◆自然语言理解◆

基于机器阅读理解的中文司法实体识别优化策略研究*

余俊晖[1,2],陈艳平[1,2**],秦永彬[1,2],黄　辉[1,2]
(1.公共大数据国家重点实验室,贵州贵阳　550025;2.贵州大学计算机科学与技术学院,贵州贵阳　550025)

摘要:针对中文司法领域信息抽取数据集中实体专业性较强、现有机器阅读理解(MRC)模型无法通过构建问句提供充足的标签语义且在噪声样本上表现不佳等问题,本研究提出了一种联合优化策略。首先,通过聚合在司法语料中多次出现的实体构建司法领域词典,将专业性较强的实体知识注入 RoBERTa-wwm 预训练语言模型进行预训练。然后,通过基于自注意力机制来区分每个字对不同标签词的重要性,从而将实体标签语义融合到句子表示中。最后,在微调阶段采用对抗训练算法对模型进行优化,增强模型的鲁棒性和泛化能力。在2021年中国法律智能评测(CAIL2021)司法信息抽取数据集上的实验结果表明:相较于基线模型,本研究方法 $F1$ 值提高了2.79%,并且模型在 CAIL2021 司法信息抽取赛道中获得了全国三等奖的成绩,验证了联合优化策略的有效性。

关键词:司法信息抽取;预训练;自注意力机制;标签语义;对抗训练

中图分类号:TP391　　**文献标识码**:A　　**文章编号**:1005-9164(2023)01-0027-08
DOI:10.13656/j.cnki.gxkx.20230308.003

在大数据时代,各大法律案件公开网站每天都会发布大量的法律文本。这些文本通常为非结构化文本,不仅专业性较强而且篇幅也较长。从这些文本中提取关键信息对实现“智慧司法”建设具有现实意义,其结果将辅助司法办案人员快速阅卷并厘清案件信息,同时也是知识图谱构建、相似案例推荐、自动量刑建议等一系列任务的重要基础。

实体识别是自然语言处理中的一类基础任务,是从非结构化或半结构化文本提取出用户指定类型的实体,并输出为结构化的信息。在法律文本中主要体现为对案件关键信息如嫌疑人、涉案物品、犯罪事实等的精确抽取。早期常用的方法是基于规则的方法[1],但基于规则的方法存在泛化能力差的缺点。有研究报道使用语法解析树的方法来提高抽取效

收稿日期:2022-09-28　　修回日期:2022-10-11
*国家自然科学基金通用联合基金重点项目(U1836205),国家自然科学基金重大研究计划项目(91746116),国家自然科学基金项目(62166007,62066007,62066008),贵州省科技重大专项计划项目(黔科合重大专项字[2017]3002)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】
余俊晖(1995-),男,在读硕士研究生,主要从事自然语言处理信息抽取、阅读理解等研究。
【**通信作者】
陈艳平(1980-),男,教授,主要从事自然语言处理、人工智能等研究,E-mail:ypench@gmail.com。
【引用本文】
余俊晖,陈艳平,秦永彬,等.基于机器阅读理解的中文司法实体识别优化策略研究[J].广西科学,2023,30(1):27-34.
YU J H,CHEN Y P,QIN Y B,et al. Research on Optimization Strategy of Chinese Judicial Entity Recognition Based on Machine Reading Comprehension [J]. Guangxi Sciences,2023,30(1):27-34.

率[2-4],然而从语法解析的角度进行嵌套命名实体识别很大程度上依赖于标记语料库。在中文语言中,句子的分块或者分词相对较困难,因为此类语言既没有已经切分的句子序列,其词汇也没有明显的边界标记,这使得语法解析方法无法取得较好的性能。

近年来,随着深度学习的发展,自然语言处理相关任务取得了优异的性能。Seo 等[5]改进了注意力机制,通过使用双向注意力流机制获得问句感知的上下文表征,从而得到更深层次的上下文语义信息,实现了机器阅读理解模型对答案片段的抽取。Li 等[6]针对命名实体识别任务,设计了一个统一的阅读理解框架,能够同时识别出文本中的非嵌套实体及嵌套实体,成为目前有效的实体识别手段之一。然而由于中文司法领域数据的专业性,司法信息抽取领域仍然存在以下问题需要解决。

