北京大学学报(自然科学版) 第 60 卷 第 1 期 2024 年 1 月
Acta Scientiarum Naturalium Universitatis Pekinensis, Vol. 60, No. 1 (Jan. 2024)
doi: 10.13209/j.0479-8023.2023.077

双向注意力文本关键词匹配法条推荐

丁娜[1] 刘鹏[2,†] 邵惠鹏[3] 王学奎[4]

1. 中国矿业大学信息与控制工程学院, 徐州 221116; 2. 矿山互联网应用技术国家地方联合工程实验室, 徐州 221008; 3. 江苏省徐州市铜山区公安局法制大队, 徐州 221100; 4. 阿里巴巴集团有限公司, 杭州 311121;
† 通信作者, E-mail: liupeng@cumt.edu.cn

摘要 提出一种双向注意力文本关键词匹配的法条推荐模型(BiAKLaw)。该模型以预训练语言模型BERT作为基础匹配模型, 利用双向注意力机制提取字符级对齐特征和关键词差异特征, 融合对齐特征、差异特征和关键词语义表征来提升匹配效果。在裁判文书交通肇事和故意伤害数据集上的实验结果表明, 与BERT模型相比, BiAKLaw在评价指标F1上分别提升3.74%和3.43%。

关键词 法条推荐; 案件事实; 文本匹配; 注意力机制

Bi-Attention Text-Keyword Matching for Law Recommendation

DING Na[1], LIU Peng[2, †], SHAO Huipeng[3], WANG Xuekui[4]

1. School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116; 2. National Joint Engineering Laboratory of Internet Applied Technology of Mines, Xuzhou 221008; 3. Legal Team of Tongshan Branch of Xuzhou Public Security Bureau, Xuzhou 221100; 4. Alibaba Group, Hangzhou 311121; † Corresponding author, E-mail: liupeng@cumt.edu.cn

Abstract This paper proposed a bi-directional attention based text-keyword matching model for law recommendation (BiAKLaw). In this model, BERT is utilized as a basic matching model, bi-directional attention mechanism is implemented to extract token-level alignment features and keyword-level differential features, and these features are fused with keyword attentive semantic representations for a better matching effect. The experimental results on the traffic accident and intentional injury datasets demonstrate that, compared with BERT, the proposed model increases F1 evaluation metric by 3.74% and 3.43% respectively.

Key words law recommendation; case fact; text matching; attention mechanism

法律领域的人工智能(legal artificial intelligence, LegalAI)旨在利用人工智能技术来处理该领域的各类任务, 已成为研究热点。法律资料多以语言文本形式呈现, 因此LegalAI广泛使用自然语言处理(natural language processing, NLP)技术来提高自动化程度, 可以在一定程度上减轻法律从业者的劳动强度, 提升法律事务处理效率, 同时降低司法应用的使用门槛, 给非专业人员提供一定的法律援助。LegalAI的典型应用包括法律判决预测(legal judgement prediction, LJP)、相似案例匹配和法律问答等。法律判决预测任务主要关注如何基于事实描述和法条内容, 来预测判决结果。法条推荐是其中重要的子任务。

我国是实行民法体系的国家, 法官基于既定的成文法条给出判决结果, 因此, 根据案件事实, 准确、全面地找到相关法条, 是法官做出公正裁决的重要前提。在早期LegalAI研究中, 法律判决预测多利用数学统计方法给出参考判决结果[1–4]。近年来, 随着神经网络技术的发展, 研究者多利用文本分类来处理法条推荐任务[5–8], 还有一些研究者融合法律专业知识来辅助判决[9–11]。已有的法条推荐研究大多将法条视为离散标签[12], 仅提取案件事实

中国矿业大学研究生创新计划(2023WLJCRCZL111)资助
收稿日期: 2023–05–18; 修回日期: 2023–07–31

的文本特征作为输入, 预测相应的法条标签, 忽略了法条的语义信息。然而, 在实际审案过程中, 法官通常以法条内容为准则对案件事实进行分析, 再确定嫌疑人行为匹配的具体法条。因此, 法条内涵的语义信息是定罪和量刑的重要依据, 需要加以充分利用。

在法律判决过程中, 找到案件相关的所有法条至关重要, 遗漏或错误推荐任一法条都会影响判决结果, 妨碍司法公正。一个案例通常包括多个犯罪事实, 如果将案例看成一个整体来预测涉及的所有法条, 得到的结果还需要花费大量人工将事实对应相应的法条, 可解释性较差。同时, 案例属于长文本, 包含各种粒度的特征, 将案例按事实划分, 有助于更好地提取文本特征信息。

本文将法条推荐任务视为案件事实与法条内容的文本匹配任务, 提出一种双向注意力文本关键词匹配的法条推荐模型(BiAKLaw)。根据 Ge 等[13]的做法, 本文将案例划分成若干独立事实, 将案例中的每个事实与法条库的每个法条进行匹配, 将一对多匹配转变成多对多匹配问题。一对多和多对多的法条匹配如图 1 所示。另外, 由于案件中犯罪事实和法条之间可能含有重合字符, 只依靠匹配标签, 模型很难学到犯罪事实与法条之间的关键差异特征。关键词序列去除了无关字符, 能最大程度地表达一段话的实际意义。因此, 本文通过提取案件和法条两者的关键词序列, 融合关键词序列的匹配特征, 提升模型的匹配结果。

