

文章编号: 2095-2163(2022)08-0123-06

中图分类号: TN929.5

文献标志码: A

基于深度学习的合同分类及要素抽取模型

张晓芳¹, 欧睿², 饶攀军¹, 郑元¹, 张雷¹, 陈科¹, 周郴莲², 王浩畅², 赵铁军³

(1 太极计算机股份有限公司, 北京 100083; 2 东北石油大学 计算机与信息技术学院, 黑龙江 大庆 163318;

3 哈尔滨工业大学 计算机科学与技术学院, 哈尔滨 150001)

摘要: 合同文本的智能化处理已成为企业信息化的一个重要需求。针对合同文本存在的凌乱、碎片化和无规则的问题, 本文提出了基于深度学习的合同分类模型及要素抽取模型。合同分类从标题分类和文本分类两个方向展开研究, 提出了基于注意力机制的 BiLSTM 模型进行标题分类, 基于改进的 HAN 深度学习模型进行文本分类, 有效地提升了文本分类的准确性; 针对存在的合同信息难以抽取的问题, 提出基于 BiLSTM-CRF 深度学习模型识别合同要素, 以准确获取合同要素信息。实验表明, 本文提出模型能够很好地应用在合同文本处理中, 能够提升分类和要素抽取的性能。

关键词: 合同文本; 文本分类; 要素抽取; 深度学习

Contract classification and element extraction model based on deep learning

ZHANG Xiaofang¹, OU Rui², RAO Panjun¹, ZHENG Yuan¹, ZHANG Lei¹, CHEN Ke¹, ZHOU Chenlian²,WANG Haochang², ZHAO Tiejun³

(1 Taiji Computer Co., Ltd., Beijing 100083, China; 2 School of Computer and Information Technology, Northeast Petroleum University, Daqing Heilongjiang 163318, China; 3 School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

[Abstract] The intelligent processing of contract text has become an important demand for enterprise informatization. In response to the problems of messy, fragmented and irregular contract texts, this paper proposes a contract classification model and element extraction model based on deep learning. Contract classification is studied from two directions: title classification and text classification. The BiLSTM model based on attention mechanism is proposed for title classification and the improved HAN model is proposed for text classification, which effectively improve the accuracy of text classification. To address the existing problem that contract information is difficult to extract, the BiLSTM-CRF model is proposed to identify contract elements to accurately obtain contract element information. Experiments show that the proposed model can be well applied in contract text processing and can improve the performance of classification and element extraction.

[Key words] contract text; text classification; element extraction; deep learning

0 引言

合同是民事主体之间设立、变更、终止民事法律关系的协议, 依法成立的合同, 受法律保护^[1]。近年来, 经济活动持续增长, 各企业需要审查的合同数量逐年上升, 合同审查陷入“数多人少”的困境。科学有效地提高合同的审查速度及准确率已然成为一个亟待解决的问题。

随着自然语言处理技术的发展, 越来越多的研究者开始关注特定领域的任务。其中, 文本分类

(Text Classification, TC) 任务对数据的管理至关重要, 是信息检索和文本挖掘的基础, 利用自然语言处理、数据挖掘和深度学习等技术, 对各类文本内容进行自动分类^[2]。信息抽取 (Information Extraction, IE) 是从海量的文本、图片和视频等数据中自动抽取用户感兴趣的结构化信息的过程^[3]。

本文研究合同文本的分类和要素抽取, 需要解决 3 个问题:

(1) 针对合同类型和要素的划分不合理的问题, 研究描述合同体系以及知识点之间的关联关系

基金项目: 太极计算机股份有限公司项目 (WBXM202101009)。

作者简介: 张晓芳 (1973-), 女, 硕士, 工程师, 主要研究方向: 企业管理创新、人工智能、知识图谱; 欧睿 (1984-), 男, 硕士, 工程师, 主要研究方向: 人工智能、自然语言处理、知识图谱等; 饶攀军 (1985-), 男, 本科, 工程师, 主要研究方向: 人工智能、自然语言处理、法务智能化; 郑元 (1984-), 男, 本科, 工程师, 主要研究方向: 人工智能、自然语言处理、知识图谱等; 张雷 (1985-), 男, 本科, 工程师, 主要研究方向: 人工智能、自然语言处理、知识图谱; 陈科 (1978-), 男, 本科, 工程师, 主要研究方向: 企业管理中台、流程智能化、人工智能; 周郴莲 (1995-), 女, 硕士研究生, 主要研究方向: 自然语言处理; 王浩畅 (1974-), 女, 博士, 教授, 主要研究方向: 自然语言处理、人工智能; 赵铁军 (1962-), 男, 博士, 教授, 博士生导师, 主要研究方向: 计算机语言学、人工智能、机器翻译。

收稿日期: 2021-10-25

哈尔滨工业大学主办 ◆ 专题设计与应用

的方法。

(2)根据合同文本的特点训练适合的分类和要素抽取的方法。

(3)确保合同要素抽取的准确率,解决合同文本训练数据少的问题。

针对上述合同文本中存在的问题,结合自然语言处理技术,本文提出了基于深度学习的合同分类及要素抽取模型,通过人工标注构建合同文本语料库。为了能够及时、有效地对合同文本进行分类,提出了改进的 Bert-BiLSTM+Attention 深度学习模型;针对合同文本存在的信息难以抽取问题,提出了基于 BiLSTM-CRF 深度学习模型获取准确的合同要素信息。实验表明,这 2 个模型在合同分类与合同要素抽取中取得了较好的效果。

1 合同文本语料库构建

1.1 数据来源

本文原始语料的收集是以《民法典》为基准、常用合同类型为中心,合同数据来源于企业已过滤掉隐私信息的真实合同,主要包括买卖、建设工程、技术服务、租赁、承揽、运输、委托、借款这 8 类合同。在人工采集合同文档时,会先去除合同文本中空白数据与不符合要求的合同数据,确保原始语料的质量。标注前,对合同文本原始数据进行预处理,主要是转换原始合同数据的格式,将其保存为 json 格式的.txt 文档。

1.2 标注过程

本文合同文本标注,遵循 3 个原则:

(1)根据不同类型合同要素的特点,进行合同数据预标注,同时汇总整理合同要素标注过程中遇到的问题,如标注要素在文本中出现多次、标注要素不准确、标注错误等,给出相应的解决方案,再进行正式标注。

(2)对于有争议的问题,参考《民法典》细则或者与领域专家商议后解决。

(3)合同标注数据一致性需达到了 90% 以上。

本文采用了多轮标注的方法,可以获得大规模合同文本标注数据,研究展开的标注过程具体如下:

(1)第一轮标注。标注人员平均分组,每组的标注人员分别单独标注同一份合同,对比每组标注人员的标注结果,对合同文本语料中存在标注不一致的问题,根据专家意见及结合《民法典》进行修正。

(2)第二轮标注。根据第一轮的标注结果,完

善标注方案后进行第二轮标注,每名标注人员将会独立标注数据,并对标注后的数据进行汇总。

(3)第三轮标注。根据前两轮标注结果进行方案完善,对不同类型的合同再次进行标注,纠正标注不一致性的数据。

为了确保合同文本标注数据的一致性,每一轮标注都会对标注人员重新进行分组,保证标注数据的准确性。

2 合同文本分类模型

文本分类是指按照一定的分类体系或者规则对文本实现自动划归类别的过程^[4]。

2.1 标题分类

合同文本的标题可以较直观地对合同文本进行分类,但是在处理大批量的合同数据时,借助分类模型可以加快分类速度,本文采用的是基于注意力机制(Attention Mechanisms, AM)的双向长短期记忆网络(Bi-directional Long Short-Term Memory, BiLSTM)模型对文本标题进行分类,模型如图 1 所示。该模型主要由 4 部分组成:

(1)采用词向量表示合同标题。

(2)利用 BiLSTM 模型获取合同标题的特征。

(3)引入 AM 表示不同合同文本标题特征的重要性。

(4)经过池化后,将利用分类器进行合同标题分类。

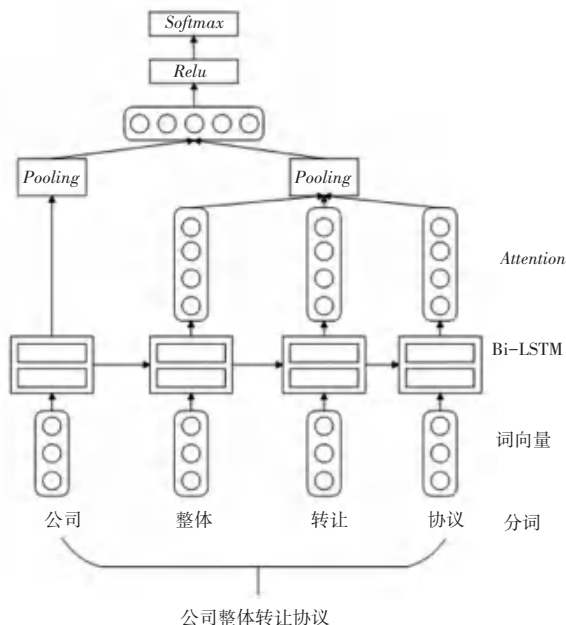


图 1 基于注意力机制的 BiLSTM 模型

Fig. 1 BiLSTM model based on attention mechanism

采用 BiLSTM 模型实现合同文本标题分类,在

训练阶段将预定义的 8 类合同文本标题嵌入到模型中。首先,将合同文本的标题进行分词处理,输入层在词向量中嵌入其维度,BiLSTM 层提取每个合同标题的特征;拼接双向的特征,将相应单词的特征作为输出;在 Attention 层对每个词语进行加权求和训练;输出层使用全连接层, Softmax 作为激活函数,输出的是合同标题的分类结果。

2.2 文本分类

合同标题虽然可以较好地完成大部分的合同文本分类任务,但也存在一些通过标题分类无法确定其类型的合同文本,此时就需要利用合同文本内容来辅助其分类。本文采用层次注意力网络(Hierarchical Attention Networks, HAN),模型结构如图 2 所示。由图 2 可知,该模型由词序列编码器、词级 Attention 层、句子序列编码器和句子级 Attention 层组成。

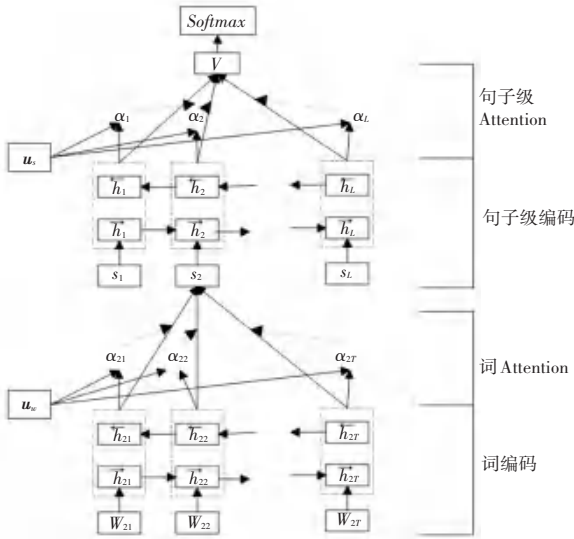


图 2 HAN 模型结构图

Fig. 2 Structure diagram of HAN model

研究中,设一份合同由 L 个句子组成,第 $s_i (i \in [1, L])$ 个句子包含 T_i 个单词。对此拟做剖析分述如下。

(1)在词编码器部分,已知句子词序列 $w_{it}, t \in [1, T]$ 组成句子,句子中每个单词的词向量经过词嵌入矩阵 W_e 把词转换成词向量,这里需用到的数学公式见式(1):

