面向借贷案件的相似案例匹配模型

曹发鑫，孙媛媛*，王治政，潘丁豪，林鸿飞
（大连理工大学计算机科学与技术学院，辽宁 大连 116024）

摘　要：相似案例匹配任务是文本匹配在司法领域的具体应用之一，目的在于区分法律文书是否相似，对类案检索具有重要意义。与传统文本匹配任务相比，法律文本通常篇幅较长，同时相似案例匹配是针对相同案由案件的匹配，案情文本之间的差异较小，以往的文本匹配方法很难计算文本相似度。针对借贷案件文本匹配存在的问题，建立一种融合借贷案件关键要素的相似案例匹配模型。为了获取文本中更丰富的语义特征，构建正则表达式获得借贷案件的特定案件要素，如借款交付形式、借款人基本属性等，并与原有的案情文本相结合，联合学习法律文本与案件关键要素的语义特征。同时，利用共享权重的预训练模型分别对不同的文书进行编码，并且对预训练模型特定编码层的输出进行融合，得到更加丰富的语义信息。引入有监督对比学习框架，更好地利用样本信息，进一步提高相似案例匹配的性能。在CAIL2019-SCM数据集上的实验结果表明，与LFESM模型相比，该模型在测试集上的准确率提高了1.05个百分点。

关键词：相似案例匹配；孪生网络；对比学习；预训练模型；法律关键要素
中图分类号：TP391　　**文献标志码：**A　　**DOI:** 10.19678/j.issn.1000-3428.0066055

Similar Case Matching Model for Lending Cases

CAO Faxin, SUN Yuanyuan*, WANG Zhizheng, PAN Dinghao, LIN Hongfei
(School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, Liaoning, China)

【Abstract】The purpose of Similar Case Matching(SCM) is to distinguish whether legal documents are similar, which is a specific application of text matching and is vital to the retrieval of similar cases. Compared with conventional texts, legal texts are typically longer, and SCM aims to realize matching for the same case. Moreover, the difference between case texts is negligible; therefore, calculating text similarity using previous text-matching methods is challenging. This study establishes a SCM model that integrates key elements of lending cases to address the issues of text matching in lending cases. To obtain richer semantic features from texts, regular expressions are constructed to obtain specific case elements of lending cases, such as the loan-delivery form and the basic attributes of borrowers, which are then combined with the original case text to jointly learn the semantic features of the legal text and key elements of the case. Additionally, pretrained models with shared weights are used to encode different instruments separately, and the outputs of specific encoding layers of the pretrained models are fused to obtain richer semantic information. Finally, the proposed model incorporates a supervised comparison learning framework to utilize the text information more effectively and further improve the performance of SCM. Experiments on the CAIL2019-SCM dataset show that this model improves the accuracy of the test set by 1.05 percentage points compared with LFESM models.

【Key words】Similar Case Matching(SCM); Siamese network; contrastive learning; pretrained model; key legal element

0　引言

相似案例匹配(SCM)旨在从以往的裁判文书中找出与当前审判案件相似的文书作为判决的重要依据，有效提升了司法部门的办案效率[1]。在之前的工作中，法律从业人员需要通过人工检索的方式寻找相似案件。截至2022年，最高人民法院在中国裁判文书网上公布了上亿篇裁判文书，对类似案件的发展提供了很大的帮助。利用相似案例匹配实现这些裁判文书的筛选，在节省成本的同时提高了查找效率。

中国裁判文书网上公布的裁判文书覆盖刑事、民事、行政等不同案件类型。同时，每类案件中又根据案由分为上百个类别。值得注意的是不同原因的案件在描述和结构上有所不同。相似案例匹配通常针对的是具有相同案由的案件。我国法律文件通常由以下部分组成：本案当事人的信息，之前的诉讼记录，案情描述，包含对案情描述分析在内的庭审意见，最终裁决以及相关的法律条文。案情描述部分包括案件的细节和相应的证据。一般而言，判断两个案例的相似性，往往等同于判断两个案例案情事实之间的相似性，这是因为案情事实描述比其他

收稿日期：2022-10-20　　**修回日期：**2023-01-11
基金项目：国家重点研发计划(2022YFC3301801)；中央高校基本科研业务费专项资金(DUT22ZD205)。
通信作者 E-mail：*syuan@dlut.edu.cn

部分包含更多的信息。

作为相似案例匹配的核心技术，文本匹配是自然语言处理的基础任务之一，可以应用到很多下游任务中，如信息检索、问答系统、查重系统等。文本匹配包含众多子任务，包括自然语言推理（NLI）、智能问答等。自然语言推理主要判断句子在语义上的关系，一般可分为蕴含、矛盾、中立。相似案例匹配和自然语言推理都是关注文本的相似度，但是与自然语言推理相比，相似案例匹配面临着一些挑战，具体如下：