1 相关工作

1.1 法律判决预测

法律判决预测包括法条推荐、罪名预测和刑期预测 3 个子任务, 最早的研究开始于 20 世纪 50 年代, 主要围绕如何使用数学和统计方法分析历史案件, 找出共性规律, 并构建模型来模拟判决流程。例如, Kort[1]通过定量分析大量已经裁决的案件来预测美国最高法院的判决。受限于人力和计算机技术, 这类基于数学统计的模型准确率和泛化能力都不高。随着机器学习技术的发展, 研究者多以分类框架为基础来构建模型, 将法条推荐和罪名预测均视为分类问题, 将历史法律文书作为训练样本, 法条和罪名作为类别标签。例如, Luo 等[9]基于 SVM, 为每个法条标签训练二元分类器, 预测法条与案件事实的关系, 并选择最相关的 k 条法条进行后续的罪名分类, Liu 等[14]基于 PAT 树和 HowNet 构建领域相关词表, 并用 K 近邻算法(K-nearest-neighbour, KNN)处理浅层文本特征来解决多标签罪名分类问题。此类机器学习模型在一定程度上提升了预测性能, 但是高度依赖手工提取模板, 且提取的只是浅层文本特征, 对高层特征的提取能力较弱。

近年来, 研究者大多使用深度学习技术处理法律判决预测任务, 大致可以分为两种类型: 一类是通过模型创新提升性能[15−16], 另一类是融合外部法律知识提升性能[9−11]。另外, 法律判决预测技术的发展离不开高质量公开数据集的支撑。Xiao 等[17]构建了首个中文法律判决预测数据集, 该数据集包

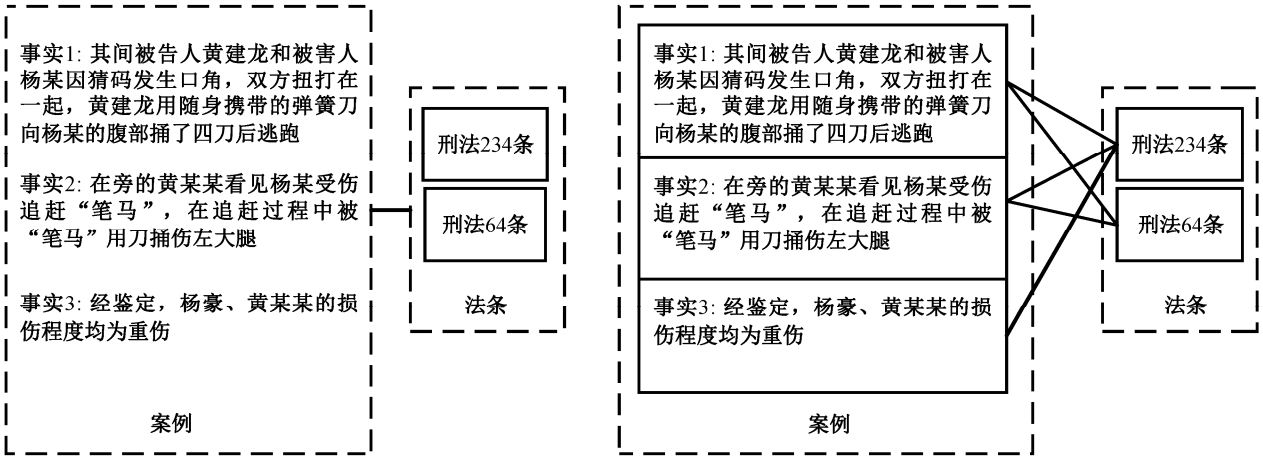

图 1 一对多和多对多法条匹配示意图

Fig. 1 Example diagram of one to many and many to many law recommendation

含中国最高人民法院公开的260万刑事案件。Chalkidis等[18]根据欧洲人权法院案件，构建一个英文法律判决预测数据集，该数据集的不足之处是规模较小。

1.2 法条文本匹配

法条推荐可作为案件事实和法条内容的文本匹配任务，提取两者文本特征并计算匹配度。判断两个文本是否匹配是许多NLP任务的基础工作，例如机器翻译、自动问答、释义识别和文本检索等。传统的文本匹配方法依赖于人工提取特征，这些模型往往针对特定的领域，泛化能力不强。例如，使用BM25算法[19]计算相似度时，解决的只是句子词汇的相似度问题，没有考虑句子深层语义的匹配，对复杂句子的效果不佳。

近年来，越来越多的研究者使用深度学习方法来完成文本匹配任务。按照模型架构不同来分类，大致可以分为表示型和交互型两种。典型的表示型模型有DSSM[20]和ARC-I[21]等，均采用孪生网络结构，参数共享，具有对称性。首先将两段文本转换为高维向量，在模型末端对两段文本进行特征交互，最后得到匹配结果。此类模型的优点是参数量较少，易于训练，缺点是两段文本在各自编码完成之前没有交互，可能会丢失重要语义信息，影响匹配结果。交互型模型是表示型模型的改进，在两段文本编码过程中进行交互匹配，然后提取文本特征，最后进行相似度计算。例如，Hu等[21]在ARC-I的基础上提出ARC-II结构；Chen等[22]结合注意力机制提出ESIM模型完成自然语言推理任务。交互式框架能够学习到更多的交互特征，匹配效果得到提升，缺点是推理速度显著变慢，难以应用于大规模语料上。

借助外部知识增强文本语义表征能力，也是文本匹配领域常用的方法。例如，Lyu等[23]将HowNet作为外部知识库，解决了单词歧义问题。Wu等[24]将分类知识作为先验知识来过滤文本中的噪声，并从多个角度进行匹配。但是，外部知识多是结构化知识库，一般需要花费大量人工来构建。随着计算能力的提升，出现基于Transformer[25]架构的预训练语言模型(pre-trained language models, PLMs)，包括BERT[26]，RoBERTa[27]和BERT-wwm[28]等。与传统方法相比，预训练语言模型在多个下游任务(包括文本匹配)中表现优异，迅速成为文本匹配领域的重要研究方法。

