文章编号：1003－0077(2022)11－0140－08

融合案件要素的相似案例匹配

刘 权[1,2]，余正涛[1,2]，高盛祥[1,2]，何世柱[3]，刘 康[3]

(1. 昆明理工大学 信息工程与自动化学院，云南 昆明 650500；
2. 昆明理工大学 云南省人工智能重点实验室，云南 昆明 650500；
3. 中国科学院 自动化研究所，北京 100190)

摘 要：相似案例匹配是智慧司法中的重要任务，其通过对比两篇案例的语义内容判别二者的相似程度，能够应用于类案检索、类案类判等。相对于普通文本，法律文书不仅篇幅更长，文本之间的区别也更微妙，传统深度匹配模型难以取得理想效果。为了解决上述问题，该文根据文书描写规律截取文书文本，并提出一种融合案件要素的方法来提高相似案件的匹配性能。具体来说，该文以民间借贷案件为应用场景，首先基于法律知识制定了 6 种民间借贷案件要素，利用正则表达式从法律文书中抽取案件要素，并形成词独热形式的案件要素表征；然后，对法律文本倒序截取，并通过 BERT 编码得到法律文本表征，解决法律文本的长距离依赖问题；接着使用线性网络融合法律文本表征与案件要素表征，并使用 BiLSTM 对融合的表征进行高维度化表示；最后通过孪生网络框架构建向量表征相似性矩阵，通过语义交互与向量池化进行最终的相似度判断。实验结果表明，该文模型能有效处理长文本并建模法律文本的细微差异，在 CAIL2019-SCM 公共数据集上优于基线模型。

关键词：相似案例匹配；案件要素；预训练语言模型

中图分类号：TP391 **文献标识码**：A

Incorporating Case Elements for Case Matching

LIU Quan[1,2], YU Zhengtao[1,2], GAO Shengxiang[1,2], HE Shizhu[3], LIU Kang[3]

(1. Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, Yunnan 650500, China;
2. Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology, Kunming, Yunnan 650500, China;
3. Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: Simiar Case matching is an important task in intelligent justice, especially for case retrieval and same-case same-judgment. Owing to the long text and the subtle difference between legal documents, existing deep matching models are difficult to achieve ideal results. To address this issue, this paper proposes a method of integrating case elements to improve the matching of similar cases with a focus on the private lending cases. First, six types of private lending case elements are formulated and extracted by regular expressions, represented in the form of one-hot word vectors. Then the legal text is filtered and formed in reverse order, represented by BERT capture the long-distance dependence. The legal text representation and the case element representation is fused by the linear network and then encoded by BiLSTM for high-dimensional representation. Finally, the vector representation similarity matrix is constructed through the twin network framework, and the final similarity is decided by semantic interaction and vector pooling. The experimental results show that the proposed model is better than the baseline model on the CAIL2019-SCM public data set.

Keywords: case matching; case elements; pre-training language model

收稿日期：2020-12-31 **定稿日期**：2021-03-10

基金项目：国家重点研发计划(2018YFC0830101, 2018YFC0830105, 2018YFC0830100)；国家自然科学基金(61972186, 61761026, 61762056)；云南省重大科技专项计划项目(202002AD080001-5)；云南省基础研究计划(202001AS070014, 2018FB104)；云南省高新技术产业专项(201606)；云南省人培项目(KKSY201703005)

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

0　引言

司法已经步入大数据的时代，网络快速传播和信息实时流转使得群体作用力越来越重要，类案类判作为以人工智能技术为基础，通过自然语言处理做出关联案件的判断的一种手段，凝聚着众多审判人员对于司法的理解，其优势在于能够规避法官主观意识下带来的不确定和偏差。类案类判是实现公正司法的重要途径，当前的大量系统无法应用于实践主要是因为相似案例匹配不够精准，因此提高相似案例匹配的精准度是一个亟待解决的问题。

相似案例匹配主要是对比两篇法律文书内容的文本相似性，从候选裁判文书中计算出最相似的，该任务本质上是文本匹配任务。文本匹配任务的实质是将两个文本的向量表征进行计算，现如今比较具有代表性的深度匹配模型种类主要分为匹配矩阵交互模型与预训练语言匹配模型两类：

（1）首先定义的匹配矩阵是基于两个句子中词和词之间的匹配程度，模型利用两个词的词向量之间的余弦相似度或者点积来定义词之间的相似度，然后句子之间两两词之间都会计算相似度，根据词在句子中的空间位置构建出一个二维的匹配矩阵，然后通过分别进行注意力的操作，将两个向量表征与匹配矩阵进行语义交互，例如，ESIM[1]（Enhanced Sequential Inference Model）模型构建了相似矩阵后，利用设计的对齐交互过程将向量表征与矩阵进行注意力计算，学习语义之间的差异性。

