◆机器学习模型◆

基于带噪预训练的刑期预测方法*

郑　洁[1],黄　辉[2],秦永彬[2**]
(1.贵阳职业技术学院信息科学系,贵州贵阳　550081;2.贵州大学计算机科学与技术学院,贵州贵阳　550025)

摘要:刑期预测模型利用自然语言处理技术自动预测当前案件的建议刑期,对提高司法工作效率,维护司法审判的公平与公正,以及实现同案同判具有重要意义。现有的研究通常采用基于预训练语言模型的方法进行刑期预测建模,但由于存在裁判文书文本较长、专业性强及部分案由标注数据不足等问题,刑期预测任务依然具有较强的挑战性。针对上述问题,本文提出了基于带噪预训练的刑期预测方法。首先,根据刑期预测任务的特点,设计了融合罪名信息的刑期预测模型;其次,结合遮蔽语言模型(Masked Language Model,MLM)任务和自蒸馏策略减少刑期预测任务预训练数据中噪声的影响;最后,改进 RoBERTa-wwm 模型中的位置嵌入,增强模型的长文本建模能力。实验结果表明,本文提出的预训练方法能够极大地提升刑期预测任务的准确率,在小样本条件下也具有很好的表现。

关键词:刑期预测;语言模型;自蒸馏;长文本建模;预训练

中图分类号:TP391.1　　**文献标识码**:A　　**文章编号**:1005-9164(2023)01-0071-08
DOI:10.13656/j.cnki.gxkx.20230308.008

司法裁判文书中蕴含着重要的案情信息,也隐含着一定的知识价值,如何高效地从这些专业领域文本数据中抽取有价值的信息,服务于司法实践,是司法领域大数据、人工智能交叉应用的关键问题。案件判决预测是司法智能化建设中的重要一环,该技术利用自然语言处理等手段分析案件文本,预测案件的罪名、法条、刑期等判决结果,可以辅助法官对案件进行快速决策分析,是当前司法实践中需要解决的关键技术问题,有着重要的研究与应用价值。

刑期预测是案件判决预测中的一项子任务,旨在根据裁判文书案情描述对被告人应判处的刑罚时间进行预测,从而为法官或普通群众提供建议刑期。传统的刑期预测模型主要依赖于人工构建的规则和特征,并结合浅层机器学习算法进行刑期数值回归预测。随着深度学习的发展,现有的研究开始采用神经网络模型从裁判文书文本中自动学习刑期预测的特征。相比人工构建特征,神经网络模型能够从司法大数据中训练出更具泛化性的模型。目前,深度学习已

收稿日期:2022-09-18　　修回日期:2022-09-28
*国家自然科学基金项目(62066008)和贵州省科学技术基金重点项目(黔科合基础[2020]1Z055)资助。
【第一作者简介】
郑　洁(1988-),女,讲师,主要从事大数据处理、机器学习与数据挖掘研究。
【**通信作者】
秦永彬(1980-),男,教授,主要从事文本计算与认知智能、大数据治理与应用研究,E-mail:ybqin@foxmail.com。
【引用本文】
郑洁,黄辉,秦永彬.基于带噪预训练的刑期预测方法[J].广西科学,2023,30(1):71-78.
ZHENG J,HUANG H,QIN Y B. Sentence Prediction Method Based on Noisy Pretraining [J]. Guangxi Sciences,2023,30(1):71-78.

经在司法领域取得显著的成果,尤其是以预训练-微调方式建模的预训练语言模型,如 BERT[1]、RoBERTa[2]、ERNIE[3] 等。基于深度学习的刑期预测模型通常采用预训练语言模型作为语义编码器,然后添加一层线性层进行刑期计算。这种方式可以有效地将通用领域知识迁移至司法领域,并为刑期预测等具体任务带来性能上的极大提升。然而,由于法律裁判文书存在专业性强、部分案由标注数据稀缺等问题,即使使用预训练语言模型,刑期预测任务的性能相较于法条、罪名预测两个判决预测任务而言仍没有得到很大提升。其中,主要面临的挑战如下。

(1)裁判文书文本过长。裁判文书中包含当事人信息、案情描述、本院认为等重要信息,其中犯罪事实、本院认为部分记载着案件的具体事实经过和法院对案件做出的评判内容。罪名预测和法条预测两个任务一般只需要犯罪事实中关键短语或句子即可作出结论,而刑期预测由于自身的复杂性,对全局文本信息的需求更大,特别是本院认为部分。刑期预测模型往往是综合本院认为和犯罪事实两个部分内容进行预测,但大部分刑事案由的犯罪事实和本院认为的文本描述加起来都远大于 512 字符。由于计算复杂度的影响,BERT[1] 系列语言模型都在预训练阶段限制了输入 token 序列最大长度不能超过 512 字符,这导致刑期预测建模时需要对输入进行裁剪,无法获得完整的案件信息。