鉴于预训练语言模型强大的性能，本文选择BERT系列模型作为基础匹配模型。

2 双向注意力文本关键词匹配的法条推荐模型

2.1 问题描述

为了充分利用法条语义信息，本文将法条推荐任务视为案件事实与法条之间的文本匹配过程。具体来说，将某个案例视为包含若干独立犯罪事实的序列$C=\{F_1, F_2, \ldots, F_m\}$，有限法条集合为$A=\{L_1, L_2, \cdots, L_n\}$，其中，$m$是某个案例包含的案件事实的数量，$n$是某类案件涉及的全部法条的数量。本文的目标就是构建一个映射函数，按照一定的规则找到与每个案例的犯罪事实相匹配的法条，可表示为

$$y=M(F_i, L_j), \tag{1}$$

其中，模型输入为案例的某个事实F_i和法条集合的某个法条L_j，M为匹配神经网络，$y\in\{0, 1\}$，输出结果为不匹配或者匹配。

2.2 模型架构

本文提出的双向注意力文本关键词匹配的法条推荐模型(BiAKLaw)包括关键词抽取层、模型输入层、语义知识交互层以及输出层，模型架构如图2所示。

1) 关键词抽取层主要基于KeyBERT[29]算法，抽取案件事实和法条两者的关键词序列。

2) 模型输入层将案件事实和法条内容拼接输入预训练模型BERT词嵌入编码层，学习文本对的相关性，得到具有语义信息的字符向量表示。

3) 语义知识交互层一方面使用双向注意力机制，得到案件事实和法条文本之间的对齐特征，使用最大池化策略来保留重要特征；另一方面基于案件事实关键词序列和法条关键词序列，对案件事实和法条序列的非关键词位置进行遮蔽。BERT模型在进行自注意力机制计算时，得到的案件事实序列向量只与法条中的关键词有关，反之，法条序列向量只与案件事实关键词有关。对两者序列向量使用平均池化策略，再相减得到关键差异特征，然后将多种特征进行拼接融合送入全连接层。

4) 输出层输出最终的法条匹配结果。

2.3 关键词抽取层

在人工判案过程中，法官会梳理案例中每个犯罪事实的关键信息，翻阅法条手册，判断符合哪个

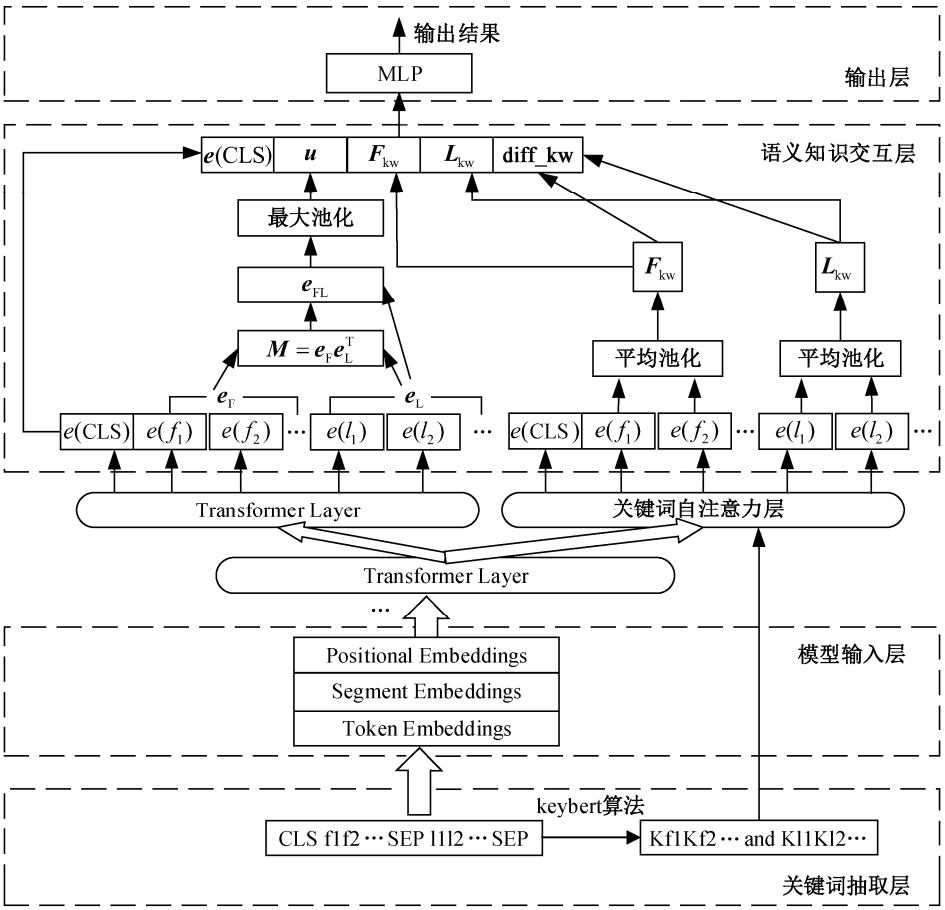

图 2 双向注意力文本关键词匹配的法条推荐模型 BiAKLaw

Fig. 2 Overall architecture diagram of bi-attention text-keyword matching model for law recommendation (BiAKLaw)

法条的前提条件, 最后结合触犯的所有法条做出最终判决。考虑到案件事实和法律条文均属于较长的专业性文本, 如果只使用两者的原序列来计算匹配度, 会引入很多无关的噪声信息, 譬如会使模型误将字面重合度高但关键信息并不匹配的文本预测为相似, 造成错误匹配, 因此我们决定引入能表达关键信息的关键词序列来辅助文本匹配。

KeyBERT是近年出现的一种无监督关键词抽取模型, 通过BERT获取文档和候选词的向量表示, 利用余弦相似度来衡量候选词与文档的相似程度, 选取相似度最高的若干词作为该文档的关键词。由于KeyBERT易于使用, 效果能满足本文方案需求, 因此将其作为关键词抽取模型。关键词抽取质量还受文本嵌入表示影响, 本文使用清华大学发布的基于大规模刑事文书预训练模型[30]来获取词嵌入表示。案件事实序列为$\{f_1, f_2, \ldots, f_p\}$, p代表事实字符数量。法条序列为$\{l_1, l_2, \ldots, l_q\}$, q代表法条字符数量。使用KeyBERT算法分别抽取案件事实和法条序列中的关键词, 得到{Kf1, Kf2…}和{Kl1, Kl2…}。