（2）由于预训练语言模型在多项 NLP 任务中表现优越，所以基于预训练语言模型的文本匹配也成为当前典型方法。例如，BERT[2] 通过将两个句子用分隔符[SEP]进行衔接输入到模型中，然后利用 Softmax 层进行匹配计算。

通用的文本匹配模型在相似案例匹配任务中同样适用，因为相似案例匹配任务与文本匹配任务都注重文本的相似性。文本匹配任务中两个句子是可以相关的，但相似案例匹配任务中的案例相似则不一样。案例涉及的当事人与案件事实可能不同，总体来说只要基本案情与法律适用一致，即属于同类案件。一是以判断两个案件描述的法律事实认定是否一致为基础，但也不是要求两个法律事实必须在所有细节情形上严丝合缝。法官对案件审理的最终目的是确定是否支持当事人的诉讼请求，因此要求法律文书中描述案情的法律要素一致。

因此与普通文本匹配任务相比，相似案例匹配主要存在以下问题：案例文档之间的区别可能很细微，很难判断两个文档是否相似。因此，我们需要根据案件的描述抽取一系列的案件要素。通常，相似或相关的原因可以共享同一组特征属性。表 1 为文书样例，对于认定的民间借贷案例的法律要素进行了着重标注，民间借贷案件要素具体包括：①原告和被告；②担保类型，抵押、担保等；③年利率；④借款方式包含现金、银行转账、电子转账等；⑤借款证据，包括借条、收款收据、合同等；⑥是否已还款。根据“法研杯”的数据集中的描述，对于民间借贷案例，官方注释了 9 种不同的要素，我们从中筛选了 6 种较为重要的要素，方便我们在后续工作中进行要素的识别与融合。

表 1　文书抽取要素样例

类型	内容
原告：王强，男，1981 年 10 月 25 日出生，汉族，住会昌县。委托诉讼代理人：胡育强，男，住会昌县。被告：王德昌，男，1976 年 11 月 2 日出生，汉族，住会昌县。被告：王晓晖，男，1972 年 9 月 27 日出生，汉族，住会昌县。 ————上部分为描述当事人信息，下部分为案情描述———— 原告王强向本院提出诉讼请求：1.判令两被告一次性归还原告借款 30000 元及利息（按月息 2%计算）；2.本案诉讼费由被告承担。事实和理由：2015 年 6 月 4 日，被告王德昌以需要资金周转为由，向原告王强借款 30000 元。被告王德昌写下《借条》和《借款合同》；被告王晓晖作为担保人，也和原告签订了《担保合同》……被告王晓晖对上述借款提供连带责任保证担保，并与原告王强签订了《保证担保合同》，合同约定保证期限自合同生效之日起至主合同全部履行完毕止。借款期限届满后，被告王德昌向原告王强支付了 2015 年 10 月之前的利息，此后的利息及本金经原告多次催取未归还。	原告王强； 借款 30000 元； 月息 2%； 借条，借款合同； 担保合同； 未归还

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

我们根据文书描述规律，发现文书的前半部分是描述案件当事人信息，后半部为案情描述。可以发现所需要的要素大部分都是出现在后半部分，为了减少文书文本的长度，从文书的后半部分开始截取，也能最大化地保留文书信息。例如，在民间借贷案件中，以下这些要素会影响最终判决：金额、利率（包含年利率、月利率等）、担保方式等。如利率存在三种情况：①在签订借贷合同时没有承诺利息，②法官不支持利息的主张，③当利率高于24%时，超出部分得不到法院的支持。最后，当利率超过36%时，借款人可以要求返还超额部分。当两种情况下的利率不同时，情况往往不同。法律文书通常以结构化格式编写，通过案件要素能更容易抓住文档的关键点。

本文根据文书描写规律，从文章结尾开始倒序进行文本截取，利用正则表达式在法律文书文本中抽取案件要素，并将其转化为向量表征。然后使用BERT作为文本编码器对法律文本进行编码，用线性网络作为融合层将法律文本编码表征与案件要素表征进行融合，并用BiLSTM网络进行高维度化表示，然后通过孪生网络框架构建向量表征相似性矩阵，并进行语义交互与最后的向量池化输出。