(2)领域性强。BERT[1] 等语言模型为了具备通用性,通常由大规模通用领域语料预训练而来。但由于司法领域数据存在大量专业词汇,与通用领域语言数据存在分布不一致的情况。在实际任务微调中,词汇分布直接从一般领域迁移到司法领域语料,预训练给下游任务带来的提升并不会像通用领域一样明显。为了将预训练模型应用于司法领域,Xiao 等[4] 在 Longformer[5] 的基础上提出了 Lawformer 预训练语言模型,但由于 Lawformer 使用的多任务联合预训练,其刑期预测采用回归方式进行建模,该方式性能较差,且多任务联合训练导致模型不能学习到刑期预测任务特有的表示。因此,本文专注于如何利用预训练提升刑期预测的性能。

(3)部分案由人工标注数据不足。首先,不同案由的案例数量分布高度不均,如刑事案件中诈骗罪、盗窃罪等案由积累的案件数量远大于非国家工作人员受贿罪、单位行贿罪等案由,这导致现有刑期预测数据集中部分案由准确率不高。其次,在现实场景中,精确标注的刑期预测数据集是不易获取的,这给小案由场景带来了一定的挑战。在长期的司法实践中,法院积累了大量的裁判文书数据,这给迁移学习带来了一定的实现途径。但经规则提取裁判文书中的刑期通常存在大量噪声,直接使用包含大量噪声的刑期标签进行继续预训练会导致模型过拟合于噪声。

针对以上问题,本文提出了基于带噪预训练的刑期预测方法。对于预训练数据中富含噪声的问题,首先,采用遮蔽语言模型(Masked Language Model,MLM)任务与带噪声的刑期预测任务相结合的预训练方式,通过 MLM 梯度引导模型过拟合于刑期标签噪声;其次,利用自蒸馏策略进行额外的平滑标签优化,进一步缓解预训练过程中的噪声影响;最后,延展 RoRERTa-wwm[6] 的绝对位置嵌入,增强模型在司法长文本上的性能。

1 相关工作

本文主要涉及司法判决预测和预训练语言模型的研究,现围绕这两个内容进行概述。

1.1 司法判决预测

随着智慧法院的提出,法律工作与人工智能技术相结合的研究越来越多,其中包括刑期预测[7]、罪名预测、法条预测等任务的司法判决预测一直备受关注。早期的研究[8-10] 采用大量法律案例训练的机器学习模型进行司法判决预测工作,但此类方法比较依赖人工构建特征,可迁移性不强。由于深度学习模型可以从数据中自动提取司法判决预测的依赖特征,研究人员开始使用神经网络方法来建模。Zhong 等[11] 针对司法判决预测子任务之间的依赖关系,提出了一种拓扑多任务学习框架 TopJudge;Chen 等[12] 提出了基于罪名的刑期预测方法,改善刑期预测性能的同时增强了模型的可解释性;Zhong 等[13] 提出了一种基于强化学习的问答模型来可视化判决预测过程,并给出可解释的判断;郭彬彬[14] 总结了司法审判流程,提出了基于 BERT 与改进 BP 神经网络的刑期预测模型。这些研究基于深度学习进行司法判决预测建模,虽然从不同的角度对方法进行了优化,但受限于基础语言模型,无法进一步提升模型性能。因此,本文聚焦于优化基础预训练语言模型在刑期预测任务上的表现。

1.2 预训练语言模型

在大规模语料上进行自监督训练得到的预训练语言模型可以将通用知识迁移至下游任务,减轻模型

对大量标注数据的依赖。早期的预训练模型[15-17]主要用于学习静态的词向量表示，但由于词向量上下文无关，导致在复杂任务中收益不大。最近，研究人员发现基于 Transformer[18]的预训练语言模型可以很好地学习到通用语料中蕴含的知识，如 GPT[18]、BERT[1]、RoBERTa[2]、DeBERTa[19]等模型，极大地提升了多项自然语言处理任务的性能。受此类研究的启发，法律智能领域开始将通用预训练语言模型迁移至法律领域进行自适应预训练。Chalkidis 等[20]使用英文法律领域语料库进行增量预训练得到 LEGAL-BERT 系列模型；Zheng 等[21]在法律案例语料上重新训练司法领域的词表，进一步提升了模型在司法领域的适应性；Xiao 等[4]为解决司法长文本问题，发布中文司法领域预训练语言模型 Lawformer，该模型在一定程度上提升了司法领域下游任务的性能，但受到噪声标注的影响，在刑期预测等任务上的表现还有待提升。本文聚焦于刑期预测任务，探索在噪声样本下的预训练策略。