(a)缺乏知识的预训练。自从 Google 于 2018 年提出 BERT[7]预训练语言模型,自然语言处理的各项任务取得了巨大的突破,预训练-微调范式成为自然语言处理任务最常用的手段。在此基础上,Cui 等[8]提出了中文任务的 RoBERTa-wwm 预训练语言模型,采用分词技术构造全词遮蔽任务进行预训练,有效提升了中文任务的性能。该系列模型通过大规模语料预训练学习通用知识,然而司法数据领域性较强,通用领域预训练语言模型(PLM)缺乏专业的司法词法结构信息,尤其是专业性强的实体,通用领域预训练语言模型在识别时准确性较差。

(b)缺乏充足的标签语义信息。Yang 等[9]研究发现现有机器阅读理解模型并不能通过构建问句获得充足的标签语义,提出将标签语义通过语义融合模块增强问句表示的方法。该方法将标签语义注入问句描述中并取得了有效的性能,然而这种方式并没有很好地区分句子中每个字对不同标签的重要性。

(c)存在噪声数据,导致泛化性差。当处理司法数据时,模型往往会因为字符的缺失或者写错而出现理解偏差,抽取出错误的实体。如给定“赵××于 2012 年 3 月 15 日撬开刘×家进行盗窃 2012 年 3 月 18 日,赵××在菜市场被公安逮捕”段落,提问“找出所有的案发时间?”原有模型会因为段落中“2012 年 3 月 18 日”之前缺失一个逗号而导致语义理解错误,从而将被捕时间抽取成作案时间。因此,需要改善模型应对噪声数据的能力,提升模型的鲁棒性和泛化性。

针对以上问题,本研究基于 RoBERTa-wwm 预训练语言模型,使用实体知识注入预训练模型、标签语义感知和对抗训练 3 种方法进行联合优化。其中实体知识注入预训练模型通过聚合在司法语料中多次出现的实体构建司法领域词典,将专业性较强的实体知识注入 RoBERTa-wwm 预训练语言模型中进行预训练,这种预训练方式能够将司法标签语义注入模型中,有效地学习司法领域的相关知识。标签语义感知通过自注意力机制区分句子中不同字对不同标签的重要性,将实体标签信息融合到句子中。最后,对抗训练阶段借助快速梯度算法(Fast Gradient Method,FGM)[10]提升模型的鲁棒性和泛化能力。

1 策略构建

1.1 基于机器阅读理解的司法信息抽取模型构建

机器阅读理解模型给出的一个问句 Query,通过自注意力机制从文本中提取答案实体所在的开始位置和结束位置,图 1 即为基于机器阅读理解的司法信息抽取模型。

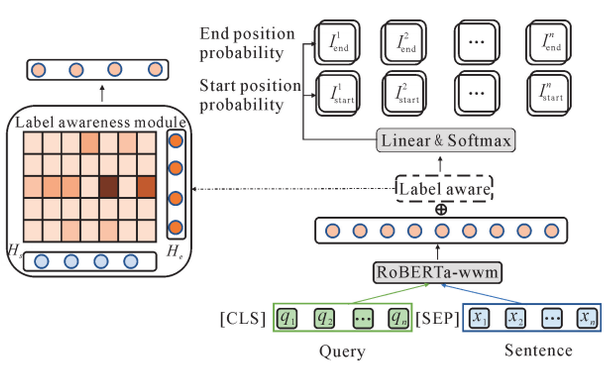

图 1 基于 RoBERTa-wwm 的机器阅读理解模型

Fig. 1 Machine reading comprehension model based on RoBERTa-wwm

1.1.1 模型主干

给出问句 Query = $\{q_1, q_2, \cdots, q_n\}$ 及长度为 n 的序列 $X = \{x_1, x_2, \cdots, x_n\}$,目标从 X 中提取实体的 $x_{\text{start,end}}$ 及其标签 y。将 X 与 Query 组合起来作为 RoBERTa-wwm 预训练语言模型的输入序列 S:{[CLS], $q_1, q_2, \cdots, q_n$, [SEP], $x_1, x_2, \cdots, x_n$},其中[CLS]用于标识整个输入的语义,[SEP]用于分割问句和句子的字符输入。经过公式(1)输出一个上下文句子表示 E_S,

$$E_S = \text{RoBERTa-wwm}(S), \tag{1}$$

其中,$E_S \in \mathbb{R}^{n \times d_{\text{RoBERTa-wwm}}}$,$d_{\text{RoBERTa-wwm}}$ 表示 RoBERTa-wwm 最后一层的维度。