1）法律文本通常具有篇幅较长的特点，以 2019 年“中国法研杯”相似案例匹配数据集[2]为例，文本的平均长度达到 679 个字符，现有的文本匹配方法很难处理长文本匹配。

2）相似案例匹配是在相同案由案件上的匹配，这些案件文本之间的差异很小，传统的文本匹配模型很难判断两个文本是否相似。主体、金额、利率等差异最终会导致判决的不同。例如，在私人贷款的情况下，利率有 3 种情况：若没有承诺利息，则法官不支持利息索赔；当利率高于 24% 时，法院不支持利率过高的部分；当利率高于 36% 时，借款人可以要求返还超额部分。

3）传统的文本匹配方法通常采用余弦相似度来衡量文本之间的相似性，但经过 BERT[3]编码的词向量因受到各向异性的影响不适合计算文本的余弦相似度。

针对上述问题，本文针对相似案例匹配任务建立融合法律关键要素的相似案例匹配模型。Sentence-BERT[4]模型是对预训练的 BERT 模型进行修改，使用孪生网络结构来获得更优的句向量，进行下游任务的相似度计算。以 Sentence-BERT 作为基本结构，案件信息在对文本编码的同时引入案情关键要素，使得模型在进行文本匹配的过程中更多地关注案件要素等细粒度信息，进而提升模型的匹配效果。在训练过程中，引入有监督对比学习框架，进一步提升模型性能。

1 相关工作

传统文本匹配注重对文本基本特征的提取[5]，需要依赖人工选择特征，可解释性好。但是这些基于词汇重合度的匹配算法只停留在字词匹配层面，泛化能力一般。对于文本匹配，更需要关注语义层的信息。在信息检索中，很多工作针对传统模型进行改进，例如将 LDA 模型应用到检索任务中[6]。

基于深度学习的文本匹配主要可以分为以下两个发展方向[7]：基于表示型的文本匹配模型和基于交互型的文本匹配模型。基于表示型的文本匹配模型主要关注构建模型的表示层。HUANG 等[8]提出 DSSM 模型，在训练阶段分别用复杂的深度学习网络构建待匹配的两个不同的文本向量，通过计算两个语义向量的余弦相似度来表征语义相似度。SHEN 等[9]提出 CDSSM 模型，在单词序列上加入卷积结构，从文本中分别提取词和句子级别的语境特征。基于交互型的文本匹配模型舍弃了以往工作在最后模块中进行语义匹配的思路，在输入层就对文本进行交互，并且将交互信息作为模型输入进行后续建模。PANG 等[10]提出 MatchPyramid 模型，借鉴了图像识别的方式进行文本匹配，首先构建了文本之间的相似度矩阵，然后采用卷积神经网络提取交互信息，因此可以学习到文本之间的多维度特征。CHEN 等[11]基于 NLI 任务提出一种有效的交互型匹配方法（ESIM），采用 BiLSTM 和 Tree-LSTM 分别对文本序列和文本解析树进行编码，在匹配层采用注意力机制来得到词向量之间的详细交互关系，聚合层通过 BiLSTM 得到各文本向量之间的编码，从而进一步增强了文本序列的信息传递。

基于预训练模型的研究方法是近年来 NLP 任务主流的研究方法之一。DEVLIN 等[12]在超大规模数据集的基础上，提出基于自注意力机制的双向 Transformer[13]框架（BERT），在 11 个 NLP 任务中达到 SOTA 效果。REIMERS 等[4]提出 Sentence-BERT，对 BERT 进行改进，使用孪生网络结构来获得语义上有意义的句向量，采用余弦相似度进行比较找到语义相似的句子。LI 等[14]在 Sentence-BERT 的基础上提出 BERT-flow，分析 BERT 句向量分布的性质，然后利用标准化流无监督地将 BERT 句向量的分布变换成更规整的高斯分布。SU 等[15]通过降低语义向量的维度提高 BERT 在语义向量相似度计算方面的效果。PEINELT 等[16]将主题模型和 BERT 结合实现了语义相似度分析，在特定领域上有优异的表现。GAO 等[17]提出 SimCSE，在 BERT 的基础上运用对比学习的框架极大提升了无监督文本匹配的性能。与传统方法相比，基于 BERT 的文本匹配方法虽然已取得不错的效果，但在判断法律文书等结构相似的文本之间的相似度方面仍然具有较大的改进空间。因此，本文在 Sentence-BERT 的基础上，融入案件文本的关键要素，提升模型对于法律文书中细微特征的识别能力，从而提升模型匹配效果。