总体来说，本文的主要贡献如下：

（1）提出了融合案件要素与截取法律文书文本的方法，一方面能够有效建模长文本，另一方面能够有效处理细微差异。

（2）在模型实验方面，采用正确的规律截断文书文本，减少了文书的文本长度，并定义了民间借贷案件文书的6个要素，并通过正则表达式进行抽取后，提出了匹配模型将要素与法律文书融合。在民间借贷案件公开数据集上实验本文方法，结果表明，相比于基线模型BERT，在测试集上提升了5.13%的准确率。

1 相关工作

1.1 文本匹配

近几年的深度匹配模型在文本匹配任务上的使用逐渐广泛。文本匹配是许多NLP任务的关键技术，如自然语言推理[4]、复述识别[5]和机器阅读理解[6]。对此人们通常使用几种模式来进行研究。其中一种模式是使用了句子编码结构，将两个句子编码成向量表示，然后将向量组合做最终预测[7-9]。然而，在编码过程中没有直接考虑两个输入序列之间的相互作用，使得模型难以捕捉复杂的关系，后续的工作用匹配聚合的方法对两句的对齐进行建模。Wang[10]使用matchLSTM对两个序列进行词级匹配；Parikh[11]提出了一种简单的注意操作，并使用前驱网络来整合对齐表示；BiMPM[5]使用多视角匹配操作对两个序列进行比较，并使用BiLSTM网络进行聚合。Gong等人[12]使用DensNet作为特征提取器，从交互张量中提取语义特征。为了更好地捕捉不同层次的句子对齐，可以将多个注意操作叠加在一起。Yang等人[13]提出了一个简单但有效的框架，具有更丰富的对齐特征。Tay等人[14]利用多级注意力机制进行更广泛的匹配，提高实验结果。ADIN[15]为多步骤推理过程堆栈异步推理层。

近几年，预训练语言模型已经在文本匹配任务上取得了最先进的结果。然而，较大的模型参数和缓慢的推断速度使得直接将这些结构部署到实际应用程序中变得困难。与上述方法不同的是，我们在预训练语言模型BERT的基础上加入了交互式匹配模型的特点，构造了匹配矩阵并进行语义交互，利用原始的词特征和上下文特征来丰富中间表示，从而指导模型产生更好的交互信息。

1.2 相似案例匹配

近年来，许多学者都在关注自然语言处理技术在法律领域的应用。2019年“法研杯”发布了第一个相似案例匹配任务的数据集[15]。除了相似案例匹配任务外，该方面还有多个任务，包括法律判决预测[16]、法律问答[17]以及从法律文书中抽取信息[18]。以上工作都试图将现有的相关模式和方法整合到法律领域，并使之适应法律文本的特点，从而更好地帮助法律专业人员。因此本文结合现有的法律工作思想，提出了一种将法律案件要素与文本匹配模型相结合的模型。

2 方法

2.1 任务定义

相似案例匹配任务主要是对案例文档进行相似度计算。将案例文书设为(a,b,c)三元组，其中a为源案例，b为与a相似的案例，c则反之。将三个

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

文书表征为 $a=\{a_1,a_2,\cdots,a_{l_a}\}$，$b=\{b_1,b_2,\cdots,b_{l_b}\}$，$c=\{c_1,c_2,\cdots,c_{l_c}\}$，目标是预测标签 y 是否属于$\{0,1\}$，当 $y=0$ 时，则 $\mathrm{sim}(a,b)>\mathrm{sim}(a,c)$，代表 a 与 b 的相似度大于 a 与 c 的相似度；当 $y=1$ 时，则 $\mathrm{sim}(a,b)<\mathrm{sim}(a,c)$，则反之。本文为方便模型输入，并减少模型的参数量，对数据进行预处理，将三元组形式的数据改写为文书对，将 a 与 b 构成正例的法律文书对$\{a,b\}$，标签为 1；a 与 c 构成负例的法律文书对$\{a,c\}$，标签为 0。

2.2　模型

图 1 为本文模型图。首先将法律文本通过 BERT 进行文本编码，通过融合层将抽取的案件要素与文本编码表征进行融合，然后利用匹配矩阵进行语义交互，最后通过池化与 Softmax 层进行输出。