$$x_{it} = W_e w_{it} \quad t \in [1, T] \quad (1)$$

其中, w_{it} 表示句子 i 中的第 t 个单词, T 为句子的总长度。

使用双向门控循环单元(Gate Recurrent Unit, GRU)对词嵌入后的句子进行编码,结合正反方向的

上下文信息,获取隐藏层的输出,经过式(2)、式(3)的操作后,得到编码向量 h_{it}, h_{it} 包括了 w_{it} 两个方向的信息。式(2)、式(3) 具体见如下:

$$\vec{h}_{it} = \overrightarrow{GRU}(x_{it}) \quad t \in [1, T] \quad (2)$$

$$\overleftarrow{h}_{it} = \overleftarrow{GRU}(x_{it}) \quad t \in [T, 1] \quad (3)$$

(2)词级 Attention 部分。首先使用一个单层的感知机,计算编码向量 h_{it} ,得到句子 i 中的第 t 个单词的隐层向量 u_{it} ,即:

$$uit = \tanh(W_w h_{it} + b_w) \quad (4)$$

其中, W_w 表示权值矩阵, b_w 表示偏置向量。

计算词级上下文向量 u_w 的相似性,测量单词在隐层向量 u_{it} 中的重要性;通过 Softmax 进行标准化操作,得到归一化权重矩阵 α_{it}, α_{it} 表示句子 i 中第 t 个词的权重。该值可由如下公式计算求出:

$$\alpha_{it} = \frac{\exp(u_{it}^T u_w)}{\sum_t \exp(u_{it}^T u_w)} \quad (5)$$

另外,标准化时,上下文向量 u_w 是随机初始化的,同时在训练过程中也不断改变。最终一个句子的表示就是权重 α_{it} 与编码向量 h_{it} 的和,也就是 s_i ,计算过程如式(6):

$$s_i = \sum_t \alpha_{it} h_{it} \quad (6)$$

需要指出的是,这里的维度与编码向量一致。

(3)句子编码器部分。连同句子级 Attention,与上面提到的 2 层本质一样,只是将单词换成了句子,研究后推演可得:

$$\vec{h}_i = \overrightarrow{GRU}(s_i) \quad i \in [1, L] \quad (7)$$

$$\overleftarrow{h}_i = \overleftarrow{GRU}(s_i) \quad i \in [T, 1] \quad (8)$$

其中, T 为句子的总长度, L 为句子的总数。

对于给定的句子 s_i ,得到了相应的句子表示 $h_i = [\vec{h}_i, \overleftarrow{h}_i]$,包含 2 个方向的上下文信息。

计算编码向量 h_i ,得到一个隐层向量 u_i ,求得的数学表述如下:

$$u_i = \tanh(W_s h_i + b_s) \quad (9)$$

其中, W_s 表示权值矩阵, b_s 表示偏置向量。

(4)句子级 Attention 部分。提出了一个句子级别的上下文向量,计算句级上下文向量的相似性,测量句子在隐层向量 u_i 里的重要性;通过 Softmax 进行标准化操作,得到归一化权重矩阵 α_i ,数学定义可表示为:

$$\alpha_i = \frac{\exp(u_i^T u_s)}{\sum_t \exp(u_i^T u_s)} \quad (10)$$

其中, u_s 表示句级上下文向量。

在此基础上,通过式(11)计算得到合同文本的特征向量 v :

$$v = \sum_i \alpha_i h_i \quad (11)$$

3 合同要素抽取模型

合同要素抽取,即抽取合同文本中用户需要的要素信息,如:合同编号、签订时间等。

3.1 数据预处理

对完成标注的合同数据进行预处理,获取合同要素抽取数据集。首先,将标注后的合同数据从标注软件中导出;其次,提取标注要素的信息与合同原文,提取的要素信息中如存在 20% 未出现任何要素的语句,则将其删除,同时定义合同要素词典;最后,根据合同要素词典将提取的标注要素信息与经过处理的合同原文融合,转换成适用在要素提取任务的 BIO 格式的.txt 文档。其中, BIO 格式是将每个合同要素标注为“B-X”、“I-X”或者“O”,这里的“B”表示实体的开头位置,“I”表示实体的中间或结尾位置,“O”表示不属于实体的标签。

