软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2021,32(10):3139−3150 [doi: 10.13328/j.cnki.jos.006028]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

融入案件辅助句的低频和易混淆罪名预测*

郭军军[1,2], 刘真丞[1,2], 余正涛[1,2], 黄于欣[1,2], 相 艳[1,2]

[1](昆明理工大学 信息工程与自动化学院,云南 昆明 650500)

[2](云南省人工智能实验室(昆明理工大学),云南 昆明 650500)

通讯作者: 余正涛, E-mail: ztyu@hotmail.com

摘 要: 由于低频罪名数据量较少和易混淆罪名案情描述相似等原因,导致低频和易混淆罪名预测效果不佳.为了解决此类问题,通过构建案件辅助句,提出一种基于双向互注意力机制的案件辅助句融合方法,实现罪名预测.主要包括以下 3 部分:首先,基于司法领域知识构建案件辅助句,将案件辅助句作为案情描述和罪名之间的映射知识;然后,基于词级和字符级表征分别提取案情描述与案件辅助句多粒度特征;同时,借助案件辅助句与案情描述双向注意机制,获得具有辅助句倾向性的案情描述表征,并最终实现低频和易混淆罪名的预测.基于中国刑事案件公共数据集的实验结果表明:所提方法在 $F1$ 值最大提升 13.2%,准确率最大提升 4.5%,低频罪名预测 $F1$ 值提升 4.3%,易混淆罪名预测 $F1$ 值提升 8.2%,所提算法显著地提升了低频和易混淆罪名的预测性能.

关键词: 低频罪名;易混淆罪名;双向互注意力;多粒度编码;案件辅助句

中图法分类号: TP18

中文引用格式: 郭军军,刘真丞,余正涛,黄于欣,相艳.融入案件辅助句的低频和易混淆罪名预测.软件学报,2021,32(10):3139−3150. http://www.jos.org.cn/1000-9825/6028.htm

英文引用格式: Guo JJ, Liu ZC, Yu ZT, Huang YX, Xiang Y. Few shot and confusing charges prediction with the auxiliary sentences of case. Ruan Jian Xue Bao/Journal of Software, 2021,32(10):3139−3150 (in Chinese). http://www.jos.org.cn/1000-9825/6028.htm

Few Shot and Confusing Charges Prediction with the Auxiliary Sentences of Case

GUO Jun-Jun[1,2], LIU Zhen-Cheng[1,2], YU Zheng-Tao[1,2], HUANG Yu-Xin[1,2], XIANG Yan[1,2]

[1](Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China)

[2](Yunnan Key Laboratory of Artificial Intelligence (Kunming University of Science and Technology), Kunming 650500, China)

Abstract: Due to the insufficiency of few shot charges and the similarity of case descriptions for the confusing charges, the prediction performance of the existing methods for few shot charges and confusing charges is not promising. To address the forementioned drawbacks, a novel few shot and confusing charges prediction method is proposed, which is based on bi-direction mutual attention mechanism with the auxiliary sentences of case. For the proposed model, firstly, the auxiliary sentence of case via the judicial field is constructed, where the auxiliary sentence of case is considered as external knowledge for mapping the description of the case to the corresponding charge. Secondly, the multi-granularity characteristics of case description and the auxiliary sentence of case are extracted at the level of both word and character, respectively. At the same time, the auxiliary sentence of case and case description are used to build

* 基金项目: 国家重点研发计划(2018YFC0830105, 2018YFC0830101, 2018YFC0830100); 国家自然科学基金(61972186, 61762056, 61472168, 61866020); 云南省科技厅省级人培项目(KKSY201703015); 云南省基础研究专项面上项目(2019FB082, 202001AT070047)

Foundation item: National Key Research and Development Program of China (2018YFC0830105, 2018YFC0830101, 2018YFC0830100); National Natural Science Foundation of China (61972186, 61762056, 61472168, 61866020); Provincial Personnel Training Project of Yunnan Science and Technology Department (KKSY201703015); Natural Science Foundation Project of Yunnan Science and Technology Department (2019FB082, 202001AT070047)

收稿时间: 2019-12-06; 修改时间: 2020-02-09; 采用时间: 2020-03-02

bi-direction mutual attention. Finally, the tendency representation of the case description with the guidance of the auxiliary sentence of case are derived, which improve the prediction accuracy of few shot and confusing charges. The experimental results conducted on the benchmark data of criminal cases show that the proposed model increases the *F*1 value and prediction accuracy by 13.2% and 4.5%, respectively, and increases the *F*1 values for the few shot charges and confusing charges by 4.3% and 8.2%, respectively, which significantly enhance the prediction performance for few shot and confusing charges.

Key words: few shot charge; confusing charge; bi-direction mutual attention; multi-granular coding; auxiliary sentence of case

罪名预测任务是法律判决任务中一个重要的子任务,在法律领域中发挥着至关重要的作用.目前常见罪名预测准确率比较高,但低频和易混淆罪名的预测准确率却不尽如人意,主要是因为低频罪名数据少和易混淆罪名案情描述相似等原因所致.据统计,截止目前,我国刑法罪名共有 469 类,罪名的分布呈典型的长尾分布(幂律分布的一种形式).在我国几千万的裁判文书数据中,我们统计了大量真实案例数据后发现:案例数据极度不均衡,部分案例的案情描述不易区分.

早期阶段,有研究者基于传统的统计学习方法来解决罪名预测任务.也有研究者试图利用字符、单词和短语等浅层文本特征预测罪名.近年来,罪名预测任务通常被形式化为文本分类任务,研究人员大多基于神经网络模型解决罪名预测任务,也有部分研究人员提出融入外部知识共同建模的方法.基于传统文本分类的方法难以从低频和易混淆罪名案例中学习到案件的关键特征,因此,低频和易混淆罪名预测仍然是罪名预测任务的难点.提升低频和易混淆罪名的预测准确率,是法律判决任务有待解决的难题之一.

(1) 低频罪名预测

我们统计裁判文书案例数据时发现:比较常见的罪名(如盗窃罪、抢劫罪等)占了大约 78%;比较低频的几十类罪名(如倒卖文物罪、高利转贷罪等)只占了不到 0.5%,此类低频罪名中大部分案例数据只有十多条,导致低频罪名可训练的案例数据特别少.因此,基于神经网络模型很难学习到足够的案件关键特征.故而在数据量有限的条件下,低频罪名的准确预测是一个严峻的挑战.