2 模型结构

2.1 相似案例匹配任务

相似案例匹配任务本质在于确定给定案例文本之间的相似性，输入数据格式为<A，B，C>，A、B、C 是 3 篇有关民间借贷案件的裁判文书，A 是已有文书，B 和 C 是两篇候选的文书。此任务的目标为确定 B 与 C 之间与 A 文本更相似的文本。通过预先定义的评价分数来衡量两篇候选文书 B、C 与 A 之间的相似度得分，当 B 的得分高于 C 的得分时，认为 B 相较于 C 与 A 更相似，标签记为 0 反之标签为 1。

2.2 模型结构

基于Sentence-BERT在句向量表示上的优秀表现，建立针对三元组进行改进的司法相似案例匹配模型，模型总体结构如图1所示。模型主要由以下4个部分组成：特征抽取层，编码层，交互层和输出层。同时，引入对比学习模块来进一步提升模型性能。

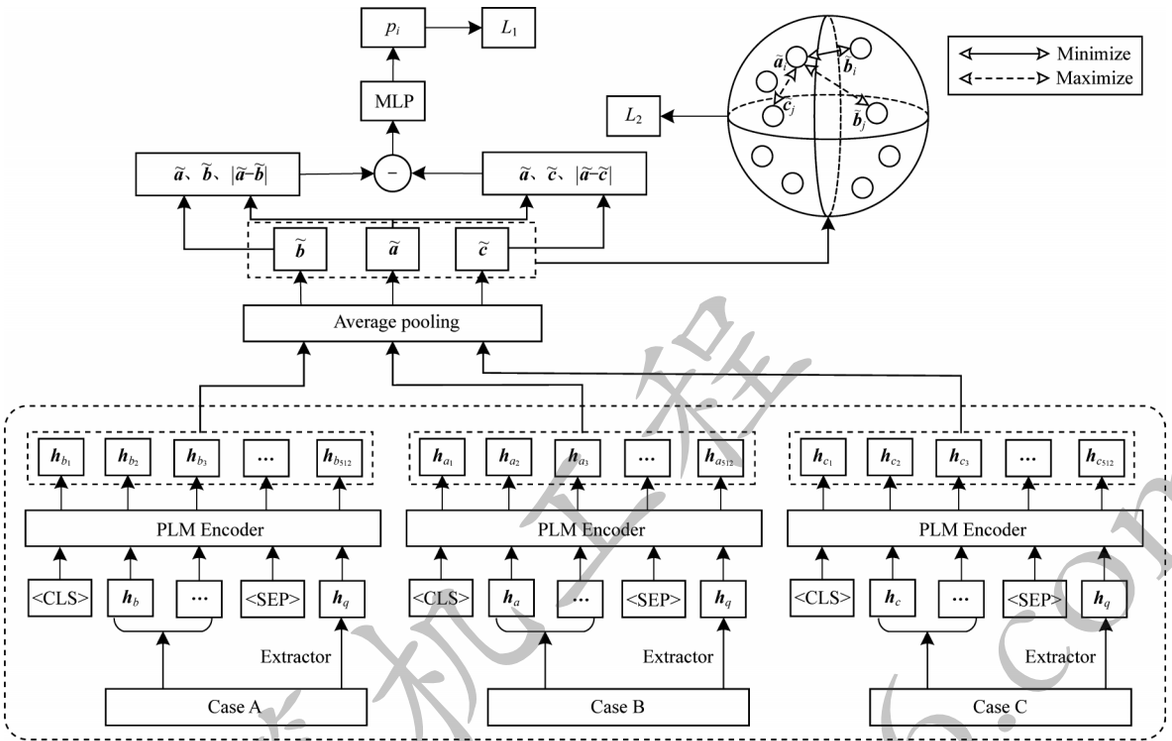

图1 相似案例匹配模型结构

Fig.1 Structure of similar case matching model

2.2.1 特征抽取层

在将案例文本进行编码前，先从案例文本中抽取法律特征属性，并且将这些属性和原文本拼接进输入模型进行学习。这么做的原因为法律文书有固定结构且用词严谨、语义唯一，导致案件文本之间的差异很微小，文书之间的相似性会聚焦于一些细粒度的特征。因此，按照案由（民间借贷）以及参考《合同法》、《担保法》、《婚姻法》相关司法解释，总结具体的法律特征属性。在通常情况下，相似案例会共享相同的法律特征。在借贷案件中，法律特征要素会影响具体判决，如表1所示，所有这些属性都是由法律术语或数字组合而成，它们在法律文本中按照一定的范式存在。因此，使用正则表达式来提取这些具体的特征，这些特征可以更好地帮助模型判断法律文本之间的相似性。

表1 法律文本特征

Table 1 Legal text features

法律特征要素	具体类型
借款交付形式	现金、网上电子汇款、银行转账、票据、未出借、网络贷款平台
借款人基本属性	法人、自然人
借款用途	个人生活、企业生产经营、违法犯罪、夫妻共同生活
借贷依据	借款合同、借条、微信等聊天记录,担保,欠条
出借人基本属性	法人、自然人
出借意图	正常出借、转贷牟利
担保类型	抵押、质押、无担保、保证
约定期内利率	24%以下、24%~36%、36%以上
约定计息方式	单利、复利、无利息
还款交付形式	部分还款、现金、未还款、网上电子汇款、银行转账