1.1.2 跨度选择

对上下文矩阵 E_S 分别使用两个二分类器,预测某个下标是否为 start 或 end 的概率,具体操作如公式(2)、公式(3)所示:

$$P_{\text{start}} = \text{softmax}(E_S \cdot T_{\text{start}}) \in \mathbb{R}^{n\times 2}, \tag{2}$$

$$P_{\text{end}} = \text{softmax}(E_S \cdot T_{\text{end}}) \in \mathbb{R}^{n\times 2}, \tag{3}$$

其中,P_{start} 表示预测下标为 start 的概率,P_{end} 表示预测下标为 end 的概率,T_{start} 和 T_{end} 为学习参数。

然而在上下文 X 中可能有多个同类实体,即有多个 start 和 end,存在重叠的情况。通过就近原则匹配 start 和 end 显然是不合理的,因此,采用公式(4)、公式(5)解决多个同类实体重叠问题:

$$\hat{I}_{\text{start}} = \{i \mid \text{argmax}(P^{(i)}_{\text{start}}) = 1, i = 1, \cdots, n\}, \tag{4}$$

$$\hat{I}_{\text{end}} = \{i \mid \text{argmax}(P^{(i)}_{\text{end}}) = 1, i = 1, \cdots, n\}, \tag{5}$$

其中,$P^{(i)}_{\text{start}}$ 的每一行表示在给定查询的情况下,每个索引作为实体起始位置的概率分布,$P^{(i)}_{\text{end}}$ 的每一行表示在给定查询的情况下,每个索引作为实体结束位置的概率分布。为解决重叠 start 与 end 匹配问题,使用 argmax 函数作用于输出概率矩阵每行的 P_{start} 和 P_{end},从而获得最大概率的实体开始位置和结束位置的组合。

1.1.3 模型训练

训练时,X 会与两组标签序列 Y_{start} 和 Y_{end} 配对,Y_{start} 和 Y_{end} 的长度均为 n,分别表示每个令牌(token)x_i 是任何实体起始位置或结束位置的真实标签。因此,针对起始位置和结束位置的预测,损失表示为

$$l_{\text{strat}} = \text{CE}(P_{\text{start}}, Y_{\text{start}}), \tag{6}$$

$$l_{\text{end}} = \text{CE}(P_{\text{end}}, Y_{\text{end}})。 \tag{7}$$

则整个跨度(span)的损失表示为

$$l_{\text{span}} = \text{CE}(P_{\text{start,end}}, Y_{\text{start,end}}), \tag{8}$$

整体的训练目标为最小化公式(9):

$$l = \alpha l_{\text{strat}} + \beta l_{\text{end}} + \gamma l_{\text{span}}, \tag{9}$$

其中 CE 表示交叉熵损失函数(CrossEntropy loss),$\alpha, \beta, \gamma \in [0,1]$ 是超参数,用于控制训练时总体的损失,这 3 种损失以端到端的方式联合进行训练,并在 RoBERTa-wwm 预训练语言模型中共享嵌入(embedding)层参数。

1.2 标签语义感知向量化表示机制

1.2.1 编码器

标签向量化表示:给定一个实体标签类型 e_i 和实体类型集 E 中的 n 个标记,利用 RoBERTa-wwm 将实体标签编码为向量化表示 h_{e_i}。向量化表示 h_{e_i} 如下所示:

$$H_{e_i} = \text{RoBERTa-wwm}(e_i), \tag{10}$$

$$h_{e_i} = \text{sum}(H_{e_i}), \tag{11}$$

其中 $H_{e_i} \in \mathbb{R}^{n\times d_{\text{RoBERTa-wwm}}}$ 是 E 中 n 个标记的向量化表示,$h_{e_i} \in \mathbb{R}^{1\times d_{\text{RoBERTa-wwm}}}$ 是来自 E 中第 i 个标签类型向量化表示,然后将每个处理过的向量化表示 h_{e_i} 连接起来得到标签向量化表示 H_E,