(2) 易混淆罪名预测

在我国刑事案件数据中,有很大一部分罪名及其案情描述不易区分,比如(抢劫罪,抢夺罪)、(盗伐林木罪,滥伐林木罪)等.此类罪名数据很难提取案例中有效区分因素,容易误导模型学习到彼此的噪声特征,干预模型的判断能力.因此,提高易混淆罪名预测的准确性,也是有待解决的一个难题.

对于低频和易混淆罪名预测准确率低这一问题,本文提出一种融入案件辅助句构建双向互注意力的方法,旨在提高低频和易混淆罪名的预测性能.不同于以往传统文本分类的方法,我们主要基于案件辅助句指导案情描述计算多粒度关键信息倾向性表征.拟基于案件辅助句与案情描述构建双向互注意力,捕捉具有案件辅助句感知的案情描述特征,最终提升低频和易混淆罪名的预测准确率.

案件辅助句作为案情描述与罪名之间的内在映射,不仅可以为低频罪名扩充关键信息,还可以为易混淆罪名提供有效区分因素.具体来说,我们分析了大量的中国刑事案件数据后,定义了几类案件辅助句.以抢劫罪和抢夺罪为例.首先,由这两类罪名的案件性质可知,这两类罪名均有“故意犯罪行为”和“以非法占有为目的”,以此可区别于其他部分案件(如过失致人死亡罪等);其次,通过案情描述对比分析,如图 1 所示,可知“抢劫罪”的案情描述中包含了“强行推倒”“刺伤”和“威胁”等暴力手段;与之相反,“抢夺罪”的案情描述更倾向于“趁其不备”,未使用暴力手段.

因此,我们可定义抢劫罪和抢夺罪的有效区分因素为该案件是否“以暴力为手段”.以此方法类推,我们分别定义其他几类案件的辅助句子.

为了验证本文所提方法对低频和易混淆罪名预测性能的提升,我们分别在 3 个不同规模的中国刑事案件公共数据集中进行实验.实验结果表明:与其他基线模型相比,本文模型在 3 个数据集上均取得了最显著的效果,评估指标均优于基线模型.与引入罪名区分属性解决低频和易混淆罪名预测模型(当前低频和易混淆罪名预测性能最佳模型)相比,本文模型在 3 个数据集上宏观 *F*1 值最大提升 13.2%,准确率最大提升 4.5%.值得一提的是:本文模型在低频罪名预测宏观 *F*1 值提升 4.3%,易混淆罪名预测宏观 *F*1 值提升 8.2%.

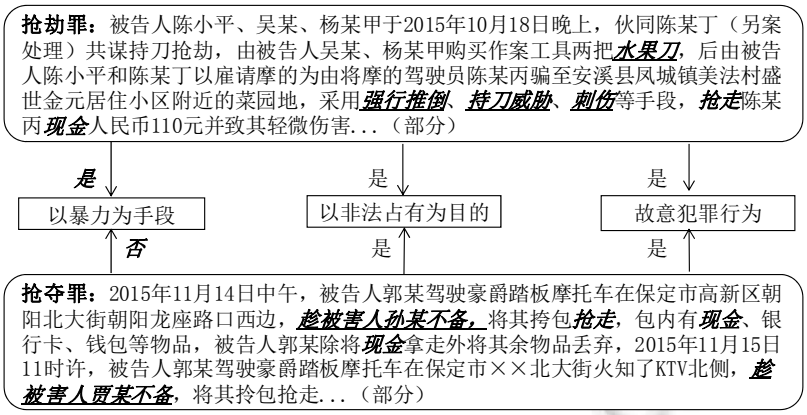

Fig.1 Comparative analysis of robbery and snatch cases

图 1 抢劫罪与抢夺罪案例对比分析

综上所述,我们的主要贡献如下:

(1) 本文的核心工作是提高低频和易混淆罪名的预测准确率,为此,我们首次引入案件辅助句这一概念,改善其预测性能;

(2) 本文采用一种融入案件辅助句构建双向互注意力的学习框架来改进低频和易混淆罪名预测性能,获取具有案件辅助句指导的案情描述多粒度倾向性特征表征;

(3) 基于 3 个不同规模的中国刑事案件公共数据集进行实验,本文模型的实验结果比其他基线模型取得了更显著的效果.

1 相关工作

罪名预测任务[1]是法律智能领域研究多年的任务.在早期工作中,由于缺少大量的标注数据,部分研究者基于传统统计学习的方法解决罪名预测任务.例如:Kort 等人[2]采用定量方法,通过计算事实元素的数值来预测判断;Nagel 等人[3]利用相关分析对重新分配的案例进行预测;Keown 等人[4]引入用于法律预测的数学模型,如线性模型等.这些方法通常被限制在只有少量标签的小数据集上.Liu 和 Hsieh[5]在罪名预测任务中考虑了短语特征,Liu 等人[6]使用 *K*NN(*K* 最近邻算法)对台湾 12 项指控进行预测,Sulea 等人[7]提出了一种基于支持向量机(SVM)的集成系统,该系统利用案情描述、判决和时间跨度作为输入.然而,以上方法只能获取浅层的语义特征,对于罪名预测任务而言,不足以捕捉到足够的案件关键信息.随着神经网络模型能够提取自然语言更深层次的语义特征,研究人员考虑基于神经网络建模,旨在提取案件更深层次的语义信息,从而提高罪名预测的准确率.

研究人员基于神经网络模型将罪名预测形式化为文本分类任务,文本分类通常着重于从案情描述中提取关键特征.Zhong 等人[8]基于神经网络模型,结合多个子任务间的拓扑结构信息提出联合模型,同时完成法律判决预测的多个子任务.Luo 等人[9]基于注意力机制的神经网络模型,在罪名预测任务中融入法条信息.该方法使罪名预测任务更加合理化.Jiang 等人[10]采用神经网络模型抽取合理的、可读的、决定性的片段信息来强化法律判决预测,提升法律判决预测的性能.由于只采用简单的文本分类可能导致案例信息交互不足,为了解决这个问题,Long 等人[11]基于判决流程,采用阅读理解的方式对罪名预测进行建模.王加伟等人[12]基于词语语义差异性完成罪名预测.基于神经网络模型的方法目前在常见罪名预测取得了较好的准确率,但对于低频和易混淆罪名而言,预测准确率仍然较低.