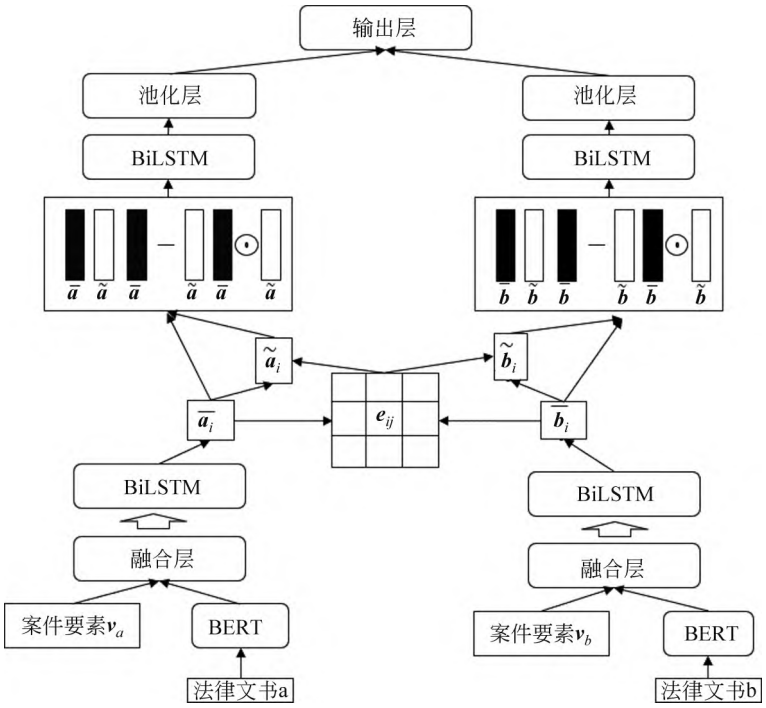

图 1　融合案件要素模型

2.2.1　案件要素抽取

首先对案件文书中的案件要素进行抽取，在制定好要素内容后，我们通过正则表达式来进行抽取。在抽取完这些要素后，我们需要对它们进行特征向量化。对于每个定义的要素，将其用词独热向量形式表征，如果用正则抽取出了内容，则表征为 1，否则表征为 0。如表 1 中的样例，因此总共设定了 6 种要素，已抽出了被告、年利率、担保类型、借款凭据等 5 个要素，没抽取到是否已还款，所以此文书的法律要素向量为[1,1,1,0,1,1]。法律文书 a 产生的要素表征为 v_a，法律文书 b 的要素表征为 v_b。

2.2.2　编码融合层

本文模型使用 BERT 作为编码器，对平行的法律文书 a 和 b 进行编码，得到编码后的法律文书表征 $\boldsymbol{a}_i$ 和 $\boldsymbol{b}_j$，如式(1)、式(2)所示。

$$\boldsymbol{a}_i=\mathrm{BERT}(a,i)\quad \forall i\in[1,\cdots,l_a] \tag{1}$$

$$\boldsymbol{b}_j=\mathrm{BERT}(b,j)\quad \forall i\in[1,\cdots,l_b] \tag{2}$$

然后将法律要素表征与 BERT 编码后的法律文书表征通过融合层进行信息融合，融合层是一个全连接层，通过融合层后，能得到两个法律文书的融合表征 $\bar{v}_a$ 和 $\bar{v}_b$，如式(3)、式(4)所示。

$$\bar{v}_a=\boldsymbol{W}\boldsymbol{v}_a+\boldsymbol{a}_i \tag{3}$$

$$\bar{v}_b=\boldsymbol{W}\boldsymbol{v}_b+\boldsymbol{b}_j \tag{4}$$

其中，$\boldsymbol{W}$ 是一个向量矩阵，表示为 $\boldsymbol{W}\in\mathbb{R}^{d\times h}$，$d$ 为 $\bar{v}_a$ 与 $\bar{v}_b$ 的维度，h 为隐状态维度。我们会使用 $BiLSTM$ 对融合的向量表征进行更高维度的特征表示化。操作如式(5)、式(6)所示。

$$\bar{\boldsymbol{a}}_i=\mathrm{BiLSTM}(\bar{v}_a,i) \tag{5}$$

$$\bar{\boldsymbol{b}}_j=\mathrm{BiLSTM}(\bar{v}_b,j) \tag{6}$$

其中，$\bar{\boldsymbol{a}}_i$ 表示法律文书 a 的高度维度化表征，$\bar{\boldsymbol{b}}_j$ 表示法律文书 b 的高度维度化表征。融合表征通过双向 LSTM 编码如式(7)～式(12)所示。