$$H_E = \text{concat}(h_{e_1}, h_{e_2}, \cdots, h_{e_{|E|}}), \tag{12}$$

其中 $h_{e_{|E|}} \in \mathbb{R}^{|E|\times d_{\text{RoBERTa-wwm}}}$,将用于标签语义感知模块。

1.2.2 解码器

利用公式(1)获得句子表示 E_S 之后,进一步使用标签语义感知模块将标签语义信息融合到句子表示中。

在标签感知模块(图 1 中 Label aware 的展开图)中,可以利用标签信息来丰富给定句子中每个 token 的信息,这将有助于实体抽取。该模块利用自注意力机制[11]计算标签感知注意力权重 $\text{Att}_{\text{label}}$,该权重表示给定句子中每个 token 嵌入与标签嵌入之间的相关性,然后得到 H_{EA}。

$$Q = W_Q E_S, \tag{13}$$

$$K = W_k H_S, \tag{14}$$

$$\text{Att}_{\text{label}} = \text{softmax}\left(\frac{QK^{\text{T}}}{\sqrt{d_k}}\right), \tag{15}$$

$$H_{EA} = \text{Att}_{\text{label}} H_E, \tag{16}$$

其中 Q、K 均为权重矩阵,由 E_S 经线性变换计算得到,$\text{Att}_{\text{label}} \in R^{m\times|E|}$,$H_{EA} \in \mathbb{R}^{m\times d_{\text{RoRERTa-wwm}}}$,$d_k = d_{\text{RoBERTa-wwm}}/N$,$W_Q, W_k \in \mathbb{R}^{d_{\text{RoBERTa-wwm}}\times d_k}$ 是可训练参数,N 表示 RoBERTa-wwm 中 Transformer 模块的个数。

然后,合并 E_S、H_{EA} 得到 H_{s_k}:

$$(H_{s\_k})^{\text{T}} = \text{concat}((H_S)^{\text{T}}, (H_{EA})^{\text{T}}), \tag{17}$$

其中,$H_{s\_k} \in \mathbb{R}^{m\times 2d_{\text{RoBERTa-wwm}}}$ 将用于 MRC 模型并进行实体的抽取。

1.3 基于对抗训练的模型鲁棒性优化

本研究使用 FGM 对模型进行对抗训练。对抗训练的本质是挑选一个能使得模型产生更大损失的扰动量作为攻击,然后将最大的扰动量添加到输入样本,朝着最小化含有扰动的损失方向更新参数[12]。

对抗训练算法之间的差别在于如何计算扰动值。在 FGM 算法中,根据 embedding 当前的梯度计算扰动值,即

$$\gamma_{adv}=\varepsilon\nabla_x L(x,y;\hat{\theta})/\|\nabla_x L(x,y;\hat{\theta})\|_2, \quad (18)$$

其中,γ 为当前输入扰动;ε 为设置参数,本研究设为 1.0;∇_x 为对 x 求取梯度,x 为模型输入;$L(x,y;\hat{\theta})$ 为样本对应的损失,y 为对应标签,$\hat{\theta}$ 为当前模型参数。

在每一次训练中,得到最大扰动 γ_{adv} 的过程类似于添加 L_2 正则化[13]。将得到的扰动注入词嵌入层的参数中,注入扰动后使用原有样本和标签计算扰动损失,与原有的真实损失一同对模型参数进行优化。

1.4 基于实体知识注入的预训练语言模型

研究表明,预训练语言模型(PLM)可以通过大规模语料库进行自监督预训练来获取知识[14-16],然后将所学知识编码到其模型参数中。然而,由于司法词汇专业性较强,现有 PLM 模型对于那些专业性较强的实体知识难以识别。为了使预训练任务更贴近下游任务且融合更多的中文语义信息,本研究使用 RoBERTa-wwm 预训练语言模型结合司法领域词典进行领域预训练。首先,爬取网上公开的、与本次任务相同的盗窃起诉书 1 万篇,经过正则清洗手段构建与 2021 年中国法律智能评测(CAIL2021)数据相同结构的司法案件语料。然后,聚合在司法语料中多次出现的实体并构建司法领域词典。最后,使用词典辅助 RoBERTa-wwm 预训练语言模型进行全词遮蔽(MASK),MASK 样例如表 1 所示。在该策略中,将实体当成一个统一的单元,相较于 BERT 基于字的 MASK,这个单元中的所有实体在训练时统一被 MASK,以此向模型注入实体的语义信息。模型可以学习到潜在的知识依赖以及更长的语义依赖来让其更具有泛化性。