因此,有研究者另辟他法,采用融入外部知识的方法构建神经网络模型.例如:Lin 等人[13]提出一种针对两类罪名定义 21 个法律要素的方法进行多任务学习,该方法局限性较大,需投入大量的人工标注,可扩展性与针对性较差;刘宗林等人[14]提出了融入罪名关键词的法律判决预测多任务学习模型,判决结果包括法条推荐和罪名预测,该方法随着数据量的增加,同样需要投入大量的人工标注工作;Hu 等人[15]提出了融入罪名区分属性预测

低频和易混淆罪名的方法,引入罪名区分属性在一定程度上减少了大数据集的标注工作,同时在低频和易混淆罪名预测任务上取得了当前最好的效果.我们认为:仅通过给定的事实描述结合简单注意力机制提取的罪名区分属性特征,未能更全面地捕捉到案件的关键信息,对于低频和易混淆罪名预测准确率的提升还存在一定的进步空间.

受此启发,我们基于司法领域的大量刑事案件数据,根据其不同罪名案例数据的特性,定义了几类代表性较强的案件辅助句,它们能够更简洁地概括案件的核心语义信息.借助于案件辅助句的指导,分别从低频和易混淆罪名的案情描述中提取更多有助于罪名识别的案件关键信息.本文模型同样基于神经网络模型将罪名预测形式化为一个多分类任务,主要借鉴 Minjoon 等人[16]提出的双向注意流建模的思想,我们提出一种融入案件辅助句构建双向互注意力机制的神经网络模型.与以往工作不同,我们同时利用案件辅助句与案情描述构建双向互注意力,强化案情描述和案件辅助句之间的信息交互,提取具有案件辅助句指导的案情描述多粒度倾向性特征,提升低频和易混淆罪名预测的准确率.基于中国刑事案件公共数据集的实验结果表明:与基线模型相比,本文模型能够更好地提升低频和易混淆罪名的预测准确率.

2 融入案件辅助句的双向互注意力罪名预测模型

针对低频和易混淆罪名预测准确率低这类问题,本文首次融入案件辅助句,并将其与案情描述相结合构建双向互注意力模型.双向互注意力网络层与前后紧密衔接,每个时刻的注意力向量与此前的嵌入层息息相关,并衔接之后的网络层.同时,我们采用多粒度特征计算的方式对案件辅助句和案情描述分别编码,获取其多层次的语义特征向量,旨在提取具有案件辅助句指导的案情描述上下文特征,最终提升低频和易混淆罪名预测准确率.

案情描述词和案件辅助句词分别用$\{X_1,\ldots,X_T\}$和$\{A_1,\ldots,A_J\}$表示,其中,T 和 J 分别表示案件辅助句和案情描述的长度.本文模型主要分为 4 部分,分别是案件辅助句的构建与多粒度特征提取、双向互注意力计算、具有案件辅助句指导的上下文特征提取和罪名预测输出网络层,如图 2 所示.

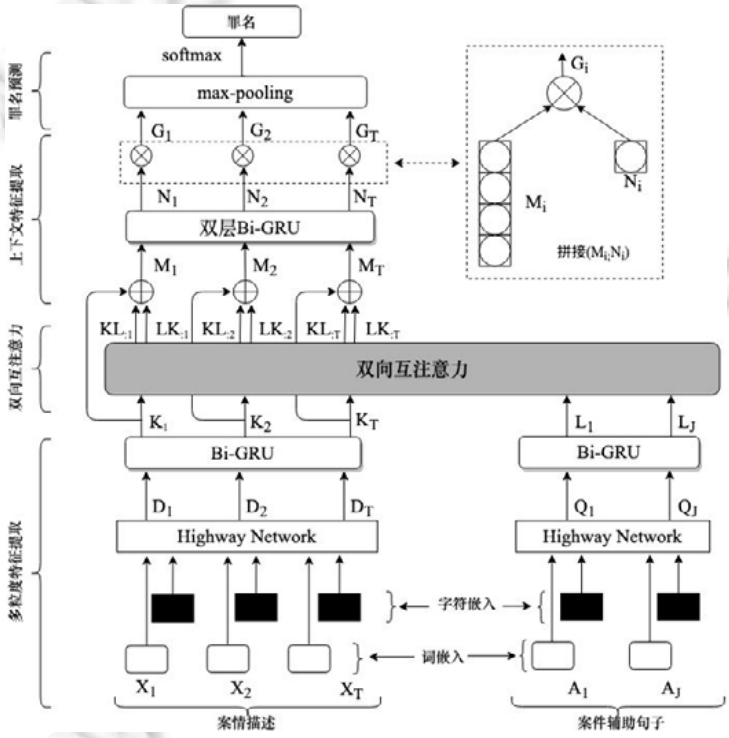

Fig.2 Bi-direction mutual attention model based on the auxiliary sentence of case

图 2 融入案件辅助句的双向互注意力模型

2.1 案件辅助句的构建与多粒度特征提取

2.1.1 案件辅助句的构建

为了改进低频罪名数据量小和易混淆罪名案情描述相似导致其预测准确率低这一问题,我们基于司法领域构建案件辅助句.借助于案件辅助句扩充低频罪名的构成元素和易混淆罪名的有效区分因素.采用如图 1 所示的方式分析了大量的中国刑事案件数据后,共定义了 9 类具有案件核心语义信息的案件辅助句,案件辅助句的详细介绍见表 1.值得一提的是,我们定义案件辅助句时均遵循法律规定和案件判决的客观事实.