2.2.2 编码层

选用BERT编码器来生成下游任务所需的词向量。利用预训练模型在进行编码时，选择将案例文本与上节中抽取到的法律要素进行拼接，同时编码得到词向量。

模型编码部分主要分为三部分：词嵌入层，段嵌入层，位置嵌入层，其中任务层和编码层是通用结构，对任何任务都适用。对于具体的长度为l的文本

$a=\{a_1,a_2,\cdots,a_l\}$，长度为 s 的特征 $q=\{q_1,q_2,\cdots,q_s\}$，因为预训练模型的输入文本限制长度不超过512，而文书中重要的信息包括法律案件的事实描述、法院裁判结果等都在文书的后半部分，因此选择从后半部分截取文书的内容作为文书输入。选择将文书与特征进行拼接，在句首加入“[CLS]”标识符，同时使用“[SEP]”将文本与特征分割，并将此作为模型的输入。需要注意的是，为了充分考虑文本之间的交互，使用3个共享权重的BERT模型。

BERT模型在不同层的Transformer编码器中能够捕获丰富的语言信息[18]，同时研究表明BERT的相似层之间能够学习到相似的语言知识。因此，选择BERT特定编码层的输出进行融合，可以得到更加丰富的文本特征。利用BERT模型的第1、5、9、12层作为增强特征，定义BERT模型的隐藏层状态向量 $\boldsymbol{H}=[\boldsymbol{H}_{i,1},\boldsymbol{H}_{i,2},\cdots,\boldsymbol{H}_{i,12}]$，其中 i 表示A、B、C文本，因此经过BERT的输出如下：

$$\bar{\boldsymbol{a}}=\sum \boldsymbol{H}_{j,a} \tag{1}$$

$$\bar{\boldsymbol{b}}=\sum \boldsymbol{H}_{j,b} \tag{2}$$

$$\bar{\boldsymbol{c}}=\sum \boldsymbol{H}_{j,c} \tag{3}$$

其中：$j\in[1,5,9,12]$；$\bar{\boldsymbol{a}}$、$\bar{\boldsymbol{b}}$、$\bar{\boldsymbol{c}}\in\mathbb{R}^{m\times n\times d}$，$m$ 为输入的batch数量，n 为一个batch中包含的所有句子的字符总和，d 为词嵌入层的隐藏层大小。

2.2.3 交互层

通过BERT编码层后，特征向量 $\bar{\boldsymbol{a}}$、$\bar{\boldsymbol{b}}$、$\bar{\boldsymbol{c}}$ 之间进行交互。在交互之前，参考Sentence-BERT模型在经过BERT编码层后添加了一个平均池化操作，以获得固定大小的句向量 $\tilde{\boldsymbol{a}}$、$\tilde{\boldsymbol{b}}$、$\tilde{\boldsymbol{c}}$。得到句子嵌入 $\tilde{\boldsymbol{a}}$、$\tilde{\boldsymbol{b}}$、$\tilde{\boldsymbol{c}}$ 后，选择将 $\tilde{\boldsymbol{a}}$、$\tilde{\boldsymbol{b}}$ 和 $\tilde{\boldsymbol{a}}$、$\tilde{\boldsymbol{c}}$ 分别进行连接，具体连接方式如下：

$$\boldsymbol{U}_{ab}=[\tilde{\boldsymbol{a}};\ \tilde{\boldsymbol{b}};\ |\tilde{\boldsymbol{a}}-\tilde{\boldsymbol{b}}|] \tag{4}$$

$$\boldsymbol{U}_{ac}=[\tilde{\boldsymbol{a}};\ \tilde{\boldsymbol{c}};\ |\tilde{\boldsymbol{a}}-\tilde{\boldsymbol{c}}|] \tag{5}$$

利用 $|\tilde{\boldsymbol{a}}-\tilde{\boldsymbol{b}}|$、$|\tilde{\boldsymbol{a}}-\tilde{\boldsymbol{c}}|$ 能够学习到词级别的相关语义，有助于进行更充分的匹配。