2 模型构建

2.1 融合罪名信息的刑期预测模型

以往的研究将刑期预测建模视为回归问题，但由于刑期分布不均，回归损失容易拟合于刑期极值点，导致模型泛化能力变差，因此本文采用多分类的方式进行建模。根据 LAIC 2021 刑期预测数据集[7]的分布，本文将刑期分为 234 个类别，分别对应 1－234 个月的刑期时长。本文以 RoRERTa-wwm[6]为基本模型结构，在预训练模型最后一层的输出后添加一个线性分类层，形成端到端的刑期预测模型，该模型整体结构见图 1。

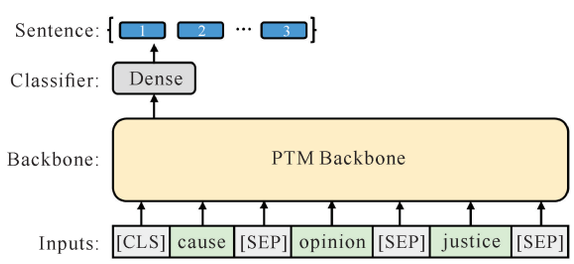

图 1 融合罪名信息的刑期预测模型结构

Fig. 1 Structure of the sentence prediction model based on the fusion of crime information

对于每一篇裁判文书，本文从中提取出本院认为、犯罪事实部分内容，并采用正则表达式进行简单的数据清洗。结果发现，案件的最终刑期不仅与案情描述内容有关联性，还与案件涉及罪名有较强的关联性，因此应在模型中融入罪名信息，增强模型对不同案由刑期的感知。为了充分利用预训练语言模型在文本数据上的优势，本文将罪名的文本描述(cause)与本院认为(opinion)、犯罪事实(justice)两个部分文本进行拼接，并与特殊标识符“[CLS]”“[SEP]”组合。其中“[CLS]”置于首位，用于表示输入文本的全局语义，“[SEP]”用于区分每个部分的输入，得到输入的字符序列之后，根据 RoBERTa 词表将字符映射为 id，输入模型进行特征提取。

RoBERTa-wwm 模型主体部分包括嵌入层以及多个堆叠的双向 Transformer 结构。首先，将输入的 id 序列转为词嵌入表示，并且与位置嵌入、句子类型嵌入相加，形成最终的特征表示。其次，蕴含罪名、本院认为以及犯罪事实 3 个部分信息的特征表示进入多重双向 Transformer 编码器，提取上下文语义，同时，由多头注意力机制组成的 Transformer 层可以在多个语义空间中对 3 个部分特征进行交互计算，增强最终语义表示的可区分性。最后，模型将“[CLS]”位置的特征向量 h_{cls} 作为输入的全局语义表示，并使用一层全连接层(Dense)和 softmax 激活函数得到刑期的概率分布：

$$p = \text{softmax}(Wh_{\text{cls}} + b),$$

式中，W 为线性层的权重参数，b 为偏置。

本文使用多分类交叉熵作为模型的损失函数：

$$L = -\sum_{i=0}^{C-1} y_i \log(p_i),$$

式中，C 为刑期类别的数量；y_i 为真实标签的独热(One-Hot)编码值，若类别为 i，则 $y_i = 1$，否则为 0；p_i 为类别 i 的预测概率值。

2.2 带噪预训练

有效的预训练可以提升模型在下游任务微调的性能。本文从中国裁判文书网收集了部分与 LAIC 2021 刑期预测数据集案由一致的案件，并使用规则提取的方式从中提取本院认为、犯罪事实及判决刑期 3 个部分信息，最终得到 20 万篇文档。同时，从 CAIL 2018[22]数据集筛选出刑期为 1－234 个月的数据，并剔除部分质量较差的裁判文书，整理出 280 万篇文档。但实验发现，规则提取的刑期标注信息中包含大量的噪声，并且 CAIL 2018[22]数据集同样存在错标数据，导致预训练后的模型在刑期预测任务上微调性能提升不大。如果对标注结果进行清洗，则会耗费大量的人力。此外，由于标注错误样本在数据中占比较大，如果直接过滤则会损失大量有效信息。为了