Table 1 Auxiliary sentence of cases

表 1 案件辅助句

罪名	编号	案件辅助句	是否相关
C_m	B_1	以牟利为目的	0 或 1
	B_2	以买或卖为目的	
	B_3	产生死亡事实	
	B_4	以暴力为手段	
	B_5	该案件和国家机关或国家相关工作人员有关	
	B_6	该案件发生于公共场合	
	B_7	以非法占有为目的	
	B_8	该案件造成人身伤害	
	B_9	故意犯罪行为	

罪名和案件辅助句分别用$\{C_m\}$和$\{B_n\}$表示(m从 0 到 148,n从 1 到 9).案件辅助句作为罪名和案情描述之间的映射,我们使用(C_m,B_n)表示它们的映射关系,其中,B_n被标注为“0”或“1”(“0”和“1”分别表示该案件辅助句和罪名相关和不相关),见表 1.我们只选择被标记为“0”的案件辅助句与案情描述计算双向互注意力向量.例如,某一案件的罪名是“抢劫罪”,其映射关系可表示为($C_{\text{抢劫罪}}$,0,1,1,0,1,1,0,1,0),当 n=1,4,7,9 时,这几条案件辅助句与“抢劫罪”相关.此外,当判断案件辅助句和案件是否相关时,C_m和 B_n必须符合该案件的客观事实描述.比如,当 C_m是“故意伤害罪”时,B_3(产生死亡事实)必须判断为“不相关”,否则违背法律规定和案件判决客观事实.值得一提的是:对于相同罪名的不同案例,案件辅助句的标注是相同的,只需投入少量的人工标注工作.

2.1.2 案件辅助句和案情描述多粒度特征提取

(1) 字符嵌入

字符嵌入是将每个字符映射到高维向量空间.类似于 Kim 等人[17]的工作,我们使用 CNN(卷积神经网络)获取每个案例的案情描述和案件辅助句的字符嵌入.字符嵌入到向量理解为 CNN 的一维输入,其大小为 CNN 的输入通道大小.CNN 的输出经过最大池化后即可获得固定大小的字符向量表征.

(2) 词嵌入

词嵌入也是把每个词映射到高维的向量空间.不同于字符嵌入层,我们采用的是 Skip-Gram[18]模型对嵌入大小为 100 的词进行预训练,将每个词映射到一个向量空间.

(3) 高速网络

引入 Highway Network(高速网络)[19]是为了平衡词向量和字符向量的相对贡献比.我们把案情描述和案件辅助句的字符向量和词向量进行简单拼接,输入到一个两层的 Highway Network,输出 2 个 d×2 维的向量序列,分别表示案情描述多层次的向量表征 $D\in\mathbb{R}^{2d\times T}$和案件辅助句多层次的向量表征 $Q\in\mathbb{R}^{2d\times J}$.

(4) 上下文嵌入

该网络层的目的是细化单词嵌入,由于此前向量表征未考虑上下文语义特征,我们将案情描述向量表征 D和案件辅助句向量表征 Q 输入一个可以理解上下文信息的嵌入机制.本文采用单层的 Bi-GRU(双向门控循环神经网络)[20]作为理解上下文信息的嵌入机制,模拟单词之间的特征交互,并将 Bi-GRU 两个方向的输出进行简单拼接,得到该网络层的输出,分别是 $K\in\mathbb{R}^{2d\times T}$和 $L\in\mathbb{R}^{2d\times J}$.$K$ 表示具有上下文特征的案情描述多粒度向量表征,L表示具有上下文特征的案件辅助句多粒度向量表征.

2.2 双向互注意力计算

双向互注意力(bi-direction mutual attention)网络层负责将具有上下文特征的案件辅助句信息和案情描述信息进行耦合,不同于以往常用的注意力计算方式,我们将每个时刻的注意力向量与之前的嵌入层相关联,且都流向之后的网络层,目的是缓解过早归纳总结而导致的信息丢失.

2.2.1 案情描述与案件辅助句相似矩阵

该网络层主要是计算具有上下文特征的案情描述表征向量 K 和案件辅助句表征向量 L 的互注意力向量.我们首先计算 K 与 L 之间的共享相似矩阵 S,再分别计算 K 和 L 之间的双向互注意力向量.相似矩阵 S 的计算如公式(1):

$$S_{tj}=\alpha(K_{:t},L_{:j})\in\mathbb{R}^{2d\times J} \tag{1}$$

其中,S_{tj} 表示第 t 个案情描述词和第 j 个案件辅助句词之间的相似性;$K_{:t}$ 表示 K 的第 t 列向量;$L_{:j}$ 表示 L 的第 j 列向量;α表示计算 K 与 L 之间相似度的可训练函数,如公式(2):

$$\alpha(k,l)=W_{(S)}^{T}(k;l;k\circ l)\in\mathbb{R} \tag{2}$$

其中,$W_{(S)}^{T}\in R^{6d}$ 是待训练的权重向量,∘表示元素依次相乘,(;)表示向量在行上进行拼接,k 与 K 的列向量对应,l 与 L 的列向量对应.

2.2.2 案情描述到案件辅助句的注意力

对每个案情描述词而言,我们捕捉案件辅助句中与其比较相关的词,如图 3(左)所示.我们对 S 中列进行 *softmax* 归一化得到 a_t,再将 a_t 与 L 中的每一列加权求和得到 KL,KL 表示案情描述与案件辅助句之间的注意力向量矩阵,如公式(3)所示:

$$\begin{aligned} a_t &= softmax(S_{t:})\in\mathbb{R}^{J}\\ KL &= \sum a_{tj}L_{:j}\in\mathbb{R}^{2d\times T}\end{aligned} \tag{3}$$

其中,$S_{t:}$表示第 t 个案情描述词与案件辅助句词的相似度,a_t 则表示第 t 个案情描述词对案件辅助句词的注意力权重.