2.4 模型输入层

模型的输入为案件事实以及法条文本, 通过BERT获得语义编码向量。BERT输入可以是一段文本或两段文本的拼接, 当输入为两段文本时, 中间用字符[SEP]连接, [CLS]和[SEP]符号会插在拼接文本的开头和结尾, 然后输入到词嵌入编码层。词嵌入编码层由3个部分组成: 字符嵌入层(token embeddings)是根据文本序列分词后的token, 查找词表得到的向量表示; 段编码层(segment embeddings)是为了区分输入是第一段文本还是第二段文本; 位置编码层(positional embeddings)引入每个token在对应segment中的位置编码信息。最后, 将三部分相加, 输入BERT词嵌入层。

2.5 语义交互层

将案件事实和法条文本拼接输入BERT词嵌入

层后，经过多层Transformer得到最终的高维语义向量。由于BERT是双向模型，序列中的每个字符相互可见，经过自注意力机制后，得到的特征向量是前后文相关的，每个序列中更重要的词语会被赋予更大的权重。

将图2中模型框架BERT最后一层的输出作为案件事实和法条文本的向量表示:

$$(\boldsymbol{e}(\mathrm{CLS}), \boldsymbol{e}_{\mathrm{F}}, \boldsymbol{e}_{\mathrm{L}}) = \mathrm{BERT}(\mathrm{fact}, \mathrm{law}), \tag{2}$$

其中，$\boldsymbol{e}(\mathrm{CLS}) \in \mathrm{R}^d$为语义匹配的一个特征，$\boldsymbol{e}_{\mathrm{F}} \in \mathrm{R}^{p\times d}$为事实描述的语义表征，$\boldsymbol{e}_{\mathrm{L}} \in \mathrm{R}^{q\times d}$为法条的语义表征，$d$代表BERT隐藏层维度，$p$和$q$分别为事实和法条的字符数。本文使用双向注意力机制，得到匹配文本表征间的对齐特征。

注意力权重矩阵的计算如下:

$$\boldsymbol{M} = \boldsymbol{e}_{\mathrm{F}} \cdot \boldsymbol{e}_{\mathrm{L}}^{\mathrm{T}}。 \tag{3}$$

$\boldsymbol{M} \in \mathrm{R}^{p\times q}$由案件事实和法条文本的语义表征矩阵相乘得到；$M_{i,j}$代表案件事实位置$i$处与法条位置$j$处字符的相关性，数值越大代表相关性越强。由于注意力权重矩阵中数值有正有负，为了规范化，使用softmax函数将权重约束在0~1之间:

$$\alpha_{i,j} = \frac{\exp(M_{i,j})}{\sum_{k=1}^{q}\exp(M_{i,k})}, \tag{4}$$

$$\beta_{i,j} = \frac{\exp(M_{i,j})}{\sum_{k=1}^{p}\exp(M_{k,j})}, \tag{5}$$

其中，$\boldsymbol{\alpha}$是法条方向的注意力权重，$\boldsymbol{\beta}$是案件事实方向的注意力权重。考虑到法条专业性较强，描述用语相对固定，而案件描述具有多样性，于是我们使用案件描述对齐法条，对$\boldsymbol{\beta}$在法条方向求和，压缩注意力权重矩阵，得到案件事实方向上总的权重向量$\boldsymbol{\delta} \in \mathrm{R}^p$:

$$\delta_i = \sum_{j=1}^{q}\beta_{i,j}, i \in [1,...,p]。 \tag{6}$$

$\boldsymbol{\delta}$与法条方向注意力权重$\boldsymbol{\alpha}$融合，得到法条方向的综合权重向量$\boldsymbol{\chi} \in \mathrm{R}^q$:

$$\chi_i = \delta \cdot \alpha_i, i \in [1,...,q]。 \tag{7}$$

使用$\boldsymbol{\chi}$对齐法条语义表征，得到$\boldsymbol{e}_{\mathrm{FL}} \in \mathrm{R}^{q\times d}$:

$$\boldsymbol{e}_{\mathrm{FL}} = \boldsymbol{\chi} * \boldsymbol{e}_{\mathrm{L}}。 \tag{8}$$

得到字符级对齐特征$\boldsymbol{e}_{\mathrm{FL}}$后，使用最大池化策略，保留重要特征得到$\boldsymbol{u} \in \mathrm{R}^d$:

$$\boldsymbol{u} = \max\_\mathrm{pool}(\boldsymbol{e}_{\mathrm{FL}})。 \tag{9}$$

为了使模型更准确地捕获匹配对象间的关键信息，进行相似性判别，并行于BERT最后一层Transformer，堆叠一个关键词自注意力Transformer层。该层的输入为BERT倒数第二层Transformer的输出，其架构与普通Transformer一样，唯一的不同之处是attention mask矩阵。Transformer内部采用自注意力机制:

$$\mathbf{self\_attention} = \mathrm{softmax}\left(\frac{\boldsymbol{Q}\cdot\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}} + \boldsymbol{M}\right), \tag{10}$$

$$M_{ij} = \begin{cases} 0, & \mathrm{mask}_{ij} = 1, \\ -\infty, & \mathrm{mask}_{ij} = 0, \end{cases}$$