充分利用大规模的噪声数据,本文在预训练阶段添加 MLM 作为辅助任务,通过构造的标准自监督标签引导模型正确训练,避免模型往错误梯度的方向优化。

预训练的模型主体结构和 2.1 节中的一致,只是增加了一个 MLM 预训练头。MLM 预训练使用了和 RoBERTa[2] 一致的方式,将输入的罪名、本院认为、犯罪事实文本进行随机遮蔽,即以 15%的概率将文本中的字符随机替换为"[MASK]",然后让模型预测遮蔽掉的字符。MLM 任务使用和刑期预测任务一致的交叉熵损失进行训练:

$$L_{\text{mlm}} = -\sum_{i=0}^{V-1} y_i^{\text{mask}} \log(p_i^{\text{mask}}),$$

式中,V 为模型的词表大小,y_i^{mask} 是遮蔽字符的标签,p_i^{mask} 表示模型预测的概率。

刑期预训练的细节和 2.1 节一致,采用分类方式进行训练,预训练的基本损失为

$$L_{\text{base}} = L_{\text{term}} + L_{\text{mlm}},$$

式中,L_{term} 为刑期任务损失。

为了缓解刑期噪声对模型的干扰,本文在预训练阶段引入自蒸馏策略,让模型从教师模型产生的软标签中学习,缓解噪声的影响。具体地,初始化一个与基础模型参数一致的教师模型,在之后的训练过程中教师模型不断演化。本文使用基础模型权重的指数移动平均(Exponential Moving Average,EMA)对教师模型进行更新:

$$v_t = \beta v_{t-1} + (1-\beta)\theta_t,$$

式中,θ_t 为 t 时刻的基础模型参数,v_t 为前 t 个时刻的移动平均参数,β 为加权参数,本文设置为 0.999。采用 EMA 方式更新教师模型参数,可以修正当前基础模型因为噪声样本带来的偏差。

在训练过程中,本文在 L_{base} 的基础上增加基础模型和教师模型的 MLM、刑期预测概率分布之间的 KL 散度损失,以此对模型的优化进行约束。蒸馏损失可以定义为

$$L_{\text{distill}} = D_{\text{KL}}(p^{\text{mask}} \| q^{\text{mask}}) + D_{\text{KL}}(p^{\text{term}} \| q^{\text{term}}),$$

式中,D_{KL} 表示 KL 散度计算函数,p^{mask} 和 p^{term} 分别为基础模型预测 MLM 和刑期预测的概率分布,q^{mask} 是 mlm 的概率分布,q^{term} 为刑期预测的概率分布。最终,增加蒸馏后的模型训练损失如下:

$$L = (1-\alpha)L_{\text{base}} + \alpha L_{\text{distill}},$$

式中,α 为基础损失和蒸馏损失的调节参数,本文将其设置为 0.4。

2.3 长文本建模

LAIC 2021 刑期预测数据集的文本长度分布见图 2。从图 2 可以看出,刑期预测文本的长度主要集中在 300－1 500 字符,这给基于深度学习的方法带来一定的挑战。Xiao 等[4] 将 Longformer 迁移到司法领域,试图解决裁判文书文本过长的问题。然而在实际使用中为了增强全局注意力,通常会将 Longformer 注意力窗口设置为 512,这导致 Longformer 长度未达到 4 096 字符时计算耗时远高于全注意力机制,并且同等长度下性能低于 BERT、RoBERTa 等基于全注意力机制的模型。因此,本文基于 RoBERTa-wwm,对模型的绝对位置嵌入进行延伸,提升模型捕获长文本上下文语义信息的能力。

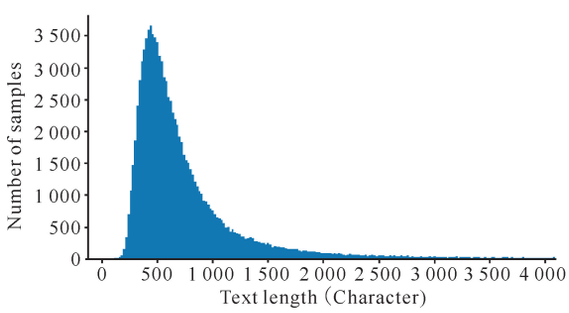

图 2 LAIC 2021 数据集文本长度分布

Fig. 2 Text length distribution of LAIC 2021 dataset

原有 RoBERTa-wwm[6] 受限于仅有 512 个位置嵌入而无法对过长文本建模,本文根据图 2 中的长度分布,将最大长度确定为 1 024 字符。为了充分利用已经预训练过的绝对位置嵌入参数,本文使用原有位置嵌入初始化扩充后嵌入的前 512 个位置特征,而对于后 512 个位置的特征,本文采用位置累加的方式进行初始化:

$$\boldsymbol{w}_n = \boldsymbol{w}_{511} + \boldsymbol{w}_{n-511},$$

式中,$\boldsymbol{w}_n$ 表示位置 id 为 n 的位置特征向量。

3 验证实验

3.1 数据集

本文将在 LAIC 2021 刑期预测数据集上验证模型的有效性。LAIC 2021 刑期预测数据集是中国司法大数据研究院承办的第四届"中国法研杯"刑期预测赛道的评测数据集,包含 40 余万篇裁判文书的犯罪事实、本院认为、刑期、年份、省份等信息。本文使用其中开源的 98 957 篇标注数据进行评测,并按 9∶1 切分训练集和测试集,切分过程中保证训练、测试数据的刑期分布一致。

LAIC 2021 数据集不同案由样本分布见图 3,共

计 50 种案由。从图 3 可以看出，不同案由的标注样本分布极度不均衡，样本主要集中在个别案由，大部分案由样本量严重不足。因此，如何解决小样本案由是刑期预测任务中的一个重点，这在具体的司法智能化实践中也有很好的应用价值。

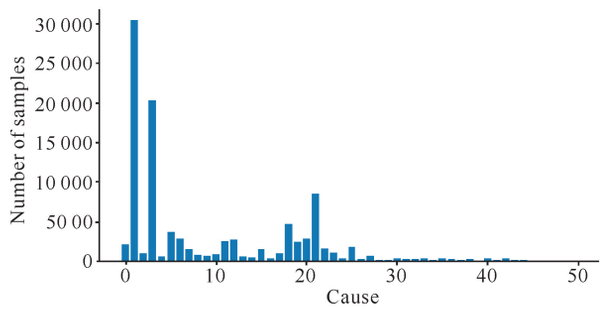

图 3　LAIC 2021 数据集不同案由样本分布

Fig. 3　Sample distribution of different cases in LAIC 2021 dataset

3.2　评价指标

本文采用和 LAIC 2021 刑期预测赛道一致的评价指标，给定样本的预测刑期 $y_i^{'}$ 和真实刑期 y_i，当前样本的分数为

$$\text{score}_i=\begin{cases}1, & |y_i-y_i^{'}|\leqslant 0.25y_i\\ 0, & |y_i-y_i^{'}|>0.25y_i\end{cases},$$

即预测刑期和真实刑期的偏离程度≤25%视为正确，否则视为错误。测试集最终准确率为所有样本分数的均值。

3.3　参数设置

本文使用 RoBERTa-wwm[6] 作为基础预训练语言模型，权重使用 base 版本。具体参数设置如表 1 所示。

表 1　参数设置

Table 1　Parameter setting

参数名称 Parameter name	预训练 Pretraining	微调 Fine-tuning
Mask probability	0.15	-
Learning rate (Lr)	5e-5	2e-5
Weight decay	0.01	0.01
Lr schedule	Linear	Linear
Warm up ratio	0.1	0.1
Batch size	4	32(16)
Gradient accumulation	32	1(2)
Epoch	1	3
Optimizer	AdamW	AdamW
Random seed	42	42

Note: "-" means that the parameter does not exist

预训练总的批次大小为 4 乘以梯度累计次数 32，即实际预训练批次大小为 128。由于文本长度扩充至 1 024 字符后预训练速度较慢，因此本文采用二阶段的预训练方式，即先进行 512 字符长度的预训练，然后使用该权重初始化继续进行 1 024 字符长度的预训练，两个阶段训练轮数均为 1。微调一列中括号外对应 512 字符长度时微调的参数，括号内对应 1 024 字符长度时的参数。

3.4　结果与分析

为了全方位对本文方法进行评估，在 LAIC 2021 刑期预测数据集设计了 3 组对比实验。实验 A 将本文方法和现有的预训练语言模型进行对比，证明本文方法的有效性；实验 B 对本文提出的预训练策略进行消融实验；实验 C 提取各个案由不同比例数据进行微调，测试本文方法在小样本条件下的表现。