在数据清洗中,需清除数据集中所有的空数据;在不破坏文本要素完整性的前提下,每 100 行处添加截断标志,避免输入语句过长。另外,不同类型的合同存在的要素信息有部分不一致,如“技术服务合同”中是出卖人和买受人,而“油工合同”则是发包方和承包方,这会影响抽取的效果。因此,本文需要对要素词典的类型进行部分删减,重复上述步骤,得到 BIO 格式的.txt 文档。

3.2 模型框架

本文针对合同要素抽取难以实现以及抽取效率不高等问题,提出了基于 BiLSTM-CRF 的合同抽取深度学习模型,如图 3 所示。该合同要素抽取模型基于字符的序列标注模型,由输入特征层、BiLSTM 中间层、条件随机场(Conditional Random Field, CRF)输出层这 3 个部分组成。首先,把合同要素拆成单个的字,其中每个字都用拼接的特征向量表示;然后,使用 BiLSTM 神经网络层对输入的合同要素文本序列进行建模;最后,利用 CRF 输出层生成对应的类别标签序列。这里,对模型中的各个组成部分拟给出重点解析如下。

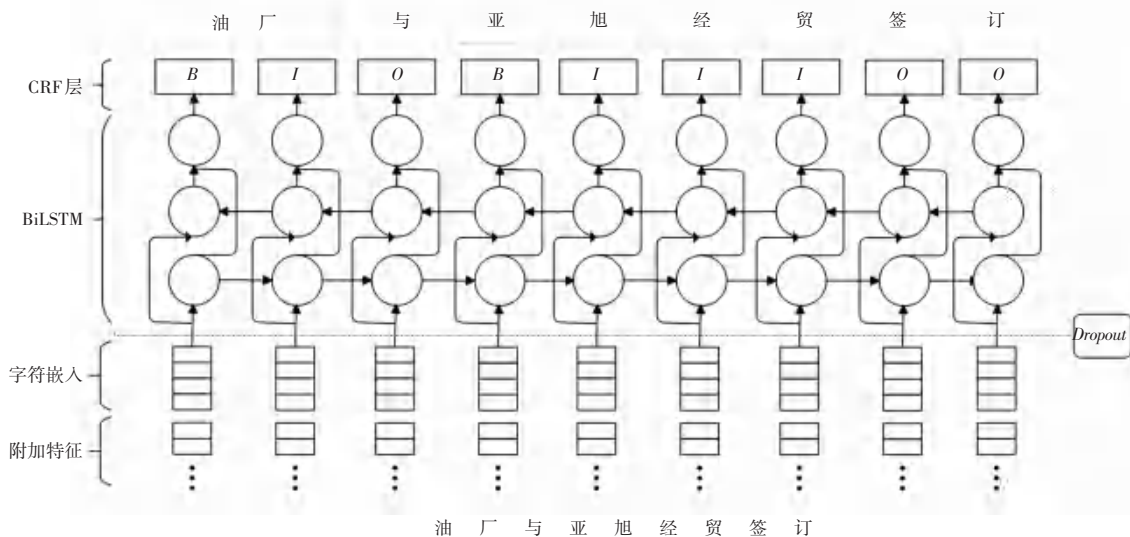


图 3 基于 BiLSTM-CRF 的合同抽取深度学习模型

Fig. 3 Contract extraction deep learning model based on BiLSTM-CRF

(1) 输入特征层。本文将训练集文本看作是字的聚合,每个字在模型的输入层由字向量(Char Embedding)与额外的特征向量(Additional Features)联结而成。