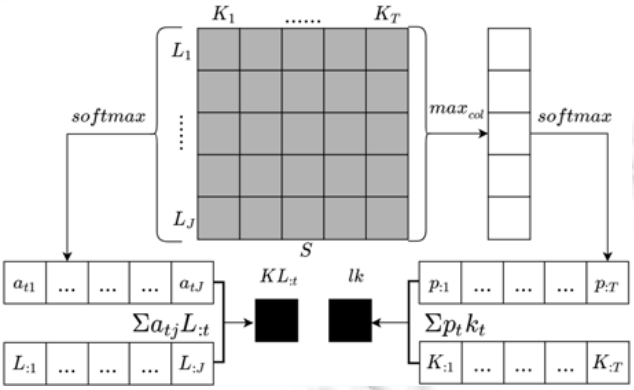

Fig.3 Bi-direction mutual attention calculation, where S represents the similarity matrix, which a_t is the normalization of column in S, and p is normalized after taking the maximum value of each column in S

图 3 双向互注意力计算,其中,S 表示相似矩阵,a_t 是 S 中列归一化所得,p 是取 S 中每列的最大值后再归一化

2.2.3 案件辅助句到案情描述的注意力

案件辅助句到案情描述的注意力计算方式是对于案件辅助句中的词而言的,案情描述中那些词与它比较相关,这些词对案件关键特征的学习尤为重要,如图 3(右)所示.取相似矩阵 S 中每列的最大值 e,再经 *softmax* 归一化后得到 p,利用 p 计算案情描述中与案件辅助句比较相关词的加权求和得到 lk;然后,lk 沿着列方向平铺 T 次得到案件辅助句到案情描述的注意力向量矩阵 LK,如公式(4):

$$\left.\begin{aligned} e &= \max_{col}(s) \in \mathbb{R}^T \\ p &= softmax(e) \in \mathbb{R}^T \\ lk &= \sum p_t K_{:t} \in \mathbb{R}^{2d} \\ LK &= Tiled_T(lk) \in \mathbb{R}^{2d \times T} \end{aligned}\right\} \tag{4}$$

其中,$\max_{col}$表示取 S 中每列的最大值,$Tiled_T$表示 lk 沿着列方向平铺 T 次.

最终,我们得到案情描述和案件辅助句的双向互注意力向量 KL 和 LK.通过 Cat 函数计算 K(案情描述表征向量)、KL 和 LK 得矩阵 M,M 中每个列向量可以视为具有案件辅助句感知的案情描述表征.如公式(5):

$$\begin{aligned} M_{:t} &= Cat(K_{:t}, KL_{:t}, LK_{:t}) \in \mathbb{R}^{d_M} \\ Cat(k, kl, lk) &= (k; kl, k \odot kl; k \odot lk) \in \mathbb{R}^{8d \times T} \\ d_M &= 8d \end{aligned} \tag{5}$$

其中,$M_{:t}$表示第 t 个列向量,与第 t 个案情描述词相对应;k 与 K 的列向量对应;LK 表示 L 中某一列向量与 K 的注意力向量;KL 表示 K 中某一列向量与 L 的注意力向量(K 和 L 与第 2.1.2 节相对应);⊙表示矩阵相乘.

2.3 具有案件辅助句指导的案情描述上下文特征提取

2.3.1 案情描述上下文特征提取

该网络层采用双层的 Bi-GRU 作为嵌入机制,主要是为了获取具有案件辅助句感知的案情描述向量表征 M 在时序上的上下文依赖关系.这与之前的上下文嵌入网络层不同,前者获取具有案件辅助句感知的案情描述上下文语义依赖特征,后者独立获取案情描述和案件辅助句上下文语义特征.我们把 Bi-GRU 两个方向的输出进行拼接,输入到下一层 Bi-GRU 后得到矩阵 N,N 中每列向量表示具有案件辅助句感知的案情描述上下文依赖特征表征.如公式(6):

$$N = BiGRU(M) \in \mathbb{R}^{2d \times T} \tag{6}$$

2.3.2 案情描述显著特征提取

这一部分主要借鉴残差网络的思想,把具有案件辅助句感知的案情描述特征表征 M 与案情描述上下文依赖特征表征 N 进行拼接后得到 G,然后采用池化操作提取 G 中的显著特征 H,如公式(7):

$$\begin{aligned} G &= (M; N) \in \mathbb{R}^{10d \times T} \\ H &= pooling(G) \in \mathbb{R}^{10d} \end{aligned} \tag{7}$$

其中,(;)表示向量在行上的拼接,$pooling$ 采用最大池化.

2.4 罪名预测输出

该网络层是根据犯罪事实预测出某一个案件的最终罪名.主要是把前网络层提取的显著特征 H 通过 $softmax$ 函数,以获取预测结果的概率分布,如公式(9):

$$P = softmax(W_{(P)} H) \tag{8}$$

其中,P 表示罪名预测结果的概率分布,$W_{(P)}$是可训练的权重向量.

3 罪名预测实验

为了证明本文所提方法在罪名预测任务上性能的提升,尤其是低频罪名和易混淆罪名预测准确率的提升,我们分别在 3 个不同规模的中国刑事案件公共数据集上进行实验,并结合本文模型做了两类对比实验:一类是与其他基线模型的性能对比实验,以验证该模型的有效性;另一类是本文模型的消融测试实验,以验证该模型的合理性.此外,还分别验证本文模型对低频和易混淆罪名预测的性能提升,以及双向互注意力机制可视化实验.

3.1 数据集

本文使用的数据集是 Hu 等人[15]2018 年公开的中国刑事案件公共数据集.该数据集共包含 149 类罪名,40 万个案例,并且只包含一项指控.原作者将其随机分为 3 个不同规模的数据集,分别包含 7 万、20 万和 38 万案

例数据,均包含相同类别的罪名及不同数量的案例.不同于他们的工作,我们加入了第 2.1 节中提到的案件辅助句,根据其映射关系,分别标注 3 个数据集的案例.根据案例数量分别命名为数据集 *S*、数据集 *M* 和数据集 *L*,其训练集、测试集和验证集的数量划分见表 2.

Table 2 Statistics of different datasets

表 2 不同数据集的统计信息

数据集	数据集 *S*	数据集 *M*	数据集 *L*
训练集	61 586	153 521	306 900
测试集	7 700	19 180	38 360
验证集	7 750	19 250	38 420

3.2 实验参数设置及评价指标

由于案例文档长度大多在 500~650 之间,我们设最大文档长度为 596.采用 Adam 算法[21]作为优化器;学习率设为 0.000 1;字符嵌入层 CNN 的过滤器大小分别设置为(1,2,3,4),Dropout(随机失活率)[22]设为丢失 0.2;单层 Bi-GRU 的随机失活率设为丢失 0.2;双层 Bi-GRU 的 Dropout 设为丢失 0.5;批次处理大小设为 10;训练轮次设置为 12.本文的评价指标主要采用准确率(Acc.)、宏观准确率(MP)、宏观召回(MR)和宏观 *F*1 值.该模型的代码会在以后公布.