表 1 MASK 样例

Table 1 MASK sample

说明 Explanation	样例 Example
Original text	嫌疑人使用钢管撬开窗户盗窃
Word segmentation text	嫌疑人 使用 钢管 撬开 窗户 盗窃
Whole word MASK	嫌疑人使用[MASK][MASK]撬开窗户盗窃

2 实验设置

2.1 数据集

本实验主要建立在 CAIL2021 司法信息抽取赛道发布的数据集上,该赛事开源了一阶段和二阶段两个数据集,本研究将其合并。实验数据主要来源于网络公开的若干罪名起诉意见书,包括犯罪嫌疑人、受害人、作案工具、被盗物品、被盗货币、物品价值、盗窃获利、时间、地点和组织机构等 10 类相关业务实体。考虑到多类罪名案件交叉的复杂性,该数据集仅涉及盗窃罪名相关信息的抽取。将数据集划分为训练集、验证集和测试集,三者间的比例为 8∶1∶1。CAIL2021 数据集分布如表 2 所示。

表 2 CAIL2021 数据集实体数量分布

Table 2 CAIL2021 dataset entity quantity distribution

实体类别(含义) Entity category (meaning)	数据集 Dataset		
	训练集 Training set	验证集 Validation set	测试集 Test set
NHCS (Criminal suspect)	7 569	907	922
NHVI (Victim)	3 564	414	429
NCSM (Stolen currency)	1 069	139	125
NCGV (Value of goods)	2 421	287	329
NCSP (Stealing profit)	555	54	58
NASI (Stolen items)	6 647	830	859
NATS (Criminal tools)	804	81	144
NT (Time)	3 195	371	425
NS (Place)	4 068	478	551
NO (Organization)	932	97	123

2.2 问句构建

MRC 模型性能依赖于问句的构建质量。针对这个特点,深入研究 CAIL2021 司法信息抽取赛道的数据集,并参考司法案件分析相关文献,通过聚合与类别标签属性相关的案件要素,有针对性地对有关标签构建司法知识问句,如表 3 所示。

表 3　司法知识问句构建

Table 3　Construction of judicial knowledge questions

实体类别 Entity category	问句构建 Question construction
Criminal suspect	Find out all criminal suspect
Victim	Find all the victims
Stolen currency	Find out all stolen currencies,such as cash,RMB,Vietnamese dong,money,Hong Kong dollars,coins,etc
Value of goods	Find out the value of all items,such as yuan,etc
Stealing profit	Find out all the stolen profits,such as yuan,etc
Stolen items	Find out all stolen items,such as electric cars,mobile phones,cars,motorcycles,etc
Criminal tools	Find out all the crime tools,such as steel pipes,knives,etc
Time	Find out all the time of the crime,including calendar time (year,month,day,etc.) and non calendar time (morning,afternoon,evening,morning,etc.)
Place	Find out all locations,such as administrative district name,street name,community name,building number,floor number,landmark address or natural landscape. In addition,include location indications,such as "in front of the house" or "behind the building"
Organization	Find out all organizations,including companies,government parties,schools,governments,news organizations,etc

2.3　评价指标

采用信息抽取领域常用的精确率(Precision,P)、召回率(Recall,R)及 $F1$ 得分作为评价指标,计算如下:

$$P=\frac{\text{识别正确的命名实体个数}}{\text{识别出的命名实体个数}}\times 100\%,\quad (19)$$

$$R=\frac{\text{识别正确的命名实体个数}}{\text{样本中的所有命名实体个数}}\times 100\%,\quad (20)$$

$$F1=\frac{2\times P\times R}{P+R}\times 100\%。\quad (21)$$

2.4　参数设置

本研究模型是基于 RoBERTa-wwm 预训练语言模型的 MRC 模型,输入 RoBERTa-wwm 预训练语言模型的最大句子长度为 512,学习率设置为 2e-5,优化器采用 AdamW,权重衰减系数设置为 0.01,Dropout 设置为 0.01,单批次大小设置为 32,所有实验都在一张 40 G 的 DGX-A100 显卡上迭代训练 15 轮。