3.4.1　实验 A：模型对比实验

为了验证本文提出的模型与预训练方法的有效性，与通用领域预训练模型 BERT[1]、MacBERT[6]、RoBERTa-wwm[6] 及司法领域预训练语言模型 LegalRoBERTa[4]、legal-ELECTRA[6]、Lawformer[4] 进行对比实验，实验权重均为 base。模型结构均为 2.1 节提出的融合罪名信息的刑期预测模型，大部分参数设置一致，而 ELECTRA 根据文献建议将其学习率设置为 3e-4，并使用了文献中的分层学习率。对比实验结果见表 2。

表 2　模型对比结果

Table 2　Comparison results of model

模型 Model	文本长度(字符) Text length (character)	准确率(%) Accuracy (%)
BERT	512	64.38
MacBERT	512	64.49
RoBERTa-wwm	512	65.12
LegalRoBERTa	512	66.03
legal-ELECTRA	512	66.53
Lawformer	512	65.63
Lawformer	1 024	66.76
This study	512	69.76
This study	1 024	72.23

从表 2 可以看出，3 个通用领域预训练语言模型的准确率接近，且都低于司法领域预训练语言模型。这说明司法领域数据与通用领域数据之间存在一定差距，通过在司法领域继续预训练可以消除这个差

距,提升模型在垂直领域的效果。对比 LegalRoBERTa 和 Lawformer 可以看出,在同等长度下采用滑窗注意力机制的 Lawformer 效果不如使用全注意力机制的 LegalRoBERTa,但 Lawformer 可以支持更长的文本,使用 1 024 字符长度后准确率比 LegalRoBERTa 高出 0.73 个百分点。这说明刑期预测任务对信息的需求很大,更长的文本能够补充更多的判决信息。本文针对刑期预测任务进行预训练方法研究,相比现有预训练模型 Lawformer (1 024 字符)提升了 5.47 个百分点。这表明本文提出的预训练策略更加适应于刑期预测任务,并且可以改善模型在富含噪声的预训练过程中的鲁棒性。

3.4.2 实验 B:预训练策略对比实验

为了进一步分析本文提出的预训练策略,本实验在 RoBERTa-wwm 模型的基础上,对不同策略组合进行对比实验,结果如表 3 所示。

表 3 预训练策略对比结果

Table 3 Comparison results of pretraining strategy

策略 Strategy	预训练长度(字符) Length of pretraining (character)	微调长度(字符) Length of fine-tuning (character)	准确率(%) Accuracy (%)
RoBERTa-wwm	512	512	63.92
RoBERTa-wwm + Crime	512	512	64.41
RoBERTa - wwm + Crime + MLM	512	512	66.34
RoBERTa-wwm + Crime + TASK	512	512	67.06
RoBERTa - wwm + Crime + MLM + TASK	512	512	68.52
RoBERTa - wwm + Crime + MLM + TASK + SD	512	512	69.76
RoBERTa - wwm + Crime + MLM + TASK + SD	512	1 024	70.73
RoBERTa - wwm + Crime + MLM + TASK + SD	1 024	1 024	72.23

Note:Crime indicates whether to add crime information, MLM indicates adding MLM pre-training task, TASK indicates using sentence label pre-training, SD indicates self-distillation

从表 3 可以看出:①增加罪名信息后刑期预测准确率提升了 0.49 个百分点,说明罪名信息可以给模型添加额外的审判知识。这个结果和具体司法经验也是一致的,如故意杀人罪在刑法上判刑高于盗窃罪。②使用 MLM 预训练任务在司法领域继续预训练可以提升 1.93 个百分点的准确率,这和现有的司法预训练语言模型性能基本一致。③即使标签存在噪声,刑期任务预训练依然能够比 MLM 任务带来更大的提升,这说明预训练任务越接近下游任务越有效,同时也说明预训练-微调范式可以缓解数据噪声的影响。④结合两个预训练任务比单独使用刑期预训练提升 1.46 个百分点,说明 MLM 损失产生的梯度可以避免模型因错误刑期标签导致的震荡,缓解噪声的影响。⑤增加自蒸馏策略后下游任务准确率提升了 1.24 个百分点,说明采用指数移动平均参数作为教师模型可以进一步缓解当前错误样本对模型的影响,并且软标签相比硬标签信息量更大,也能给预训练带来一定增益。⑥直接对 512 字符长度预训练的位置嵌入进行扩充,使用 1 024 字符长度微调提升准确率 0.97 个百分点,这说明长文本可以为刑期预测提供更充足的信息,也证明了本文提出的位置嵌入初始化方式的有效性。⑦继续使用 1 024 字符长度预训练,相比 512 字符长度预训练权重提升 1.5 个百分点,说明预训练后的位置嵌入能够更清楚地捕捉到文本的时序关系。另外,使用更长文本可以加快预训练的收敛。