2.2.4 输出层

交互层得到的 $\boldsymbol{U}_{ab}$ 和 $\boldsymbol{U}_{ac}$ 分别表示A、B的相似度和A、C的相似度，将两者相减可以对比出在各种不同维度上A、B还是A、C的相似度更高，最终线性层对这些维度进行加权，得到一个综合的相似度判定。在模型中，将两者差值输入到多层感知分类器(MLP)中进行预测：

$$p_i=\mathrm{MLP}(\boldsymbol{U}_{ab}-\boldsymbol{U}_{ac}) \tag{6}$$

其中：p_i 表示预测标签为 y_i 的概率值，y_i 的值为0或1，0代表A、B文本的相似度大于A、C文本的相似度，1则反之。整个模型是端到端模型，采用交叉熵损失函数进行训练：

$$L_1=-\sum_{i=0}^{1}[y_i\log_a p_i+(1-y_i)\log_a(1-p_i)] \tag{7}$$

2.2.5 对比学习层

除交叉熵损失外，另外添加基于对比学习的损失来计算文本之间的相似度得分。对比学习的核心思想是拉近相似样本的距离，拉远不相似样本的距离。相似样本的构造可以分为有监督和无监督两种，本文采用有监督对比学习，通过将监督样本中的相似样本作为正样本，不相似样本作为负样本，进行对比学习。

选择经过池化层后的向量 $\tilde{\boldsymbol{a}}$、$\tilde{\boldsymbol{b}}$、$\tilde{\boldsymbol{c}}$ 为对比学习框架的输入，需要注意的是 $\tilde{\boldsymbol{a}}$、$\tilde{\boldsymbol{b}}$、$\tilde{\boldsymbol{c}}$ 各自包含的是一个batch内的文本，$\tilde{\boldsymbol{a}}_i$、$\tilde{\boldsymbol{b}}_i$、$\tilde{\boldsymbol{c}}_i(0\leqslant i\leqslant N)$ 分别表示一个样本内的A、B、C文本，N 表示batchsize。在一个样本内，A和B文本为默认相似的文本，A与C则相反。损失函数如下：

$$L_2=\frac{\mathrm{e}^{\mathrm{Sim}(\tilde{\boldsymbol{a}}_i,\tilde{\boldsymbol{b}}_i)/\tau}}{\sum_{j=1}^{N}(\mathrm{e}^{\mathrm{Sim}(\tilde{\boldsymbol{a}}_i,\tilde{\boldsymbol{b}}_j)/\tau}+\mathrm{e}^{\mathrm{Sim}(\tilde{\boldsymbol{a}}_i,\tilde{\boldsymbol{c}}_j)/\tau})} \tag{8}$$

其中：τ 为超参数，是一个大于0的常数；Sim为两个文本的相似度计算，采用余弦相似度进行计算。采用联合学习方法进行模型训练，最终损失如下：

$$L=L_1+\alpha L_2 \tag{9}$$

其中：α 为权重，本文设置为0.5。

3 实验与结果分析

3.1 数据集

CAIL2019-SCM[2]是用于相似案例匹配的数据集，该数据集由8 138个三元组裁判文书构成，所有的裁判文书来自中国裁判文书网，并且与民间借贷案件有关。每个三元组的格式为<A，B，C>，其中，A为待比较文书，B、C为候选文书。在训练集中，A和B的相似度默认大于A和C的相似度，验证集和测试集中标签是不确定的。裁判文书中包含事实描述部分，文本长度为500~800个字符。三元组内文本之间的相似度由专业的法律从业者根据以往经验进行标注。表2显示了数据集的数量以及样本分布，可以看出数据的标签分布基本是平衡的。

表2 CAIL2019-SCM数据集设置

Table 2 Setting of the CAIL2019-SCM dataset 单位：个

数据集	正样本数	负样本数	总计
训练集	2 596	2 506	5 102
验证集	837	663	1 500
测试集	803	733	1 536

3.2 评价标准

对于测试数据，通过对B和C文本的相似度的计算预测结果标签，A和B文本相似则标签为0，A和C文本相似则标签为1(数据集可以保证每个三元组中B和C文本必然包含一篇与A相似的文本和一篇不相似的文本)。通过与真实的标签比较，计算模型的准确率，即标签一致的占比：

$$A_{ACC}=\frac{N_C}{N_P}\times 100\% \tag{10}$$

其中：N_C是关系预测正确的样本数量；N_P是预测的样本数量。

3.3 基线模型

对于基线模型，首先选择原数据集论文中提供的3个基准模型：BERT（选择用RoBERTa[19]代替），LSTM[20]和CNN[21]，在后两者的实现细节上，选择用RoBERTa作词嵌入，选择BiLSTM、CNN做编码层，然后通过分类器进行训练。除了官方提供的3个基准模型以外，还选择了以下模型作为基线模型：

1）ESIM。采用BiLSTM和注意力机制来获得文本更好的交互信息，并且采用多种交互形式将特征进行融合。

2）Roformer[22]。应用了旋转式位置编码的思想，这是一种配合注意力机制达到“使用绝对位置编码的方式实现绝对位置编码”的设计，能够使RoBERTa等模型可以编码的序列长度增大，对长文本具有很好的处理效果。