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

$$\boldsymbol{i}_t=\sigma(\boldsymbol{w}_{ixt}*\mathrm{wd}+\boldsymbol{U}_{iht-1}) \tag{7}$$

$$\boldsymbol{f}_t=\sigma(\boldsymbol{W}_{fxt}*\mathrm{wd}+\boldsymbol{U}_{fht-1}) \tag{8}$$

$$\boldsymbol{o}=\sigma(\boldsymbol{W}_{oxt}*\mathrm{wd}+\boldsymbol{U}_{oht-1}) \tag{9}$$

$$\tilde{\boldsymbol{c}}_t=\tanh(\boldsymbol{W}_{cxt}*\mathrm{wd}+\boldsymbol{U}_{cht-1}) \tag{10}$$

$$\boldsymbol{c}_t=\boldsymbol{i}\circ\tilde{\boldsymbol{c}}_t+\boldsymbol{f}_t\circ\boldsymbol{c}_{t-1} \tag{11}$$

$$\boldsymbol{h}_t=\boldsymbol{o}_t\circ\tanh(c_t) \tag{12}$$

其中，i_t 是输入门，决定 LSTM 存储什么信息；f_t 是遗忘门，决定 LSTM 丢弃什么信息；o 是输出门，决定这个 LSTM 单元要输出什么信息；“∘”代表点乘操作。通过不同的门控机制，LSTM 将输入的词向量 wd 转化为与上文相关的语义向量 $\boldsymbol{h}_t$。

2.2.3 语义交互层

对通过后得到的文本表征进行局部推理，在进行局部推理之前，首先要将两个融合文本表征进行对齐操作，即计算两个向量的相似度。操作如式(13)所示。

$$\boldsymbol{e}_{ij}=\bar{\boldsymbol{a}}_i^{\mathrm{T}}\bar{\boldsymbol{b}}_j \tag{13}$$

将文书 a 的融合表征向量 $\bar{a}_i$ 进行转置操作，与文书 b 的融合表征向量 $\bar{b}_j$ 进行相乘操作得到相似度矩阵 $\boldsymbol{e}_{ij}$。然后将其进行推理操作，将两个不同的法律文书表征 $\bar{a}_i$、$\bar{b}_j$ 分别与相似度矩阵 $\boldsymbol{e}_{ij}$ 进行局部推理操作，将得到的相似度矩阵，结合 $\bar{a}_i$，$\bar{b}_j$ 两个表征，互相生成彼此相似性加权后的表征 $\tilde{a}_i$ 和 $\tilde{b}_j$。具体如式(14)、式(15)所示。

$$\tilde{\boldsymbol{a}}_i=\sum_{j=1}^{l_b}\frac{\exp(\boldsymbol{e}_{ij})}{\sum_{k=1}^{l_b}\exp(\boldsymbol{e}_{ik})}\bar{\boldsymbol{b}}_j,\quad\forall i\in[1,\cdots,l_a] \tag{14}$$

$$\tilde{\boldsymbol{b}}_j=\sum_{j=1}^{l_a}\frac{\exp(\boldsymbol{e}_{ij})}{\sum_{k=1}^{l_a}\exp(\boldsymbol{e}_{kj})}\bar{\boldsymbol{a}}_i,\quad\forall i\in[1,\cdots,l_b] \tag{15}$$

对相似性加权后的表征 $\tilde{a}_i$ 和 $\tilde{b}_j$ 再进行一次信息增强的操作，具体如式(16)、式(17)所示。

$$\boldsymbol{m}_a=[\bar{\boldsymbol{a}};\tilde{\boldsymbol{a}};\bar{\boldsymbol{a}}-\tilde{\boldsymbol{a}};\bar{\boldsymbol{a}}\odot\tilde{\boldsymbol{a}}] \tag{16}$$

$$\boldsymbol{m}_b=[\bar{\boldsymbol{b}};\bar{\boldsymbol{b}};\bar{\boldsymbol{b}}-\bar{\boldsymbol{b}};\bar{\boldsymbol{b}}\odot\bar{\boldsymbol{b}}] \tag{17}$$

两个句子表征进行差和点积(每个元素单独相乘)操作，获取例如“对立”的推断关系。其中，“;”表示向量拼接，“⊙”表示向量点积。

再次使用 BiLSTM 表示 $\boldsymbol{m}_a$ 和 $\boldsymbol{m}_b$，公式与将词嵌入向量进行编码相同，但目标变成了获取局部推理得到的两个向量的上下文信息。具体如式(18)、式(19)所示。

$$\boldsymbol{v}_{a,t}=\mathrm{BiLSTM}(F(\boldsymbol{m}_{a,t})), \tag{18}$$

$$\boldsymbol{v}_{b,t}=\mathrm{BiLSTM}(F(\boldsymbol{m}_{b,t})), \tag{19}$$

为了控制模型复杂度，使用了 1 层网络和 Relu 激活函数处理 $\boldsymbol{m}_a$ 和 $\boldsymbol{m}_b$，经 BiLSTM 之后得到的句子矩阵表示分别是 $\boldsymbol{v}_{a,t}$ 和 $\boldsymbol{v}_{b,t}$。因为 $\boldsymbol{m}_a$、$\boldsymbol{m}_b$ 会显著增加参数的数量，所以这里采用 F 激活函数进行映射，F 是 ReLU 的前馈神经网络。推理模型将上述得到的结果向量池化成定长向量，提供给最终的分类器以确定整个推理关系。