3.3 基线模型

这一部分内容本文主要采用了几类非常典型的文本分类模型和两类比较新颖的罪名预测模型作为基线模型.其中,Few-Shot Attributes 模型是 Hu 等人 2018 年提出解决低频和易混淆罪名预测的模型,达到了当前该任务的最好效果,基线模型如下所示.

- TFIDF+SVM 模型:我们使用词频逆文档频率(TFIDF)[23]提取输入特征,并采用支持向量机(SVM)[24]作为分类器;
- CNN 模型:我们设置具有多个过滤器宽度的卷积神经网络(CNN)[17]作为分类器;
- LSTM 模型:我们使用两层的长短期记忆网络(LSTM),并采用最大池化提取最大特征;
- Fact-Law Attention 模型:Luo 等人[9]于 2017 年结合相关法律文献提出了一种基于注意力的罪名预测神经网络模型;
- Few-Shot Attributes 模型:Hu 等人[15]于 2018 年提出了一种结合罪名区分属性提升低频罪名预测性能的神经网络模型.

对于 TFIDF+SVM 模型,本文将特征大小设为 2 000,对于其他神经网络模型,本文使用 Skip-Gram[18]模型预先训练词向量,并设置嵌入大小为 100.将 LSTM 模型的隐藏状态大小设置为 100.基于 CNN 的模型,为了保持一致性,我们将过滤器的宽度都设置为(2,3,4,5),每个过滤器的大小设置为 25.值得注意的是:当词向量和字符向量拼接后,本文模型的表示大小变成了 200.为了更公平地加以比较,我们在 CNN 和 LSTM 的池化层之后添加了一个 100×200 的全连接层,记为 CNN-200 和 LSTM-200.

3.4 实验结果及分析

这一部分工作主要结合本文模型做了两类对比实验:一类是与 CNN、LSTM 和 Few-Shot Attri.等基线模型的性能对比,旨在验证本文模型对罪名预测,尤其是低频和易混淆罪名预测性能优于当前最先进的基线模型,见表 3;另一类是为了验证本文模型的案件辅助句、多粒度特征计算以及高速网络的有效性,见表 4.特别说明:“(–)字符嵌入”表示未使用字符编码,只使用词嵌入;“(–)高速网络”表示未使用高速网络平衡字符向量和词向量的贡献比;“(–)案件辅助句”表示未使用我们定义的案件辅助句.此外,还分别验证了本文模型在低频和易混淆罪名预测性能的提升.

第 1 类对比实验中,由表 3 可知:本文模型的 Acc.、MP、MR 和 *F*1 值均超过所有基线模型,*F*1 值最大提升为 32.5%.这可以证明:基线模型在低频和易混淆罪名预测效果略有不足;反之,本文模型对低频和易混淆罪名预

测性能实现了有效的改进.与当前 Few-Shot Attri.模型(低频和易混淆预测当前最优)对比,本文模型在 3 个数据集上的 *F*1 值分别提升 7.6%、13.2%和 12.5%,准确率最大提升 4.5%.验证了本文模型的鲁棒性和有效性,也证明本文模型可以有效地提升低频和易混淆罪名预测的准确率.

消融测试实验结果见表 4,由表 4 可知,本文模型 Acc.、MP、MR 和 *F*1 值均超过表 3 中其他基线模型.当我们未使用字符编码计算多粒度特征、高速网络平衡词向量和字符向量贡献比时,宏观 Acc.、MP、MR 和 *F*1 值均略微下降.由此可以证明,多粒度特征计算及高速网络是本文模型重要的一环.在案件辅助句的消融实验部分,我们从案情描述中随机抽取一个句子代替案件辅助句构建互注意力机制.从表 3 可知,本文模型的宏观 *F*1 值至少下降 6.1%.因此可以证明:融入的案件辅助句对低频和易混淆罪名预测准确率的提升是非常重要的,也是本文模型中必不可少的一环.

Table 3 Comparison of experimental results between the model and baselines

表 3 本文模型与基线模型实验结果对比

数据集	案例(小)				案例(中)				案例(大)			
评价指标	Acc.	MP	MR	*F*1	Acc.	MP	MR	*F*1	Acc.	MP	MR	*F*1
TRIDF+SVM	85.8	49.7	41.9	43.5	89.6	58.8	50.1	52.1	91.8	67.5	54.1	57.5
CNN	91.9	50.5	44.9	46.1	93.5	57.6	48.1	50.5	93.9	66.0	50.3	54.7
CNN-200	92.6	51.1	46.3	47.3	92.8	56.2	50.0	50.8	94.1	61.9	50.0	53.1
LSTM	93.5	59.4	58.6	57.3	94.7	65.8	63.0	62.6	95.5	69.8	67.0	66.8
LSTM-200	92.7	60.0	58.4	57.0	94.4	66.5	62.4	62.7	95.1	72.8	66.7	67.6
Fact-Law Att.	92.8	57.0	53.9	53.4	94.7	66.7	60.4	61.8	95.7	73.3	67.1	68.6
Few-Shot Attri.	93.4	66.7	69.2	64.9	94.4	69.2	69.2	67.1	95.8	75.8	73.7	73.1
本文模型	**97.9**	**72.6**	**75.0**	**72.5**	**98.9**	**82.4**	**80.5**	**80.3**	**99.2**	**89.2**	**84.5**	**85.6**

Table 4 Experimental results of ablation test

表 4 消融测试实验结果

数据集	案例(小)				案例(中)				案例(大)			
评价指标	Acc.	MP	MR	*F*1	Acc.	MP	MR	*F*1	Acc.	MP	MR	*F*1
(–)字符嵌入	97.1	69.3	72.8	70.0	97.9	77.4	77.6	76.1	98.3	86.2	81.1	82.3
(–)高速网络	97.3	72.1	73.0	71.1	98.2	80.1	78.6	78.8	98.6	87.6	82.9	84.2
(–)案件辅助句	94.5	66.3	67.2	66.4	96.8	72.7	70.9	70.2	97.8	80.8	78.2	78.6
本文模型	**97.9**	**72.6**	**75.0**	**72.5**	**98.9**	**81.7**	**80.9**	**79.8**	**99.2**	**88.7**	**85.1**	**85.9**

为了进一步验证本文模型对低频和易混淆罪名预测性能的有效改进,我们分别验证该模型在低频罪名预测任务和易混淆罪名预测任务的性能提升.选取比较经典的分类模型 LSTM(维度大小设置为 200,保持与本文模型一致)和目前低频及易混淆罪名预测效果最好的 Few-Shot Attri.模型当作基线模型,选用宏观 *F*1 值当作评价指标.