3）ERNIE-DOC[23]。采用回溯式Feed机制和增强的循环机制，使模型获取更长的有效上下文长度，以获得整个文档的相关信息。

4）Backward。为2019年“中国法研杯”相似案例匹配任务第二名的模型。采用Cross-Encoder的方式对文本进行编码，即文书A、B和A、C分别用同一个BERT模型进行编码，然后将编码结果相减，用分类器进行训练。

5）AlphaCourt。为2019年“中国法研杯”相似案例匹配任务第一名的模型。选择Triplet Loss损失函数进行训练，并且采用模型融合的方法提升训练结果。

6）LFESM[24]。在ESIM的基础上进行改动，用BERT编码层代替GloVe[25]+BiLSTM进行编码，同时从案例文本中提取一部分法律特征，以one-hot的方式进行编码，与经过BERT编码的词向量融合进行训练。该模型是目前CAIL2019-SCM数据集上的SOTA模型。

3.4 环境与参数设置

实验在PyTorch环境中进行，在Ubuntu系统上采用GPU（RTX3090）进行模型训练。优化器为AdamW。为了减少过拟合现象的发生，采用权重衰减策略，权重衰减系数为0.01。学习率为1×10^{-5}，同时采用学习率衰减策略，模型每处理200个batch，学习率减少20%。Dropout设置为BERT模型的默认参数0.1。考虑到显存的问题，batchsize设置为3。超参数τ设置为0.08。考虑到文书中重要的信息都位于后半部分，选择从后截断文本，规定输入序列与抽取特征的和经过分词后的字符总数必须小于512。在词嵌入层使用的预训练模型为包含12个Transformer的Chinese-RoBERTa-wwm-ext（base），RoBERTa[19]（base）的隐藏层大小为768。

3.5 结果分析

表3显示了所提模型和基线模型的准确率比较。由表3可以看出，所提模型在验证集和测试集上都取得了当前最好的效果。在原数据集论文中提供的3个基准模型中，表现最好的是CNN模型，原因在于CNN模型能够充分考虑到文本局部的特征，挖掘到更细粒度的信息，从而提升匹配效果。ESIM模型在验证集上的效果相对于原数据集论文中提供的基线模型有所提升，但是在测试集上表现一般，原因在于虽然该模型充分考虑了文本之间的交互，但是忽略了文书中细粒度特征，而法律文书之间的差异通常体现在一些细微的法律要素上。Roformer模型通过对编码位置的更改，能够将BERT模型输入的序列长度增加至1 024，充分提取文本内的信息，相较于BERT模型的效果有所提升，但忽略了文本之间的交互。ERNIE-DOC同样通过循环机制使模型具有更长的有效上下文长度，以获得整个文本的相关信息，但是并没有关注不同文本之间的交互匹配信息。Backward采用BERT模型传统的文本匹配方式计算文本之间的相似度，虽然能够充分考虑文本之间的交互，结果相较于之前未进行交互的模型提升明显，但是也会进一步限制文本输入BERT的长度，因此会损失掉一些文本信息。AlphaCourt选择Rank模型解决本文中相对相似的问题，损失函数采用Triplet Loss，同时融入法律特征等辅助模型进行相似性比较，但是该模型不能充分利用法律特征的语义信息，仅利用了词频、词数等信息。LFESM相对于ESIM模型进行了进一步改进，利用BERT模型进行编码能够充分挖掘文本的上下文信息，但是利用one-hot编码对法律特征要素进行处理，同样会损失特征的语义信息。相较于基线模型表现最好的LFESM模型，所提模型在验证集和测试集上的结果分别提升了0.46和1.05个百分点。

表3 CAIL2019-SCM数据集上的实验结果

Table 3 Experimental results on the CAIL2019-SCM dataset %

模型	验证集准确率	测试集准确率
BERT	61.73	67.25
LSTM	62.00	68.00
CNN	62.27	69.53
ESIM	63.33	67.84
Roformer	66.07	69.79
ERNIE-DOC	65.60	68.80
Backward	67.73	71.81
AlphaCourt	70.07	72.66
LFESM	70.01	74.15
所提模型	70.47	75.20

与所有基线模型相比，本文提出的基于Sentence-BERT的相似案例匹配模型充分考虑文书的序列信息，融入法律特征要素，同时采用对比学习的方法对BERT输出的词向量进行进一步优化，最终通过分类任务完成整个匹配过程，使模型具备较好的泛化能力，在CAIL2019-SCM数据集上取得了较好的效果。

3.6 消融实验

为了更好地证明模型中各部分的作用，进行了消融实验研究，实验结果如表4所示。由表4可以看出：将人工提取的法律特征要素去掉（w/o法律要素）后，只将法律文本送入编码层，然后通过分类器完成匹配，验证集和测试集的准确率比融入法律特征要素的模型分别下降了1.74和2.00个百分点，说明了法律特征要素融入模型中进行训练的有效性；直接取RoBERTa最后一层Transformer的隐藏层输出，而不取中间第1、5、9、12层（w/o预训练中间层），验证集和测试集的准确率分别下降了3.54和2.22个百分点，说明了BERT的中间层能够学习到一些有用的语义信息，帮助模型更好地区分文本的细粒度；去掉对比学习的辅助任务（w/o对比学习）后，验证集和训练集的效果分别下降了0.27和0.90个百分点，说明了加入对比学习后，对于模型的泛化能力有了显著提升。