因为求和会对序列长度敏感，所以用平均和最大池化合并向量形成最后固定长向量 $\boldsymbol{V}$。如式(20)～式(22)所示。

$$\boldsymbol{v}_{a,\mathrm{ave}}=\sum_{i=1}^{l_a}\frac{\boldsymbol{v}_{a,i}}{l_a},\quad\boldsymbol{v}_{a,\max}=\max_{i=1}^{l_a}\boldsymbol{v}_{a,i}, \tag{20}$$

$$\boldsymbol{v}_{b,\mathrm{ave}}=\sum_{j=1}^{l_b}\frac{\boldsymbol{v}_{b,}}{l_b},\quad\boldsymbol{v}_{b,\max}=\max_{j=1}^{l_b}\boldsymbol{v}_{b,j} \tag{21}$$

$$\boldsymbol{V}=[\boldsymbol{v}_{a,\mathrm{ave}};\boldsymbol{v}_{a,\max};\boldsymbol{v}_{b,\mathrm{ave}};\boldsymbol{v}_{b,\max}] \tag{22}$$

最后的向量 $\boldsymbol{V}$ 进入多层感知机(tanh 激活函数＋Softmax 层)，计算最后的语义相似度。

3 实验

3.1 数据集

本文采用了公共数据集，由“法研杯”2019 年任务之相似案例匹配发布，数据集包含了 8 138 个法律案件文书的三元组，所有文书均来自于裁判文书网，主题均为民间借贷案件。该数据集由法律专业人士标注，包括了当事人描述与法律事实描述两部分。该数据集具体统计如表 2 所示。

表 2 数据集统计

数据集规模	数量
训练集	5 102
测试集	1 500
验证集	1 536

3.2 基线模型

ESIM[1]：该模型充分利用唯一的对齐过程，采用丰富的外部语法特性作为对齐层的额外输入。

BERT[2]：预训练语言模型，本文将 BERT 作为文本编码器，获取法律文本的词嵌入向量。

CNN[19]：将 KIM 提出的文本分类模型修改为文本匹配模型，经过一层卷积后使用最大池化，最后使用 Softmax 层作为输出。

LSTM[20]：长短时记忆(Long short-term memory，

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

LSTM)是一种特殊的 RNN，主要是为了解决长序列训练过程中的梯度消失和梯度爆炸问题。

RE2[12]：包含了原始点对齐特性、先前对齐特性和上下文特性等三个特性的深度匹配框架，可以简化所有剩余组件。其特点是参数少、速度快。

3.3　实验设置

3.3.1　实验设置

BERT 采用的是谷歌公司发布的中文预训练 BERT 模型，其中学习率设置为 0.000 02，Transformer 的层数设置为 12 层。序列最大长度为 64，学习率设置为 0.000 03，训练轮次设置为 100，忍耐度为 4，由于 BERT 对输入序列长度的最大限制为 512，结合数据格式特点，我们从前面截断输入序列。这是因为在民间借贷案例文书描述中，后面的部分比前面的部分提供更多的信息。

ESIM、RE2 的 batch size 设置为 32，使用了法律领域的预训练词向量，维度为 300。

CNN、LSTM、BERT 的实验结果均取自于数据集发表文章[15]中展示的结果。

3.3.2　评价指标

为了验证实验的有效性，我们使用了准确率(Acc)作为评价指标，准确率是指所有的预测正确(正类负类)的占总的比重，如式(23)所示。

$$\mathrm{Acc}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \tag{23}$$

其中，TP 表示将正类预测为正类，FN 表示将正类预测为负类，FP 表示将负类预测为正类，TN 表示将负类预测为负类。

3.4　实验

表 3 展示了 4 种基线模型与本文模型在数据集的验证集和测试集的表现。

表 3　模型性能实验　(单位：%)

模型	验证集	测试集
CNN	62.27	69.53
LSTM	62.00	68.00
RE2	62.08	68.94
BERT	61.93	67.32
ESIM	62.34	69.27
本文模型	65.21	72.45