3.4.3 实验 C:小样本对比实验

实验对训练集中每一个案由的样本进行采样,随机从每个案由的全部样本中抽取 10% - 100%的样本量,然后对比 LegalRoBERTa、Lawformer 和本文预训练权重在不同数量训练样本下的性能。本实验使用 512 字符长度对模型进行微调,结果如图 4 所示。

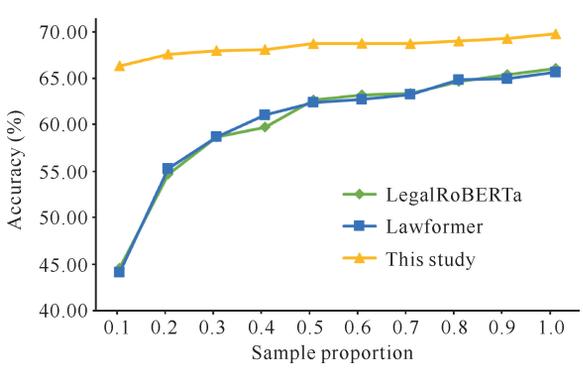

图 4 模型在不同样本量条件下的性能对比

Fig. 4 Performance comparison of models under different sample sizes

实验结果表明,使用本文预训练策略得到的权重在小样本条件下表现稳定,并且仅用 10%的标注样本量就超过了其他两种权重的全量数据性能。这说明在大规模噪声数据集上学习到的刑期预测知识可以迁移至下游任务中,帮助模型在小样本情况下快速

拟合。在实际的司法智能化实践中，可以使用本文提出的方法提升模型在案件不足的案由上的性能表现。

4 结论

本文针对司法领域预训练任务中数据富含噪声、司法文本过长等问题，提出了一种基于带噪预训练的刑期预测方法。首先，为了验证本文预训练方法的有效性，提出了一种融合罪名信息的刑期预测模型作为基础模型；其次，采用 MLM 自监督任务缓解刑期任务预训练中噪声的影响，并使用自蒸馏策略提升预训练模型的鲁棒性；最后，改进 RoBERTa-wwm 的位置嵌入，增强模型对长文本的建模能力。本文的方法相比现有的司法预训练语言模型提升刑期预测准确率 5.47 个百分点，并且在小样本情况下表现优异。在未来的工作中将继续对司法领域预训练语言模型进行探索，引入更多的司法判决预训练任务，使模型能够在多种判决任务中同时具备优秀的性能表现。

参考文献

[1] DEVLIN J，CHANG M W，LEE K，et al. BERT：pre-training of deep bidirectional transformers for language understanding [C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics：Human Language Technologies，Volume 1. Seattle，USA：Association for Computational Linguistics，2019：4171-4186.

[2] LIU Z，LIN W，SHI Y，et al. A robustly optimized BERT pre-training approach with post-training [C]//China National Conference on Chinese Computational Linguistics. Hohhot，China：Chinese Computational Linguistics，2021：1218-1227.

[3] ZHANG Z Y，HAN X，LIU Z Y，et al. ERNIE：enhanced language representation with informative entities [Z/OL]. (2019-05-17)[2022-09-18]. https://arxiv.org/abs/1905.07129v1.

[4] XIAO C，HU X，LIU Z，et al. Lawformer：a pre-trained language model for chinese legal long documents [Z/OL]. (2021-05-09)[2022-09-18]. https://arxiv.org/abs/2105.03887.

[5] BELTAGY I，PETERS M E，COHAN A. Longformer：the long-document transformer [Z/OL]. (2020-04-10)[2022-09-18]. https://arxiv.org/abs/2004.05150.

[6] CUI Y M，CHE W X，LIU T，et al. Revisiting pre-trained models for chinese natural language processing [Z/OL]. (2020-04-29)[2022-09-18]. https://arxiv.org/abs/2004.13922.

[7] 中国司法大数据研究院. LAIC 2021 数据集(刑期预测)，第四届中国法研杯司法人工智能挑战赛[EB/OL]. [2022-09-18]. http://data.court.gov.cn/pages/laic2021.html.

[8] LIU C L，HSIEH C D. Exploring phrase-based classification of judicial documents for criminal charges in chinese [C]//International Symposium on Methodologies for Intelligent Systems. Berlin，Heidelberg：Springer，2006：681-690.