表4 CAIL2019-SCM数据集上的消融实验结果

Table 4 Ablation experiment results on the CAIL2019-SCM dataset %

模型	验证集准确率	测试集准确率
所提模型	70.47	75.20
w/o法律要素	68.73	73.20
w/o预训练中间层	66.93	72.98
w/o对比学习	70.20	74.30

3.7 超参数分析

为了探索对比学习权重α取值对于实验结果的影响，设置了一组实验来进行验证，表5展示了在联合训练阶段取不同权重的实验结果。由表5可以看出，在α=0.50时模型准确率最高，并且在α>0.50之后准确率有较大下降，说明对比学习框架不能直接加入所提模型，需要找到一个合适的权重值。

表5 不同学习权重下的准确率

Table 5 Accuracy under different weights

α	准确率/%
0.00	74.2
0.25	74.6
0.50	75.2
0.75	74.0
1.00	71.2

4 案例分析

所提模型的目的之一是希望通过关注案件文本中的关键要素来找出不同案件文本之间的细微差异。为了进一步验证融合案件要素的有效性，随机从测试集中选取案例进行说明，具体如下：

案例A 被告姚某某，**福建某有限公司法定代表人**，被告因**生产经营需要**，向原告借款700 000元，原告陆续将借款以**现金**和**银行转款**的方式足额交付给被告后，于2016年2月1日，被告**法定代表人**向原告出具**收款收据**，并加盖被告公司财务专用章，内容记载借到原告700 000元，**年息12%**计算，未约定借款期限。借款后，被告**未支付利息**，经原告催要未果，引起纠纷。

案例B 被告杨某某，**秦岭某机械制造有限公司法定代表人**，以**生产经营缺少资金**为由分6次向原告处借款63万元，约定**年利率为12%**，原告以**转账或现金**的方式将63万元借款交付给了被告，被告向原告出具了6份**收款收据**，从2016年6月份利息结算至2017年12月底，被告又拖欠原告13.43万元利息及63万元**借款本金未付**。原告催要未果，现诉至本院。

案例C 被告孙某某以**个人生活需要**为由，于2016年1月7日分2次向原告借款120 000元，2017年1月7日被告为原告重新出具了**收款收据**，总计利息为37 800元，原/被告未约定借款期限。原告向被告提供借款后，经原告多次催要，被告**未偿还原告借款本金及利息**。原/被告均认可2017年1月7日之后借款利息调整为**月息2%**。上述事实有收款记录、庭审笔录为证。

在案例内包含A、B、C 3个文本，其中黑体部分是案例中出现的要素。从案例中可以直观地看出，虽然3个案例同时均有收款收据作为说明，但A和B具有共同的法律要素，例如被告人均为某公司的法定代表人，借钱用途均为生产经营需要，年利率均为12%。C中被告是以个人生活为由借钱，身份为自然人。同时，借款利率与前2个案例不同，考虑到A和B的相同法律要素较多，测试集中将A和B标为相似案例，证明了本文中考虑要素匹配的必要性。

5 结束语

本文针对由民间借贷案件构成的相似案例匹配任务，建立一种基于法律关键要素的相似案例匹配模型。相似案例匹配存在篇幅较长、结构相似、语义差距小等挑战，所提模型通过融合特定的法律要素，并且与案情文本相结合，使模型学习到充分的语义信息。同时，参考孪生网络的设计，建立适配三元组的相似案例匹配模型。最后，通过增加对比学习损失，增强了模型识别不同文本的能力，进而提高模型的泛化性。实验结果表明，所提模型在CAIL2019-

SCM[2]数据集上相比于对照模型具有明显优势。后续将更加注重法律先验知识的应用,同时进一步提高模型检索效率,以便能够更好地应用于类案推荐任务。

参考文献

[1] 王景林,吴宜霖. 类案检索制度在司法实践中的应用研究[J]. 法制博览,2022(2):100-102.
WANG J L, WU Y L. Research on the application of similar case retrieval system in judicial practice [J]. Legality Vision, 2022(2): 100-102. (in Chinese)

[2] XIAO C J, ZHONG H X, GUO Z P, et al. CAIL2019-SCM: a dataset of similar case matching in legal domain[EB/OL]. [2022-09-16]. https://arxiv. org/abs/1911. 08962. pdf.

[3] CONNEAU A, KIELA D, SCHWENK H, et al. Supervised learning of universal sentence representations from natural language inference data [C]//Proceedings of 2017 Conference on Empirical Methods in Natural Language Processing. Philadelphia, USA: Association for Computational Linguistics, 2017: 670-680.

