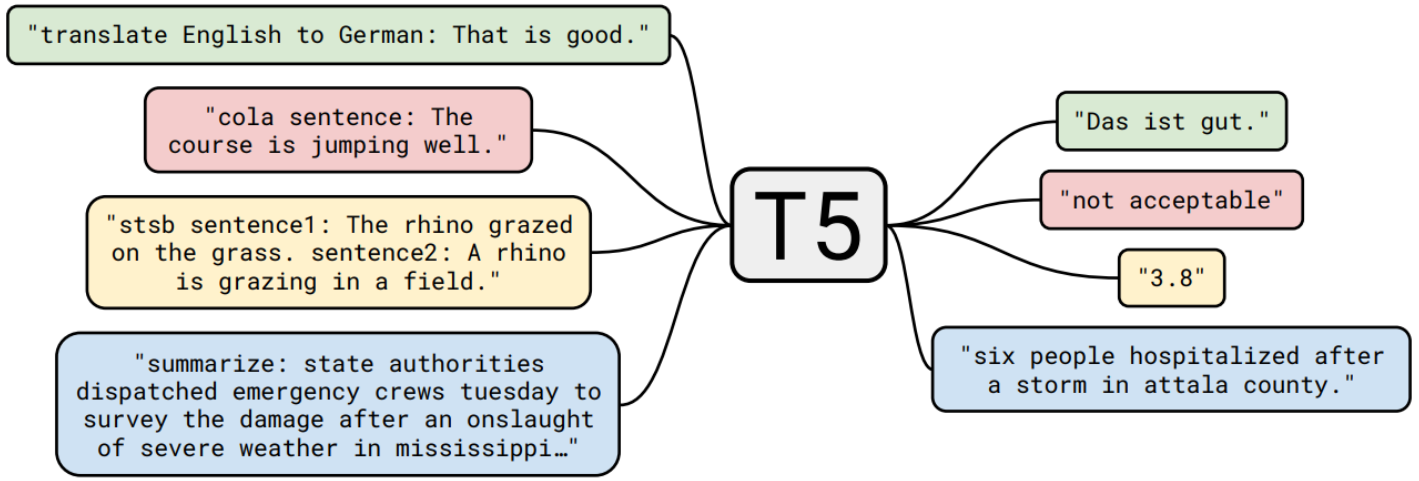
- **模型选择**

➢权重：T5-PEGASUS

➢优势：

1. 词典增加结巴分词，更贴合本次比赛任务；

2. mt5基础上预训练，在中文生成任务上性能更佳。



方案设计

- **领域预训练**

➢数据来源：罪名法务智能项目及和鲸社区法律问答数据集

问题	没有签订合同，没买保险，是在工地受伤的，别人说是他违规操作这样去走司法程序是会理亏吗	农村私人雇佣导致工伤，但是没有签订合同，都是同村的人雇佣做工，这个可以维权吗
候选答案	不理亏 您好，建议协商不成可以到法院起诉 需要把案情仔细说一遍	可以双方协商赔偿，雇佣方需要承担一定责任。 可以的，属于提供劳务者受害责任纠纷 可以要求赔偿但不是通过工伤的名义。
标准回复	无	您好，可以维权，建议直接提起诉讼，由雇佣方承担一定责任，但不算工伤。

方案设计

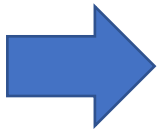
- **领域预训练**

➢训练样本构造

没有签订合同，没买保险，是在工地受伤的，别人说是他违规操作这样去走司法程序是会理亏吗

不理亏
您好，建议协商不成可以到法院起诉
需要把案情仔细说一遍

无



没有签订合同，没买保险，是在工地受伤的，别人说是他违规操作这样去走司法程序是会理亏吗

不理亏

您好，建议协商不成可以到法院起诉
需要把案情仔细说一遍

方案设计

- **领域预训练**

➢方案存在问题

1. 赛题训练集中标准答案部分字符来自于候选答案，构造的数据中会存在标准答案和候选答案无公共序列问题；
2. 构造的数据会出现无候选答案情况；
3. 给定标准答案非人工标准最佳答案，存在一定噪声。

➢采用预训练-微调方式可以在一定程度上缓解预训练数据噪声的影响

方案设计

- **对抗训练**

对模型的 embedding层添加扰动，让模型在增加扰动的情况继续向减小损失的方向进行优化，可以有效地提升模型的鲁棒性和泛化能力，尤其是在面对对抗样本的时候能够有稳定的表现。

➢FreeLB > FGM > PGD

➢预训练和微调阶段都使用FreeLB效果最好

方案设计

- **稀疏SoftMax**

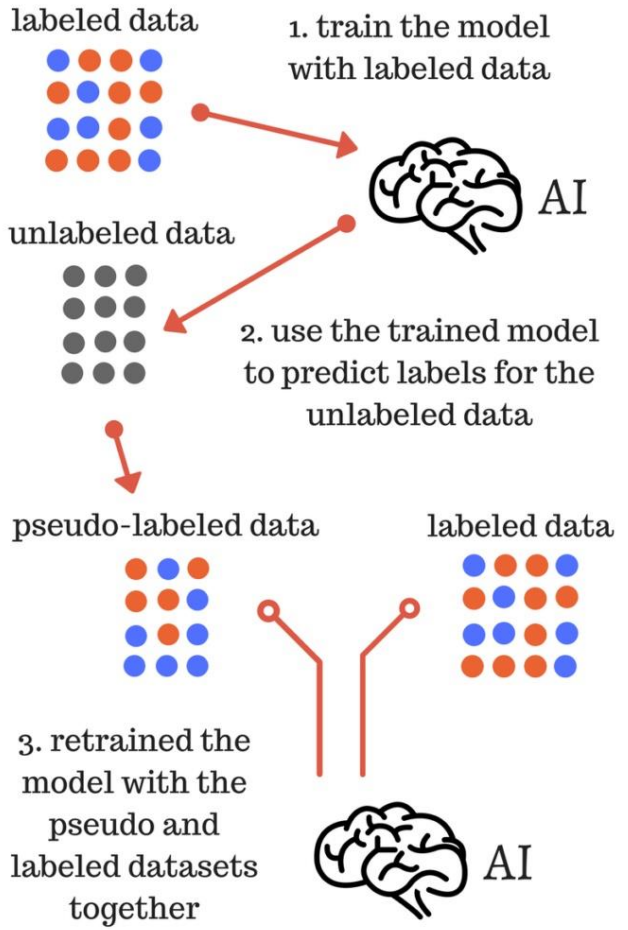
  使用稀疏化SoftMax替换SoftMax，避免SoftMax过度学习而导致过拟合。
  稀疏化即计算概率的时候，只保留前k个，后面的直接置零。

➢仅在微调阶段使用，预训练阶段使用效果下降。

方案设计

- **伪标签**

1. 使用模型集成后结果创建伪标签数据；
2. 和原有训练集混合进行五折单模训练；
3. 五折模型预测结果进行集成。



赛题总结

- **性能对比**

序号	模型	对抗	Sparse	伪标签	折数	A榜
1	T5	-	-	-	1	0.3589
2	T5	FreeLB	-	-	1	0.3616
3	预训练T5	FreeLB	-	-	1	0.3669
4	预训练T5	FreeLB	√	-	1	0.3702
5	预训练T5	FreeLB	-	-	5	0.3732
6	预训练T5	FreeLB	√	-	5	0.3757
7	预训练T5	FreeLB	√	4、5、6集成	5	0.3802

- **展望：** Copy机制、构造更优预训练数据、搭建异构模型。

THANKS