首先统计“数据集 *S*”中同一罪名对应的不同案例数量,根据案例数量将其分为 3 部分:当同一罪名的案例数据小于 10 时,该罪名归类为低频;当同一罪名案例数据大于 100 时,该罪名归类为高频;其余部分罪名归类为中频.分别计算不同频率罪名对应的宏观 *F*1 值,见表 5.

Table 5 Macro *F*1 values for charge of different frequencies on case dataset *S*

表 5 数据集 *S* 上不同频率罪名的宏观 *F*1 值

罪名类别	低频	中频	高频
罪名数量	55	47	47
LSTM-200	32.1	54.5	82.7
Few-shot Attri.	48.6	59.2	85.5
本文模型	**52.9**	**70.1**	**94.2**

由表 5 中实验结果可知:与 LSTM-200 模型和 Few-Shot Attri.模型相比,对于案例数据小于 10 的低频罪名预测,本文模型的宏观 *F*1 值分别提升 20.2%和 4.3%.由此可以证明,本文模型对低频罪名预测性能的提升是有效的;同时也可证明:融入案件辅助句,可在一定程度上改善案例数据极度不均衡这一问题.此外,本文模型对中

高频罪名预测的性能也有很好的提升,宏观 $F1$ 值最少提升分别为 10.9%和 8.7%.

其次,我们从“数据集 S”中选取案例数据均大于 1 000 的 4 类易混淆罪名,避免被低频数据干扰.主要是“放火罪”和“失火罪”、“抢夺罪”和“抢劫罪”、“行贿罪”和“受贿罪”以及“盗伐林木罪”和“滥伐林木罪”,验证本文模型对易混淆罪名预测性能的提升,实验结果见表 6.

Table 6 Macro $F1$ values for confusing charge on case dataset S

表 6 数据集 S 中易混淆罪名的宏观 $F1$ 值

罪名类别	易混淆罪名
LSTM-200	79.7
Few-Shot Attri.	88.1
本文模型	**96.3**

从表 6 中的实验结果可知:对于易混淆罪名区分预测,本文模型比 LSTM-200 模型和 Few-Shot Attri.模型的宏观 $F1$ 值分别提升 16.6%和 8.2%.由此可以证明:本文模型能够更好地捕捉易混淆案例的区分特征,进一步提高易混淆罪名预测的准确性.

3.5 案例分析

这一部分工作主要为了验证融入案件辅助句对低频和易混淆罪名预测性能的提升,我们选择一个如图 7(左)所示的易混淆案例.被告人在此案中被判定的最终罪名是“过失致人重伤罪”.该案例很具代表性,容易被“争执”和“厮打”等噪声数据误导,包括人工阅读的第一直觉也会误认为是“故意伤害罪”.此外,该罪名对应的案件辅助句如图 7(右)所示.

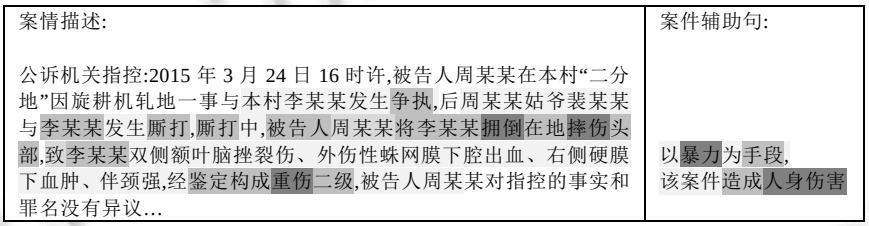

Fig.7 Visualization of bi-direction mutual attention mechanism

图 7 双向互注意力机制可视化

由于“过失致人重伤罪”和“故意伤害罪”都与暴力行为和人身伤害相关,因此很难将该案件判定为过失伤害还是故意伤害,两者的重要区别是:前者对应的案件辅助句还包含“故意犯罪行为”,而后者没有.因此我们认为,该案件辅助句是两者的重要区分.从图 7 中可以直观地看到该案例的案情描述和案件辅助句的注意力机制可视化分布(背景颜色较深的词具有较高的注意力权重),双向互注意力机制可以捕获案情描述和案件辅助句中的核心语义信息.

针对上述案例,Few-Shot Attri.模型和本文模型的预测结果见表 7.本文模型正确预测为“过失致人重伤罪”,Few-Shot Attri.模型预测结果为“故意伤害罪”.由此可以证明:融入具有案件核心语义信息的辅助句,可在很大程度上帮助我们的模型捕捉案件的关键信息,提升低频和易混淆罪名预测的准确率.

Table 7 Charge prediction result of the selected case

表 7 选取案例罪名预测结果

模型	预测结果
真实罪名	过失致人重伤罪
Few-Shot Attri.	故意伤害罪
本文模型	**过失致人重伤罪**

4 总结与展望

本文研究旨在提升低频易混淆罪名的预测准确率.针对低频罪名训练数据少和易混淆罪名案情描述不易区分等原因导致两者预测准确率低这一问题,我们引入了案件辅助句这一概念,提出了一种融入案件辅助句与案情描述构建双向互注意力建模的方法.此外,我们分别计算案情描述与案件辅助句不同粒度的特征.本文模型在中国刑事案件数据集上取得当前最显著的效果.当然,我们的工作目前还有很多的不足之处,比如案件辅助句的定义比较广泛,以及是否可自动构建方案更细化的案件辅助句,这是有待解决的工作之一.

未来的研究工作主要分为 3 个方面.

(1) 考虑基于神经网络等模型自动构建更为细化的案件辅助句,提升案件辅助句与案情描述更深层次的语义信息交互;

(2) 当前工作只考虑了单项罪名指控,将来的工作设想结合多项指控案例数据,改进该模型为多罪名预测,这更符合我国刑事案件罪名自动判决的目的;

(3) 法律判决预测任务包含法条推荐、罪名预测、刑期预测等多项子任务,将来的工作还考虑结合多项子任务,提升低资源数据的判决预测准确率.