[9] ALETRAS N，TSARAPATSANIS D，PREOTIUC P D，et al. Predicting judicial decisions of the european court of human rights：a natural language processing perspective [J]. PeerJ Computer Science，2016，2：e93. DOI：10.7717/peerJ-cs.93.

[10] SULEA O M，ZAMPIERI M，VELA M，et al. Predicting the law area and decisions of French supreme court cases [Z/OL]. (2017-08-04)[2022-09-18]. https://arxiv.org/abs/1708.01681.

[11] ZHONG H，GUO Z，TU C，et al. Legal judgment prediction via topological learning [C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels，Belgium：Association for Computational Linguistics，2018：3540-3549.

[12] CHEN H，CAI D，DAI W，et al. Charge-based prison term prediction with deep gating network [Z/OL]. (2019-08-30)[2022-09-18]. https://arxiv.org/abs/1908.11521.

[13] ZHONG H，WANG Y，TU C，et al. Iteratively questioning and answering for interpretable legal judgment prediction [C]//Proceedings of the AAAI Conference on Artificial Intelligence. New York，USA：AAAI，2020，34(1)：1250-1257. DOI：10.1609/aaai.v34i01.5479.

[14] 郭彬彬. 基于 BERT 与改进 BP 神经网络的盗窃案刑期预测方法研究[J]. 软件工程，2022，25(2)：6-9.

[15] MIKOLOV T，YIH W，ZWEIG G. Linguistic regularities in continuous space word representations [C]//Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics. Atlanta，Georgia，USA：Association for Computational Linguistics，2013：746-751.

[16] PENNINGTON J，SOCHER R，MANNING C D. Glove：global vectors for word representation [C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha，Qatar：Association for Computational Linguis-

tics,2014:1532-1543.

[17] VASWANI A,SHAZEER N,PARMAR N,et al. Attention is all you need [C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach,CA,USA:Curran Associates Inc.,2017:6000-6010.

[18] RADFORD A,NARASIMHAN K,SALIMANS T,et al. Improving language understanding by generative pre-training [Z/OL]. (2018-06-11)[2022-09-18]. http://www.nlpir.org/wordpress/2019/06/16/improving-language-understanding-by-generative-pre-training.

[19] HE P,LIU X,GAO J,et al. DeBERTa:decoding-enhanced bert with disentangled attention [Z/OL]. (2020-06-05)[2022-09-18]. https://arxiv.org/abs/2006.03654.

[20] CHALKIDIS I,FERGADIOTIS M,MALAKASIOTIS P,et al. LEGAL-BERT:the muppets straight out of law school [Z/OL]. (2020-10-06)[2022-09-18]. https://arxiv.org/abs/2010.02559v1.

[21] ZHENG L,GUHA N,ANDERSON B R,et al. When does pretraining help? Assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings [Z/OL]. (2021-04-18)[2022-09-18]. https://arxiv.org/abs/2104.08671v3.

[22] XIAO C,ZHONG H,GUO Z,et al. CAIL2018:a large-scale legal dataset for judgment prediction [Z/OL]. (2018-07-04)[2022-09-18]. https://arxiv.org/abs/1807.02478.

Sentence Prediction Method Based on Noisy Pretraining

ZHENG Jie[1],HUANG Hui[2],QIN Yongbin[2**]

(1. Department of Information Science,Guiyang Vocational and Technical College,Guiyang,Guizhou,550081,China;2. College of Computer Science and Technology,Guizhou University,Guiyang,Guizhou,550025,China)

Abstract: The sentence prediction model uses natural language processing technology to automatically predict the recommended sentence of the current case,which is of great significance to improve the efficiency of judicial work,maintain the fairness and justice of judicial trial,and realize the same sentence in the same case. The existing studies usually adopt the method based on pre-training language model to model the sentence prediction. However,due to the problems of long text of judgment documents,strong professionalism,and insufficient labeling data for some cases,the task of sentence prediction is still quite challenging. In view of the above problems,this paper proposes a sentence prediction method based on noisy pre-training. Firstly,according to the characteristics of sentence prediction task,a sentence prediction model integrating crime information is designed. Secondly,the influence of noise in the pre-training data of sentence prediction task is alleviated by combining the Masked Language Model (MLM) task and the self-distillation strategy. Finally,the position embedding in the RoBERTa-wwm model is improved to enhance the long text modeling ability of the model. The experimental results show that the pre-training method proposed in this paper can greatly improve the accuracy of the sentence prediction task,and also has good performance under small sample conditions.

Key words:sentence prediction;language model;self-distillation;long text modeling;pretrain

责任编辑:陆媛峰