从表 3 可以看出，本文模型在验证集与测试集上均获得了优异结果。与 BERT 相比，本文模型在验证集上准确率提高了 3.28%，在测试集上准确率提高了 5.13%，因为案件要素是案件内容的抽象化表达，能体现案件文书描述案情的细节，通过将案件要素与案件文本进行关联，能够实现法律文本细微差别的判断；并且为了解决裁判文书长度过长，无法抓取重要信息的问题，在不大于 BERT 限制的序列长度 512 下，我们从文本的结尾开始排序截取，减少了序列的无用信息。BERT 的本质是多个 Transformer 的堆叠，它能通过多头注意力机制和位置编码，很好地解决句子的长距离依赖问题；相比于 ESIM，本文模型在验证集上准确率提高了 2.87%，在测试集上准确率提高了 3.18%。因为本文模型在结合两种模型的基础上再融入了案件要素，对法律文本通过 BERT 进行文本编码，而不是采用 ESIM 使用的词嵌入编码方式，表征法律文本会更加详细；对案件要素与法律文本通过线性层进行融合，增强了法律文本的表征，也扩大了不同文书之间的差异性，然后通过 BiLSTM 网络进行高维度化向量表征；对融合的文书文本表征通过借鉴 ESIM 的构建匹配矩阵后，再进行差异性注意力学习。ESIM 相比于 LSTM 与 CNN，在验证集与测试集表现略优，ESIM 模型中运用到了 BiLSTM 网络编码文本，且利用匹配矩阵对两个法律文书文本进行语义交互，且模型参数量也有了很大的提升，所有表现更加优异；RE2 作为最新的交互文本匹配模型，相比于 ESIM，并未表现得更好，但是模型的参数量小，训练速度更快；尽管 LSTM 与 CNN 在长文本的学习过程中会产生长距离依赖问题，但是 BERT 对于文本的最大长度仅限于 512，因此当对法律文书文本进行截断操作时，会造成一定的信息丢失，因此 BERT 的效果略低于 LSTM 与 CNN。

由于 BERT 允许的最大序列长度为 512，但文书文本长度往往大于 512，因此在输入模型之前通常都需要截断文本。而如何正确地截取文本与验证本文截取文本的有效性，我们进行了文本截取部位的实验，结果如表 4 所示。

表 4　截取实验结果　(单位：%)

	验证集	测试集
从后开始截取	65.21	72.45
从前开始截取	63.12	70.16
去掉首位各 1 句，取文本中段内容	**64.67**	**70.76**

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

我们采用了本文模型作为实验模型，因为案件要素抽取步骤是根据输入文本进行，因此文本长度不一样，抽取的要素有一定的差异。我们发现从后开始截取的方法最为有效，根据表1分析文本可知，文书的前半部分是描述当事人的信息，包括出生日期、户籍等。这类信息对于判断案例是否相似毫无帮助；而从文书中间取序列效果也略差，因为在民间借贷案件中，最后几句话往往交代当事人是否已还款，这对于判断案例匹配是一个重要的要素。

此外，为进一步验证本文模型的有效性，进行了消融实验，实验结果如表5所示。在本文模型的基础上，分别去掉了案件要素的抽取与融合、BiLSTM层对融合文本表征的高维度化、不使用BERT作为文本编码器。

表5 消融实验结果

（单位：%）

	验证集	测试集
w/o 案件要素	63.13	70.21
w/o BiLSTM 高维度化	64.15	71.23
w/o BERT	62.32	69.36
本文模型	**65.21**	**72.45**

注：w/o 代表 without。

从表5结果显示，去掉各个部分对实验结果均有一定的影响。去掉案件要素相比于本文模型，在验证集上下降了2.08%，测试集上下降了2.24%，看出案件要素在判断两个法律文书是否相似的过程中具有一定的作用，不仅提升了模型的参数量，而且能使模型更加关注这些法律要素代表的词汇，而不是其他词语；去掉BiLSTM层网络对输入表示编码相比于本文模型，在验证集上下降了1.06%，测试集上下降了1.22%，融合法律文本表征只是通过线性层简单地进行了向量融合，使用BiLSTM层两次对向量编码，可以增强模型的参数，且使向量蕴含信息更加丰富；不使用BERT文本编码器与本文模型，在验证集上下降了2.89%，测试集上下降了3.09%，当不使用BERT作为文本编码器后，我们使用BiLSTM作为文本编码器，BERT的词嵌入本身可以表达每个词的语义，并且句中词与词之间的上下文信息得以保存，在模型后半部分中获得更复杂的交互时，模型的性能将得到提高。因此本文在保证了法律文本的丰富语义情况下，将案件要素与法律文本表征进行融合，并通过语义多层交互，丰富了法律文本语义表示及对案件要素的挖掘，结果证明了本文模型的合理性、有效性。