3　实验结果与分析

3.1　实体知识注入预训练模型对模型的影响

表 4 对比了有无实体知识注入进行预训练对模型性能的影响结果。从表 4 中可以看出,在司法领域预训练时注入实体知识信息,相较于原始模型的 $F1$ 得分提高了 1.57%,同时也证明了将预训练任务迁移到司法领域上继续预训练可以进一步提升模型的效果。

表 4　实体知识注入预训练模型对模型的影响

Table 4　Effect of entity knowledge injection pre-training model on model

模型 Model	数据集 Dataset	精确率(%) Precision (%)	召回率(%) Recall (%)	$F1$ 得分(%) $F1$-score (%)
RoBERTa-wwm (No entity knowledge injection)	Validation set	93.22	94.01	93.61
	Test set	93.34	93.92	93.63
RoBERTa-wwm (Entity knowledge injection)	Validation set	94.57	95.63	95.10
	Test set	94.76	95.65	95.20

3.2　标签语义嵌入对模型性能的影响

通过自注意力机制计算原始文本与标签语义之间的相关性,将标签语义嵌入模型中来获取充足的标签语义信息,表 5 在使用 3.1 节预训练的编码器 RoBERTa-wwm + pretrain 基础上对比了有无标签语义嵌入对模型性能的影响。由于 RoBERTa-wwm 预训练语言模型具备丰富的语义知识,表 5 的 $F1$ 值虽然有所提升但是提升不明显。因此,为进一步验证有无

标签语义嵌入对模型性能的影响,使用 Word2Vec 和 Bi-LSTM 替换掉 RoBERTa-wwm 预训练语言模型,并进行相关显著性实验(表 6)。本研究提出的标签语义嵌入机制通过自注意力机制区分不同实体类别标签对句子中每个词语的重要性,将标签语义信息融入句子嵌入中,模型性能得以提升,在验证集和测试集上 $F1$ 值分别提高 0.23%和 0.30%,显著性实验 $F1$ 值分别提升 5.61%和 5.42%,证明为模型添加更多有效的标签语义能够帮助其进一步提升性能。

表 5　标签语义嵌入对模型性能的影响

Table 5　Effect of tag semantic embedding on model performance

模型 Model	数据集 Dataset	精确率(%) Precision (%)	召回率(%) Recall (%)	$F1$ 得分(%) $F1$-score (%)
Tagless knowledge embedding (RoBERTa-wwm + pretrain)	Validation set	94.57	95.63	95.10
	Test set	94.76	95.65	95.20
Tagged knowledge embedding (RoBERTa-wwm + pretrain)	Validation set	94.89	95.78	95.33
	Test set	95.13	95.88	95.50

表 6　替换预训练编码器后标签语义嵌入对模型性能影响的显著性实验

Table 6　Significant experiment on the effect of tag semantic embedding on model performance after replacing the pre-training encoder

模型 Model	数据集 Dataset	精确率(%) Precision (%)	召回率(%) Recall (%)	$F1$ 得分(%) $F1$-score (%)
Tagless knowledge embedding (Word2Vec + Bi-LSTM)	Validation set	59.34	61.19	60.25
	Test set	58.89	60.96	59.91
Tagged knowledge embedding (Word2Vec + Bi-LSTM)	Validation set	65.52	66.20	65.86
	Test set	64.78	65.88	65.33

3.3　对抗训练对模型的影响

对抗训练通过对嵌入层矩阵添加扰动来增强模型在对抗样本下的表现,提升模型整体的鲁棒性和泛化能力。为验证对抗训练对模型的影响,在 3.1 节的预训练权重基础上,对比了训练阶段有无对抗训练的模型性能。从表 7 的实验结果可以看出,在 3.2 节的实验基础上,验证集和测试集上的 $F1$ 值分别提高了 0.73%和 0.58%。这说明在模型训练阶段增加对抗训练可以在一定程度上提升模型的性能。如图 2 所示,模型在无对抗训练的第 7 轮和第 11 轮时性能发生抖动,在加入对抗训练后,模型训练稳定,这表明加入对抗训练能提高模型训练的稳定性,增强模型的鲁棒性和泛化能力。