致谢 本文工作是在作者导师的悉心指导下完成的,深表感谢.同时,也真诚地感谢团队的其他老师和同学的耐心解答,以及在法律判决任务方面做了大量工作的前辈们.

References:

[1] Xiao CJ, Zhong HX, Guo ZP, Tu CC, Liu ZY, Sun MS. Cail2018: A large-scale legal dataset for judgment prediction. 2018. https://arxiv.org/abs/1807.0247

[2] Kort, F. Predicting Supreme Court decisions mathematically: A quantitative analysis of the “right to counsel” cases. The American Political Science Review, 1957,51(1):1−12.

[3] Nagel SS. Applying correlation analysis to case prediction. Texas Law Review, 1963,42:1006.

[4] Keown R. Mathematical models for legal prediction. Computer/Law Journal, 1980,2:829.

[5] Liu CL, Hsieh CD. Exploring phrase-based classification of judicial documents for criminal charges in Chinese. In: Proc. of the ISMIS. 2006. 681−690.

[6] Chao LL, Cheng TC, Jim HH. Some case-refinement strategies for case-based criminal summary judgments. Journal of Information Science and Engineering, 2004,20(4):783−800.

[7] Sulea OM, Zampieri M, Vela M, Josef VG. Exploring the use of text classication in the legal domain. In: Proc. of the ASAIL. 2017.

[8] Zhong HX, Guo ZP, Tu CC, Xiao CJ, Liu ZY, Sun MS. Legal judgment prediction via topological learning. In: Proc. of the 2018 Conf. on Empirical Methodsin Natural Language Processing. Brussels: ACL, 2018. 3540−3549.

[9] Luo BF, Feng YS, Xu JB, Zhang X, Zhao DY. Learning to predict charges for criminal cases with legal basis. In: Proc. of the 2017 Conf. on Empirical Methodsin Natural Language Processing. Copenhagen: ACL, 2017. 2727−2736.

[10] Jiang X, Ye H, Luo ZC, Chao WH. Interpretable rationale augmented charge prediction system. In: Proc. of the 27th Int’l Conf. on Computational Linguistics. Santa Fe: ACL, 2018. 146−151.

[11] Long SB, Tu CC, Liu ZY, Sun MS. Automatic judgment prediction via legal reading comprehension. In: Proc. of the 18th China National Conf. on Computational Linguistics. Kunming: CCL, 2019. 558−572.

[12] Wang JW, Zhang H, Tan HY, Wang YL, Zhao HY, Li R. Multi-label charge prediction based on semantic differences of words. Journal of Chinese Information Processing, 2019,33(10):127−134 (in Chinese with English abstract).

[13] Lin WC, Kuo TT, Chang TJ, Yen CA, Chao JC, Lin SD. Exploiting machine learning models for chinese legal documents labeling, case classification, and sentencing prediction. Int’l Journal of Computational Linguistics & Chinese Language Processing, 2012, 17(4):49−67 (in Chinese with English abstract).

[14] Liu ZL, Zhang MS, Zhen RR, Gong ZQ, Yu N, Fu GH. Multi-task learning model for legal judgment predictions with charge keywords. Journal of Tsinghua University (Science and Technology), 2019,59(7):497−504 (in Chinese with English abstract).

[15] Hu ZK, Li X, Tu CC, Liu ZY, Sun MS. Few-Shot charge prediction with discriminative legal attributes. In: Proc. of the 27th Int'l Conf. on Computational Linguistics. Santa Fe: ACL, 2018. 487−498.

[16] Minjoon S, Aniruddha K, Farhadi A, Hajishirzi H. Bidirectional attention flow for machine comprehension. In: Proc. of the 5th Int'l Conf. on Learning Representations. Toulon: ICLR, 2017.

[17] Kim Y. Convolutional neural networks for sentence classification. In: Proc. of the 2014 Conf. on Empirical Methodsin Natural Language Processing. Doha: ACL, 2014. 1746−1751.

[18] Mikolov T, Sutskever I, Chen K, Greg SC, Dean J. Distributed representations of words and phrases and their compositionality. In: Proc. of the NIPS. 2013. 3111−3119.

[19] Srivastava RK, Greff K, Schmidhuber J. Highway networks. Computer Science, 2015.

[20] Cho K, Caglar G. Learning phrase representations using RNN encoder-decoder for statistical machine translation. Computer Science, 2014.

[21] Diederik K, Jimmy B. Adam: A method for stochastic optimization. In Proc. of the 3rd Int'l Conf. on Learning Representations. San Diego: ICLR, 2015.

[22] Srivastava N, Geoffrey EH, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 2015,15(1):1929−1958.

[23] Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. Information Processing & Management, 1988,24(5): 513−523.

[24] Johan AK, Suykens JV. Least squares support vector machine classifiers. Neural Processing Letters, 1999,9(3):293−300.

附中文参考文献:

[12] 王加伟,张虎,谭红叶,王元龙,赵红燕,李茹.基于词语语义差异性的多标签罪名预测.中文信息学报,2019,33(10):127−134.

[13] 林琬真,郭宗廷,張桐嘉,顏厥安,陳昭如,林守德.利用機器學習於中文法律文件之標記、案件分類及量刑預測.中文計算語言學期刊,2012,17(4):49−67.

[14] 刘宗林,张梅山,甄冉冉,公佐权,余南,付国宏.融入罪名关键词的法律判决预测多任务学习模型.清华大学学报(自然科学版),2019,59(7):497−504.

郭军军(1987－),男,博士,讲师,CCF 专业会员,主要研究领域为自然语言处理,信息检索,机器翻译.

黄于欣(1983－),男,博士,CCF 专业会员,主要研究领域为自然语言处理,文本摘要,文本生成.

刘真丞(1997－),男,学士,主要研究领域为自然语言处理,事件抽取.

相艳(1979－),女,讲师,主要研究领域为自然语言处理,情感分析,信息抽取.

余正涛(1970－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为自然语言处理,机器翻译,信息检索.