4 结论

本文针对相似案例匹配任务，设计了6种案件要素帮助模型更好地区分相似案例，并且根据文书描写规律，正确的截断法律文书文本，减少文本长度。本文在查询相关法律后，制定了6种民间借贷案件要素，利用正则表达式制定规则，从法律文书中进行抽取，形成词独热形式的案件要素表征，采用BERT作为编码层来捕获法律文书中的长距离依赖问题。用线性网络作为融合层将法律文本编码表征与案件要素表征进行融合，并用BiLSTM网络对融合表征进行高维度化表示，然后通过孪生网络框架构建向量表征相似性矩阵，进行语义交互与最后的向量池化输出。本文模型与其他基线模型进行了性能对比实验，并且针对模型的各个模块进行消融实验。实验结果显示了本文方法的有效性与合理性。

未来工作中，旨在挖掘更多的法律文本特征，而不是单薄的案件要素，并寻找比BERT更好的方式来丰富表征法律文书文本。

参考文献

[1] CHEN Q, ZHU X, LING Z, et al. Enhanced LSTM for natural language inference[C]//Proceedings of the Meeting of the Association for Computational Linguistics, 2017: 1657-1668.

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

[3] XIAO C, ZHONG H, GUO Z, et al. Cail2019-SCM: A dataset of similar case matching in legal domain[J]. arXiv preprint arXiv:1911.08962, 2019.

[4] BOWMAN S, ANGELI G, POTTS C, et al. A large annotated corpus for learning natural language inference[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2015: 632-642.

[5] WANG Z, HAMZA W, FLORIAN R. Bilateral multi-perspective Matching for natural language sentences [C]//Proceedings of the 26th International Joint Conference on Artificial Intelligence, 2017: 4144-4150.

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

[6] RAJPURKAR P, ZHANG J, LOPYREV K, et al. SQuAD: 100,000+ questions for machine comprehension of text[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2016: 2383-2392.

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

[8] YIN W, SCHÜTZE H. Convolutional neural network for paraphrase identification[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2015: 901-911.

[9] MUELLER J, THYAGARAJAN A. Siamese recurrent architectures for learning sentence similarity[C]//Proceedings of the 30th AAAI Conference on Artificial Intelligence, 2016.

[10] WANG W, YAN M, WU C. Multi-granularity hierarchical attention fusion networks for reading comprehension and question answering[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018: 1705-1714.

[11] PARIKH A, TÄCKSTRÖM O, DAS D, et al.A decomposable attention model for natural language inference[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2016: 2249-2255.

[12] GONG Y, LUO H, ZHANG J. Natural language inference over interaction space[C]//Proceedings of International Conference on Learning Representations, 2018: 1-15.

[13] YANG R, ZHANG J, GAO X, et al. Simple and effective text matching with richer alignment features[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 4699-4709.

[14] TAY Y, LUU A T, HUI S C. Co-stack residual affinity networks with multi-level attention refinement for matching text sequences[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2018: 4492-4502.

[15] LIANG D, ZHANG F, ZHANG Q, et al. Asynchronous deep interaction network for natural language inference[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2019: 2692-2700.

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

[17] FAWEI B, PAN J Z, KOLLINGBAUM M, et al. A methodology for a criminal law and procedure ontology for legal question answering[C]//Proceedings of the Joint International Semantic Technology Conference. Springer, Cham, 2018: 198-214.

[18] VACEK T, SCHILDER F. A sequence approach to case outcome detection[C]//Proceedings of the 16th Edition of the International Conference on Articial Intelligence and Law, 2017: 209-215.

[19] KIM Y. Convolutional neural networks for sentence classification[C]//Proceedings of EMNLP, 2014: 1746-1751.

[20] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

刘权(1995—)，**硕士研究生，主要研究领域为自然语言处理。**

E-mail: lq826roo@gmail.com

高盛祥(1977—)，**通信作者，副教授，硕士生导师，主要研究领域为自然语言处理、信息检索、机器翻译。**

E-mail: gaoshengxiang.yn@foxmail.com

余正涛(1970—)，**教授，博士生导师，主要研究领域为自然语言处理、信息检索、机器翻译。**

E-mail: ztyu@hotmail.com

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net