表 7　对抗训练对模型 $F1$ 值的影响

Table 7　Effect of antagonism training on $F1$ value of model

数据集 Dataset	$F1$ 得分(%) $F1$-score (%)	
	无对抗 No adversarial	有对抗 Adversarial
Validation set	95.10	95.83
Test set	95.20	95.78

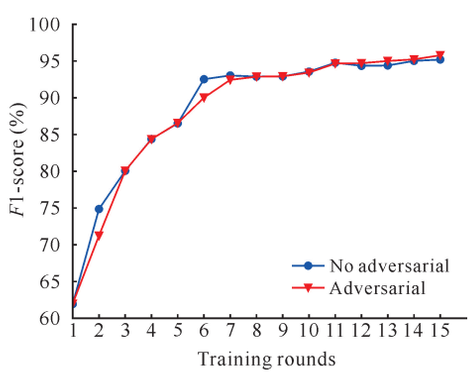

图 2　对抗训练对训练稳定性的影响

Fig. 2　Effect of adversarial training on training stability

3.4　模型对比实验

为进一步验证联合优化策略的有效性,本研究与多种机器阅读理解模型及预训练模型进行对比实验。

①BiADF[5]:使用双向注意力流机制来获得问句感知的上下文表示,对上下文和问句之间的复杂交互关系进行建模。

②BERT-MRC[6]:通过构建问句获得实体知识,以阅读理解的方式对实体进行抽取。

③LEAR[9]:BERT-MRC 的一种改进模型,将待

抽取的实体标签描述为查询语句,与原始文本进行拼接,然后基于 BERT 对实体进行抽取,通过这种方式让模型学习实体标签本身的语义信息。

④BERT[7]:一种预训练语言模型,使用 Transformer 架构的编码器,用于学习在给定上下文的情况下词的向量化表示。

⑤RoBERTa-wwm[8]:一种使用全词 MASK 的中文预训练语言模型,并且使用了更多的预训练数据。

相关性能指标如表 8 所示。从表 8 中可以看出预训练模型的性能远远优于基于 BiDAF 的机器阅读理解模型,这说明经过大规模数据预训练后的模型可以得到很好的语义表示,通过预训练-微调范式可以有效地提高模型在特定任务上的性能;同时,预训练阶段注入额外的实体知识能够使得预训练模型与外部知识进行交互,从而提升 MRC 任务的性能。本研究模型基于 RoBERTa-wwm 预训练语言模型进行设计,采用了多任务联合的机制,相比基线模型有较好的性能表现;进一步使用基于注意力机制的标签知识嵌入机制,将标签知识融合到句子向量化表示中来提高模型的性能,最终模型性能整体提升了 2.79%。在实施联合优化策略后,本研究模型在 CAIL2021 司法信息抽取赛道的数据集上获得了更好的 $F1$ 值,表明了实体知识注入预训练模型、标签知识嵌入及对抗训练对司法 MRC 任务有重要的影响,本研究的优化方法可以有效改善模型的性能。

表 8　模型 $F1$ 值对比实验结果

Table 8　Model $F1$ value comparison test results

模型 Model	验证集(%) Validation set (%)	测试集(%) Test set (%)
BiDAF	86.93	86.84
BERT	91.77	91.72
RoBERTa-wwm	92.38	92.43
LEAR	93.39	93.38
BERT-MRC	92.96	92.91
Before model optimization	93.61	93.63
After model optimization	96.47	96.42

4　结论

本研究针对 CAIL2021 司法信息抽取赛道的数据集设计了基于 RoBERTa-wwm 预训练语言模型的司法信息抽取机器阅读理解模型,并联合基于实体知识注入任务预训练、标签知识嵌入机制和对抗训练 3 种方法对模型进行优化,本研究模型的 $F1$ 值比基线模型提高了 2.79%。为探索更多的优化策略,未来的工作可以从以下 3 个方面入手:针对司法信息抽取出来的案件信息构建知识图谱,从而为模型提供更多的知识;增加更多的数据,引入更多的训练方式,对本研究模型进行司法领域预训练;探索引入更多的标签语义提升模型的性能。

参考文献

[1] RAU L F. Extracting company names from text [C]// The Seventh IEEE Conference on Artificial Intelligence Application. Piscataway, NJ: IEEE, 1991: 29-32.

[2] FINKEL J R, MANNING C D. Joint parsing and named entity recognition [C]//Proceedings of Human Language Technologies: the 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2009: 326-334.