其中，**mask**是attention mask矩阵。进行自注意力计算时，可以通过更改attention mask矩阵，使字符间信息相互可见或不可见。基于关键词抽取层抽取的关键词序列，更改attention mask矩阵。如图3所示，假设案件事实有4个字符，Kf2和Kf3为关键词字符，法条有3个字符，Kl2和Kl3为关键词字符，attention mask矩阵中为1的位置代表信息可见，为0的位置代表信息不可见。计算案件事实语义表征时，只关注事实自身信息和法条关键词信息；计算法条语义表征时，只关注法条自身信息以及案件事实关键词信息。

经过关键词注意力层后，得到两个关键词相关语义表征矩阵，然后使用平均池化压缩语义信息，得到两个新的语义表征向量$\boldsymbol{F}_{\mathrm{kw}} \in \mathrm{R}^d$和$\boldsymbol{L}_{\mathrm{kw}} \in \mathrm{R}^d$。将两个向量相减，得到**diff_kw**$\in \mathrm{R}^d$，用来表示关键差

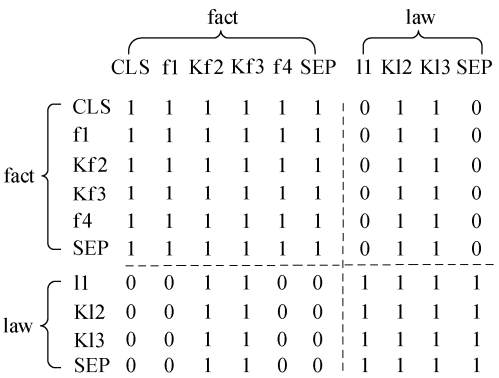

图3　关键词注意力掩码矩阵

Fig. 3　Matrix of keyword attention mask

异特征:

$$\mathbf{diff\_kw} = \boldsymbol{F}_{\mathrm{kw}} - \boldsymbol{L}_{\mathrm{kw}} 。 \tag{11}$$

最后将原始 Transformer 层最后一层的输出 $\boldsymbol{e}$(CLS)、对齐特征 $\boldsymbol{u}$、关键语义表征 $\boldsymbol{F}_{\mathrm{kw}}$, $\boldsymbol{L}_{\mathrm{kw}}$ 以及关键差异特征 **diff_kw** 拼接, 作为匹配特征 $\boldsymbol{H}\in\mathrm{R}^{5d}$, 输入全连接层进行结果预测:

$$\boldsymbol{H} = \mathrm{CAT}(\boldsymbol{e}(\mathrm{CLS}), \boldsymbol{u}, \boldsymbol{F}_{\mathrm{kw}}, \boldsymbol{L}_{\mathrm{kw}}, \mathbf{diff\_kw}) 。 \tag{12}$$

2.6 输出层

将匹配特征输入全连接层进行变换, 然后使用 softmax 激活函数得到预测结果 $\tilde{y}_i \in \{0,1\}$:

$$\tilde{y}_i = \mathrm{argmax}(\mathrm{softmax}(\mathrm{MLP}(\boldsymbol{H}))) , \tag{13}$$

预测结果为 0 代表不匹配, 为 1 代表匹配。

本文使用交叉熵损失函数优化预测结果与真实标签间的误差, 交叉熵损失函数为

$$L = -\sum_{i=0}^{1}[y_i \log p_i + (1-y_i)\log(1-p_i)] , \tag{14}$$

其中, y_i 为真实标签, p_i 是预测概率值。

3 实验设置与结果分析

3.1 实验数据集

本文数据来自中国裁判文书网公开的刑事裁判文书。我们选取两类常见刑事案件(交通肇事和故意伤害案件)作为研究对象, 每类案件下载近 600 份文书, 每份文书中包含犯罪事实和涉及法条等内容。首先进行文本预处理, 然后提取犯罪事实, 与法条建立对应关系。交通肇事类案件有 589 份文书, 包含 4711 个独立犯罪事实, 平均字符数为 63, 涉及法条总数为 56, 平均字符数为 68。故意伤害类案件有 600 份文书, 共 4067 个独立犯罪事实, 平均字符数为 57, 涉及 30 个法条, 平均字符数为 65。裁判文书语料统计信息见表 1。将同类案件中每个犯罪事实与法条集合里的每个法条拼接, 构成若干二元组 $\{F_i, L_j\}$, 标签 $\in\{0, 1\}$, 0 代表不匹配, 1 代表匹配。按照 8:1:1 的比例划分训练集、验证集和测试集。交通肇事数据集包含 197177 条训练数据、26433 条验证数据和 28281 条测试数据。故意伤害数据集包含 74761 条训练数据、8611 条验证数据和 9571 条测试数据。数据集统计信息见表 2。实验数据样例如图 4 所示。

3.2 实验设置

案件事实以及法条序列的最大长度都设置为 60, 词向量基于所有下载的文书和法条内容, 使用 CBOW[31]模型进行训练, 维度设置为 128。选择清华大学发布的基于大规模刑事文书预训练的 XS-BERT[30]作为编码器, 匹配阈值设置为 0.5, 训练轮数设置为 10, 学习率设置为 1×10^{-5}, dropout 设置为 0.1。连续训练 3 轮验证集后, 若损失仍然不下降, 则停止训练, 使用 softmax 函数作为激活函数, 使用交叉熵损失函数和 Adam 优化器更新模型参数。考虑到所用的法条匹配数据集存在数据不平衡现象,

表 1 裁判文书语料统计信息

Table 1 Statistics of judgement documents corpus

裁判文书案件类型	案例	犯罪事实(平均字符数)	法条(平均字符数)
交通肇事类	589	4711 (63)	56 (68)
故意伤害类	600	4067 (57)	30 (65)