[4] REIMERS N, GUREVYCH I. Sentence-BERT: sentence embeddings using Siamese BERT-networks [EB/OL]. [2022-09-16]. https://arxiv. org/abs/1908. 10084. pdf.

[5] DAS D, SMITH N A. Paraphrase identification as probabilistic quasi-synchronous recognition [C]//Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. Philadelphia, USA: Association for Computational Linguistics, 2009: 468-476.

[6] 卜质琼,郑波尽. 基于LDA模型的Ad hoc信息检索方法研究[J]. 计算机应用研究,2015,32(5):1369-1372.
BU Z Q, ZHENG B J. Ad hoc information retrieval method based on LDA[J]. Application Research of Computers, 2015, 32(5): 1369-1372. (in Chinese)

[7] 吕正东,李航. 深度匹配学习在语言匹配中的应用[J]. 中国计算机学会通讯,2015,8(8):30-38.
LÜ Z D, LI H. Apply deep matching learning in language matching[J]. Communication of China Computer Federation, 2015, 8(8): 30-38. (in Chinese)

[8] HUANG P S, HE X D, GAO J F, et al. Learning deep structured semantic models for web search using click through data[C]//Proceedings of the 22nd ACM International Conference on Information & Knowledge Management. New York, USA: ACM Press, 2013: 2333-2338.

[9] SHEN Y L, HE X D, GAO J F, et al. A latent semantic model with convolutional-pooling structure for information retrieval[C]//Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. New York, USA: ACM Press, 2014: 101-110.

[10] PANG L A, LAN Y Y, GUO J F, et al. Text matching as image recognition[C]//Proceedings of AAAI Conference on Artificial Intelligence. Palo Alto, USA: AAAI Press, 2016: 2793-2799.

[11] CHEN Q, ZHU X D, LING Z H, et al. Enhanced LSTM for natural language inference[EB/OL]. [2022-09-16]. https://arxiv. org/abs/1609. 06038. pdf.

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

[13] 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. New York, USA: ACM Press, 2017: 6000-6010.

[14] LI B H, ZHOU H, HE J X, et al. On the sentence embeddings from pre-trained language models[EB/OL]. [2022-09-16]. https://arxiv. org/abs/2011. 05864. pdf.

[15] SU J L, CAO J R, LIU W J, et al. Whitening sentence representations for better semantics and faster retrieval[EB/OL]. [2022-09-16]. https://arxiv. org/abs/2103. 15316. pdf.

[16] PEINELT N, NGUYEN D, LIAKATA M. tBERT: topic models and BERT joining forces for semantic similarity detection[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, USA: Association for Computational Linguistics, 2020: 7047-7055.

[17] GAO T Y, YAO X C, CHEN D Q. SimCSE: simple contrastive learning of sentence embeddings[EB/OL]. [2022-09-16]. https://arxiv. org/abs/2104. 08821. pdf.

[18] JAWAHAR G, SAGOT B, SEDDAH D. What does BERT learn about the structure of language?[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, USA: Association for Computational Linguistics, 2019: 3651-3657.

[19] LIU Y H, OTT M, GOYAL N, et al. RoBERTa: a robustly optimized BERT pretraining approach[EB/OL]. [2022-09-16]. https://arxiv. org/abs/1907. 11692. pdf.

[20] PALANGI H, DENG L, SHEN Y L, et al. Deep sentence embedding using long short-term memory networks: analysis and application to information retrieval[J]. ACM Transactions on Audio, Speech, and Language Processing, 2016, 24(4): 694-707.

[21] HU B T, LU Z D, LI H, et al. Convolutional neural network architectures for matching natural language sentences[C]//Proceedings of the 27th International Conference on Neural Information Processing Systems. New York, USA: ACM Press, 2014: 2042-2050.

[22] SU J L, LU Y, PAN S F, et al. RoFormer: enhanced Transformer with rotary position embedding[EB/OL]. [2022-09-16]. https://arxiv. org/abs/2104. 09864. pdf.

[23] DING S Y, SHANG J Y, WANG S H, et al. ERNIE-doc: a retrospective long-document modeling Transformer[EB/OL]. [2022-09-16]. https://arxiv. org/abs/2012. 15688. pdf.

[24] HONG Z L, ZHOU Q F, ZHANG R, et al. Legal feature enhanced semantic matching network for similar case matching [C]//Proceedings of 2020 International Joint Conference on Neural Networks. Washington D. C. , USA: IEEE Press, 2020: 1-8.

[25] PENNINGTON J, SOCHER R, MANNING C. GloVe: global vectors for word representation[C]//Proceedings of 2014 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, USA: Association for Computational Linguistics, 2014: 1532-1543.

编辑 陆燕菲