[3] ZHANG X T, LI D C, WU X H. Parsing named entity as syntactic structure [C]//INTERSPEECH 2014, Fifteenth Annual Conference of the International Speech Communication Association. Singapore: ISCA, 2014: 278-282.

[4] JIE Z M, MUIS A O, LU W. Efficient dependencyguided named entity recognition [C]//Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. Menlo Park, CA: AAAI, 2017: 3457-3465.

[5] SEO M J, KEMBHAVI A, FARHADI A, et al. Bi-directional attention flow for machine comprehension [Z/OL]. (2018-06-21)[2022-09-07]. https://arxiv.org/pdf/1611.01603.pdf.

[6] LI X Y, FENG J R, MENG Y X, et al. A unified MRC framework for named entity recognition [C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2020: 5849-5859.

[7] 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. Stroudsburg, PA: Association for Computational Linguistics, 2019, 1: 4171-4186.

[8] CUI Y M, CHE W X, LIU T, et al. Pre-training with whole word masking for chinese BERT [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3504-3514.

[9] YANG P, CONG X, SUN Z Y, et al. Enhanced language representation with label knowledge for span extraction [C]//Proceedings of the 2021 Conference on Empirical

Methods in Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2021: 4623-4635.
[10] MIYATO T,DAI A M,GOODFELLOW I. Adversarial training methods for semi-supervised text classification [Z/OL]. (2017-05-06) [2022-09-08]. https://arxiv. org/pdf/1605. 07725v3. pdf.
[11] 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: NIPS,2017.
[12] MADRY A,MAKELOV A,SCHMIDT L,et al. Towards deep learning models resistant to adversarial attacks [Z/OL]. (2019-09-04)[2022-09-14]. https://arxiv. org/pdf/1706. 06083. pdf.
[13] GOODFELLOW I J,SHLENS J,SZEGEDY C. Explaining and harnessing adversarial examples [Z/OL]. (2015-03-20) [2022-09-14]. https://arxiv. org/pdf/1412. 6572. pdf.
[14] TENNEY I,XIA P,CHEN B,et al. What do you learn from context? probing for sentence structure in contextualized word representations [Z/OL]. (2019-05-15) [2022-09-15]. https://arxiv. org/pdf/1905. 06316. pdf.
[15] PETRONI F,ROCKTÄSCHEL T,RIEDEL S,et al. Language models as knowledge bases? [C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Stroudsburg,PA: Association for Computational Linguistics,2019:2463-2473.
[16] ROBERTS A,RAFFEL C,SHAZEER N. How much knowledge can you pack into the parameters of a language model? [C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg,PA: Association for Computational Linguistics,2020:5418-5426.

Research on Optimization Strategy of Chinese Judicial Entity Recognition Based on Machine Reading Comprehension

YU Junhui[1,2],CHEN Yanping[1,2**],QIN Yongbin[1,2],HUANG Hui[1,2]
(1. State Key Laboratory of Public Big Data,Guiyang,Guizhou,550025,China;2. College of Computer Science and Technology,Guizhou University,Guiyang,Guizhou,550025,China)

Abstract: Aiming at the problems that the entities in the Chinese judicial information extraction dataset are highly professional,the existing Machine Reading Comprehension (MRC) model cannot provide sufficient label semantics by constructing questions and performs poorly on noise samples,a joint optimization strategy is proposed in this study. Firstly,a judicial domain dictionary is constructed by aggregating entities that appear many times in the judicial corpus,and professional entity knowledge is injected into the RoBERTa-wwm pre-training language model for pre-training. Then,the entity label semantics are integrated into the sentence representation by distinguishing the importance of each word to different label words based on the self-attention mechanism. Finally,in the fine-tuning stage,the adversarial training algorithm is used to optimize the model to enhance the robustness and generalization ability of the model. The experimental results on the 2021 China Legal Intelligence Evaluation (CAIL2021) judicial information extraction dataset show that compared with the baseline model,the $F1$ value of this research method is increased by 2. 79%. And the model in the CAIL2021 judicial information extraction track won the national third prize,which verified the effectiveness of the joint optimization strategy.

Key words: judicial information extraction; pre-training; self-attention mechanism; label semantics; adversarial training

责任编辑:米慧芝