表 2 数据集统计信息

Table 2 Statistics of datasets

数据集	交通肇事数据集	故意伤害数据集
训练集	197177	74761
验证集	26433	8611
测试集	28281	9571

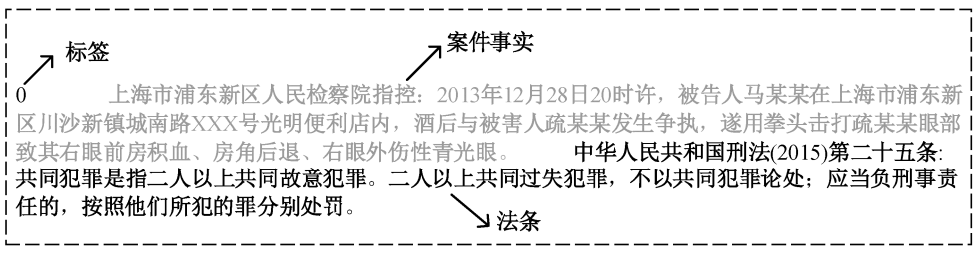

图 4 实验数据样例

Fig. 4 Example of experimental data

负例样本多于正例, 将交通肇事数据集和故意伤害数据集正负样本采样比分别设置为 1:12 和 1:6, 以便获得最优 F1 值, 对比匹配模型采用同样的采样策略。本文实验中用的 BERT 模型为 BERT-Base-Chinese。

3.3 实验结果与分析

3.3.1 对比实验

为了验证本文模型 BiAKLaw 的有效性, 从文本分类和文本匹配两个方面设置对比实验。文本分类模型选择主流分类模型 TextCNN[32]、LSTM[33]和 BERT, 模型的输入为案件事实, 涉及的法条为标签, 是多标签分类任务。本研究将法条之间看成是相互独立的, 将问题转化为多个单法条推荐的二分类问题。选择 sigmoid 作为激活函数。文本匹配模型选择经典深度语义匹配模型 ARC-II[21], ESIM[22]和 RE2[34]以及预训练模型 BERT, XS-BERT 和 SBERT[35]。ARC-II 使用 CNN 提取文本特征, 在第一层卷积后交互两个文本特征。RE2 模型包含原始词嵌入特征、先前对齐特征和上下文特征 3 个关键特征, 并且简化了其他组件。ESIM 使用 LSTM 获取句子语义表示向量, 再用注意力机制对齐特征。BERT 是通用领域的预训练语言模型。XS-BERT 是基于大规模刑事文书预训练的领域专用预训练模型, SBERT 为拥有孪生结构的预训练模型。

模型的输出有两种情况, 匹配或不匹配, 实际上是二分类问题。因此, 我们采用 P, R 和 F1 作为模型评价指标:

$$\mathrm{F1}=\frac{2PR}{P+R}, \tag{15}$$

其中, P 是精确度, R 是召回率, F1 为综合评价指标。

将本文模型 BiAKLaw 和对比模型在交通肇事和故意伤害数据集上进行对比实验, 在测试集上的实验结果如表 3 所示。可以看出, BiAKLaw 在各项指标上都取得最优结果, 在交通肇事数据集上, 与表现次之的 XS-BERT 模型相比, BiAKLaw 的 F1 评价指标提升 3.74%, 在故意伤害数据集上提升 3.43%。这是因为与 XS-BERT 模型相比, BiAKLaw 具有更强的识别案件事实和法条关键信息差异的能力, 通过双向注意力机制得到匹配对之间更深层次的交互特征, 关键词信息的注入使得模型能够过滤干扰信息, 更关注匹配对中的关键语义信息。孪生结构预训练模型 SBERT 在训练时间上有一定的优势, 但模型性能弱于交互式的通用型 BERT 和法律领域专用的 XS-BERT。从表 3 还可以看出, 文本分类模型的性能整体上明显弱于文本匹配模型, 这是因为同类型案件事实描述差异较小, 分类模型难以区分细微的语义差别。法条类别众多, 每个犯罪事实涉及法条个数不定, 也给分类模型带来更多的挑战。

表 3　BiAKLaw 与对比模型实验结果

Table 3　Experiment results of BiAKLaw and comparative models

模型	交通肇事数据集			故意伤害数据集		
	P	*R*	F1	*P*	*R*	F1
TextCNN	0.734	0.713	0.723	0.742	0.708	0.724
LSTM	0.725	0.706	0.715	0.730	0.712	0.721
BERT	0.755	0.718	0.736	0.774	0.713	0.742
ARC-II	0.923	0.831	0.875	0.869	0.877	0.873
RE2	0.892	0.780	0.832	0.869	0.805	0.836
ESIM	0.896	0.781	0.835	0.855	0.851	0.853
SBERT	0.900	0.848	0.873	0.884	0.912	0.898
BERT	0.903	0.852	0.877	0.890	0.909	0.899
XS-BERT	0.900	0.866	0.882	0.894	0.915	0.904
BiAKLaw	**0.924**	**0.905**	**0.915**	**0.932**	**0.938**	**0.935**

说明: 粗体数字表示性能最优, 下同。

3.3.2 模块消融实验

为了探索各个模块的加入对模型性能产生的影响, 我们将 XS-BERT、XS-BERT+双向注意力机制、XS-BERT+关键词注意力机制以及 XS-BERT+双向注意力机制+关键词注意力机制 4 个模型在交通肇事和故意伤害数据集上进行对比实验, 结果如表 4 所示。可以看出, 双向注意力机制和关键词注意力机制模块的加入都对模型产生积极的影响, 其中关键词注意力机制的影响最大。在交通肇事数据集上, 双向注意力机制的加入使得评价指标 F1 比 XS-BERT 提升 0.79%, 在故意伤害数据集上提升 0.77%。在交通肇事数据集上, 关键词注意力机制模块的加入使得评价指标 F1 比 XS-BERT 提升 1.13%, 在故意伤害数据集提升 2.76%。实验结果说明, 通过关键词信息的加入, 模型能够更多地关注匹配对象间的关键差异特征, 减少无关信息的干扰。

3.3.3 法条推荐实例分析

我们选取故意伤害测试集中的一个案例来说明本文方法的有效性, 案例分析结果如表 5 所示。

表 4 消融实验结果

Table 4 Result of ablation experiments

模型	交通肇事数据集			故意伤害数据集		
	P	*R*	F_1	*P*	*R*	F1
XS-BERT	0.900	0.866	0.882	0.894	0.915	0.904
XS-BERT+双向注意力机制	0.903	0.877	0.889	0.898	0.925	0.911
XS-BERT+关键词注意力机制	0.912	0.872	0.892	0.908	**0.951**	0.929
XS-BERT+双向注意力机制+关键词注意力机制(BiAKlaw)	**0.924**	**0905**	**0.915**	**0.923**	0.938	**0.935**

表 5 故意伤害案例分析

Table 5 Case analysis of intentional injury

案例	法律条文	真实标签	预测概率值	
			BiAKLaw	XS-BERT
事实 a**:** …被告人范祚南因琐事与被害人邓某在佛山市三水区乐平镇范湖东源村某巷 2 号发生**争执**并**扭打**起来, 随后被害人邓某从屋内拿出羊角锤**追赶**被告人范祚南, 被告人范祚南使用垃圾铲**击打**被害人邓某, 造成被害人邓某左顶部和左额发际处被**打击致伤**。	刑法 234 条: **故意伤害**他人身体的, 处三年以下有期徒刑、拘役或者管制。犯前款罪, 致人**重伤**的, 处三年以上十年以下有期徒刑; 致人**死亡**或者以特别**残忍手段**致人**重伤**造成**严重残疾**的, 处十年以上有期徒刑、无期徒刑或者死刑。本法另有规定的, 依照规定。	1	0.923	0.902
事实 b**:** 经**鉴定**, 被害人邓某的**损伤程度**为**轻伤**。	刑法 72 条: 对于被判处拘役、三年以下有期徒刑的犯罪分子, 同时符合下列条件的, 可以宣告缓刑, 对其中不满十八周岁的人、怀孕的妇女和已满七十五周岁的人, 应当宣告缓刑: (一)犯罪**情节较轻**; (二)有**悔罪表现**; (三)没有**再犯罪**的危险; (四)宣告缓刑对所居住社区没有重大不良影响。宣告缓刑, 可以根据犯罪情况, 同时禁止犯罪分子在缓刑考验期限内从事特定活动, 进入特定区域、场所, 接触特定的人。被宣告缓刑的犯罪分子, 如果被判处附加刑, 附加刑仍须执行。	1	0.919	0.895
	刑法 234 条: **故意伤害**他人身体的, 处三年以下有期徒刑、拘役或者管制。犯前款罪, 致人**重伤**的, 处三年以上十年以下有期徒刑; 致人**死亡**或者以特别**残忍手段**致人**重伤**造成**严重残疾**的, 处十年以上有期徒刑、无期徒刑或者死刑。本法另有规定的, 依照规定。	1	0.907	0.873
事实 b**:** 经**鉴定**, 被害人邓某的**损伤程度**为**轻伤**。	刑法 234 条: **故意伤害**他人身体的, 处三年以下有期徒刑、拘役或者管制。犯前款罪, 致人**重伤**的, 处三年以上十年以下有期徒刑; 致人**死亡**或者以特别**残忍手段**致人**重伤**造成**严重残疾**的, 处十年以上有期徒刑、无期徒刑或者死刑。本法另有规定的, 依照规定。	0	0.102	0.134

说明: 粗体文字为案件事实和法条的关键词。

该案例包括两个犯罪事实: 事实 a 涉及刑法 234 条, 事实 b 涉及刑法 72 和 234 条。事实 a 中, 被告人与被害人有“争执”和“击打”等行为, 符合刑法 234 条“故意伤害”。事实 b 中, 被害人伤势鉴定为“轻伤”, 符合刑法 234 条“故意伤害”行为和刑法 72 条中的“情节较轻”部分。表 5 还给出案件事实与法条在 BiAKLaw 和 XS-BERT 中的匹配概率值。可以看出, 两个模型预测的概率值都大于 0.5, 均能给出正确的预测标签。本文模型 BiAkLaw 有关键词信息的注入, 能更多地关注关键词信息, 提升了匹配性能。对于不匹配的事实 b 与刑法 234 条, 本文模型 BiAKLaw 比 XS-BERT 更易辨别出两者是不相匹配的。

需要指出的是, 法条推荐任务的目的是根据案件事实的细节描述, 客观地推荐与之相关联的一个或多个法条。最终如何判决, 需要法官根据推荐的法条及其他因素综合决定。从案例分析结果可以看出, 案件事实和法条序列各自的关键词信息不仅可以帮助模型关注犯罪事实和法条关键信息的细节差异, 还能直观地体现为什么这个犯罪事实会对应这

个法条, 有效地增强了推荐结果的可解释性。

4 结语

本文将法条推荐任务转换成犯罪事实与法条语义匹配度计算, 提出一种基于双向注意力文本关键词匹配的法条推荐模型 BiAKLaw。通过双向文本关键词注意力机制, 既保留了原本完整的语义信息, 又使模型更多地关注两者之间的关键差异特征, 捕获 token 级对齐特征和 keyword 差异特征, 输出端融合多粒度匹配特征, 增强了模型匹配效果。由于双向融合了法条文本和案件事实的各自关键词作为推荐结果的重要依据, 使得本文法条推荐方法具有良好的可解释性。在真实数据集上的实验结果表明, 与主流分类模型和深度经典语义匹配模型相比, BiAKLaw 模型的性能均有不同程度的提升, 因此, 对智慧司法领域法条推荐任务的研究有一定的启示意义。

在未来的工作中, 我们将尝试融合外部知识 (例如案例法律特征、法条结构特征和庭审观点等), 持续优化法条推荐模型。

参考文献

[1] Kort F. Predicting supreme court decisions mathematically: a quantitative analysis of the “Right to Counsel” cases. American Political Science Review, 1957, 51(1): 1–12

[2] Ulmer S S. Quantitative analysis of judicial processes: some practical and theoretical applications. Law and Contemporary Problems, 1963, 28(1): 164–184

[3] Nagel S S. Applying correlation analysis to case prediction. Tex L Rev, 1963, 42: 1006

[4] Keown R. Mathematical models for legal prediction. Computer/Law Journal, 1980, 2(1): 829

[5] Tsoumakas G, Katakis I. Multi-label classification: an overview. International Journal of Data Warehousing and Mining (IJDWM), 2007, 3(3): 1–13

[6] Nam J, Kim J, Loza Mencía E, et al. Large-scale multi-label text classification — revisiting neural networks // Proceedings of Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Nancy, 2014: 437–452

[7] Wang T, Liu L, Liu N, et al. A multi-label text classification method via dynamic semantic representation model and deep neural network. Applied Intelligence, 2020, 50: 2339–2351

[8] Liu H, Chen G, Li P, et al. Multi-label text classification via joint learning from label embedding and label correlation. Neurocomputing, 2021, 460: 385–398

[9] Luo B, Feng Y, Xu J, et al. Learning to predict charges for criminal cases with legal basis [EB/OL]. (2017–07–28) [2023–03–10]. https://doi.org/10.48550/arXiv.1707.09168

[10] Zhong H, Guo Z, Tu C, et al. Legal judgment prediction via topological learning // Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, 2018: 3540–3549

[11] Hu Z, Li X, Tu C, et al. Few-shot charge prediction with discriminative legal attributes // Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe, 2018: 487–498

[12] Wang S, Jiang J. A compare-aggregate model for matching text sequences [EB/OL]. (2016–11–06) [2023–02–26]. https://arxiv.org/abs/1611.01747

[13] Ge J, Huang Y, Shen X, et al. Learning fine-grained fact-article correspondence in legal cases. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2021, 29: 3694–3706

[14] Liu C L, Hsieh C D. Exploring phrase-based classification of judicial documents for criminal charges in Chinese // International Symposium on Methodologies for Intelligent Systems. Bari, 2006: 681–691

[15] Chen H, Cai D, Dai W, et al. Charge-based prison term prediction with deep gating network [EB/OL]. (2019–08–30) [2023–03–13]. https://doi.org/10.48550/arXiv.1908.11521

[16] Pan S, Lu T, Gu N, et al. Charge prediction for multi-defendant cases with multi-scale attention // Computer Supported Cooperative Work and Social Computing: 14th CCF Conference, ChineseCSCW 2019. Kunming, 2019: 766–777

[17] Xiao C, Zhong H, Guo Z, et al. Cail2018: a large-scale legal dataset for judgment prediction [EB/OL]. (2018–07–04) [2023–03–12]. https://doi.org/10.48550/arXiv.1807.02478

[18] Chalkidis I, Androutsopoulos I, Aletras N. Neural legal judgment prediction in English [EB/OL]. (2019–06–05) [2023–03–21]. https://doi.org/10.48550/arXiv.1906.02059

[19] Robertson S, Zaragoza H. The probabilistic relevan-

ce framework: BM25 and beyond. Foundations and Trends in Information Retrieval, 2009, 3(4): 333–389

[20] Huang P S, He X, Gao J, et al. Learning deep structured semantic models for web search using clickthrough data // Proceedings of the 22nd ACM International Conference on Information & Knowledge Management. San Francisco, 2013: 2333–2338

[21] Hu B, Lu Z, Li H, et al. Convolutional neural network architectures for matching natural language sentences // Advances in Neural Information Processing Systems. Montreal, 2014: 2042–2050

[22] Chen Q, Zhu X, Ling Z, et al. Enhanced LSTM for natural language inference [EB/OL]. (2016–09–20) [2023–03–24]. https://doi.org/10.48550/arXiv.1609.06038

[23] Lyu B, Chen L, Zhu S, et al. LET: linguistic knowledge enhanced graph transformer for Chinese short text matching [C/OL]. (2021–02–05) [2023–02–20]. https://doi.org/10.48550/arXiv.2102.12671

[24] Wu Y, Wu W, Xu C, et al. Knowledge enhanced hybrid neural network for text matching [C/OL]. (2016–11–14) [2023–03–28]. https://doi.org/10.48550/arXiv.1611.04684

[25] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need // Advances in Neural Information Processing Systems. Long Beach, 2017: 5998–6008

[26] Devlin J, Chang M W, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, 2019: 4171–4186

[27] Liu Y, Ott M, Goyal N, et al. RoBERTa: a robustly optimized bert pretraining approach [EB/OL]. (2019–07–26) [2023–03–01]. https://doi.org/10.48550/arXiv.1907.11692

[28] Cui Y, Che W, Liu T, et al. Pre-training with whole word masking for chinese BERT. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3504–3514

[29] Grootendorst M. KeyBERT: minimal keyword extraction with BERT [EB/OL]. (2020–02–09) [2023–01–09]. https://github.com/MaartenGr/KeyBERT

[30] Zhong H, Zhang Z, Liu Z, et al. Open Chinese language pre-trained model zoo [EB/OL]. (2019–07–01) [2023–02–11]. https://github.com/thunlp/OpenCLaP

[31] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space [EB/OL]. (2013–01–16) [2023–02–11]. https://doi.org/10.48550/arXiv.1301.3781

[32] Cho K, Van Merrienboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation [EB/OL]. (2014–09–02) [2023–03–11]. http://arxiv.org/abs/1406.1078

[33] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735−1780

[34] Yang R, Zhang J, Gao X, et al. Simple and effective text matching with richer alignment features // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, 2019: 4699–4709

[35] Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using Siamese BERT-Networks [EB/OL]. (2019–08–27) [2023–03–23]. http://arxiv.org/abs/1908.10084