(2) BiLSTM。LSTM 是具有记忆长短期信息的能力的神经网络,前向的 LSTM 与后向的 LSTM 结合成为 BiLSTM 神经网络,网络结构如图 4 所示,可以更好地捕捉双向的语义依赖。由图 4 可知,对

LSTM 的计算过程可做分析表述如下。

设 t 时刻的遗忘门、输入门和输出门的输出为 f_t 、 i_t 、 o_t ,由此计算得到的数学公式为:

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \quad (12)$$

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \quad (13)$$

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \quad (14)$$

其中, x_t 是 t 时刻的输入向量; h_{t-1} 是 t 时刻的隐藏层的向量; σ 表示使用的激活函数是 Sigmoid 函

数; W_f 、 W_i 、 W_o 分别为 t 时刻的遗忘门、输入门和输出门对应的权重矩阵; b_f 、 b_i 、 b_o 分别为 t 时刻的遗忘门、输入门和输出门对应的偏置向量。

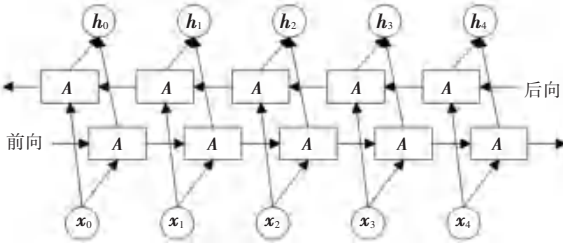


图 4 BiLSTM 神经网络结构

Fig. 4 Structure of BiLSTM neural network

利用 \tanh 层创建候选向量 \tilde{C}_t , 该向量将会被加载细胞的状态中, \tilde{C}_t 为 t 时刻单元状态, 研究得到的数学公式为:

$$\tilde{C}_t = \tanh(W_c [h_{t-1}, x_t] + b_c) \quad (15)$$

其中, W_c 是权重矩阵, b_c 是偏置向量。

将 \tilde{C}_t 与 W_c 、 b_c 两个向量结合创建更新值。

紧接着, 计算长记忆 C_t 与短记忆 h_t 的值, C_t 是 t 时刻的单元状态, h_t 是隐节点的输出。具体数学公式可写为:

$$C_t = f_t C_{t-1} + i_t \tilde{C}_t \quad (16)$$

$$h_t = o_t \tanh(C_t) \quad (17)$$

其中, f_t 、 o_t 分别是 t 时刻的遗忘门、输出门的输出。

输入序列信息是 BiLSTM 循环神经网络提取的合同要素, 结合前后向的 LSTM 输出, 传进 CRF 层。

(3) CRF 模型输出层, 得到合同文本的要素抽取结果。在 BiLSTM-CRF 模型中, 首先定义了句子 X 输出标签序列 y 的分值 $s(X, y)$, 这里推得的数学运算式可写为^[5]:

$$s(X, y) = \sum_{i=0}^n A_{y_i, y_{i+1}} + \sum_{i=1}^n P_{i, y_i} \quad (18)$$

其中, A 是转移矩阵, 表示将所有状态向下一步转移的概率; P 是 LSTM 输出的矩阵。

利用维特比算法推断合同要素标签, 搜索具有最高条件概率的合同要素标签序列 y^* , 计算公式的数学表述如下:

$$y^* = \arg \max_{y \in Y_X} s(X, \tilde{y}) \quad (19)$$

其中, $\tilde{y} \in Y_X$, Y_X 表示 y 所有可能的标签序列。

在输入层和 BiLSTM 层中间, 使用一个 Dropout 正则化, 使得模型中神经网络的节点随机失活, 在训练时不会出现某一个节点权重过大的情形。

4 实验结果与分析

4.1 评价指标

本文在合同文本分类和要素抽取的任务中, 采用准确率 (P)、召回率 (R) 和 F_1 值作为评价指标, 各指标的数学运算公式见如下:

$$P = \frac{TP}{TP + FP} \quad (20)$$

$$R = \frac{TP}{TP + FN} \quad (21)$$

$$F_1 = \frac{2PR}{P + R} \quad (22)$$

其中, TP 为实际值和预测值均为正值时数据的数量; FP 为实际值为负值、预测值为正值时的数据数量; FN 为实际值为正值、预测值为负值时数据的数量。

4.2 合同分类实验

在合同分类中, 选用合同文本的中文文本数据集。从合同数据中抽取了 2 000 条标题和 2 000 条文本数据, 其中的 1 500 条作为训练集, 500 条作为验证集。合同文本包括 8 个类别, 即买卖、建设、服务、租赁、承揽、运输、委托、借款。

神经网络模型取得良好性能的关键是选择合适的超参数。合同分类模型参数设置: 层数为 30 层、词向量维度为 768 维、Dropout 的值为 0.5 和学习衰减率为 0.001。

4.2.1 标题分类

为了验证本文提出的基于注意力机制的 BiLSTM 模型用于合同标题分类中的有效性, 将本文模型 BiLSTM + Attention 与 TextRNN、LSTM 和 BiLSTM 模型进行合同标题分类的对比实验, 其实验结果见表 1。

表 1 标题分类不同模型的实验结果

Tab. 1 Titles test results of different models			%
模型	准确率	召回率	F_1 值
TextRNN	88.24	82.05	84.86
LSTM	89.36	83.95	86.57
BiLSTM	91.38	88.89	89.86
BiLSTM+Attention	90.55	90.18	90.33

由表 1 可知, 在合同标题分类中, 本文模型的召回率和 F_1 值比其他 3 个对比模型高。TextRNN 模型的准确率、召回率和 F_1 值分别比本文模型低 2.37%、8.13% 和 5.67%; LSTM 模型准确率、召回率和 F_1 值比本文模型低 1.19%、6.23% 和 3.76%;

BiLSTM 模型准确率比本文模型高 0.83%, 但召回率和 F_1 值却比本文模型低 1.29% 和 0.47%。综合来看, 在标题分类性能对比中, 本文提出的 BiLSTM+Attention 模型标题分类效果最好。

4.2.2 文本分类

将本文模型 BiLSTM+HAN 与 TextRNN、CNN 和 RCNN 模型进行文本分类的实验, 验证该模型用于文本分类的有效性, 实验结果见表 2。

表 2 文本分类不同模型的实验结果

Tab. 2 Text test results of different models %

模型	准确率	召回率	F_1 值
TextRNN	84.56	87.29	85.73
CNN	90.01	85.55	86.59
RCNN	91.35	85.57	88.34
BiLSTM+HAN	90.74	90.83	90.76

由表 2 可知, BiLSTM+HAN 模型的要素抽取 F_1 值达到了 90.76%, 且在准确率和召回率上的值都达到了最佳。本文模型比 TextRNN 模型准确率、召回率和 F_1 值分别提高了 6.18%、3.54% 和 5.03%; 比 CNN 准确率、召回率和 F_1 值分别提高了 0.73%、5.28% 和 4.17%; 虽比 RCNN 准确率低 0.61%, 但召回率和 F_1 值却提高了 5.26% 和 2.42%。本文提出的 BiLSTM+HAN 模型在合同文本分类上效果明显优于其他模型。

4.3 合同要素抽取

合同要素抽取模型中各参数设置如下: 学习率为 0.015、隐藏层大小为 200、Dropout 的值为 0.5、层数为 32 层和学习衰减率为 0.05。

为了验证本文提出的基于 BiLSTM-CRF 的合同抽取模型的有效性, 将本文模型与 CRF、LSTM、Flat 和 Lattice LSTM 模型进行合同要素抽取的对比实验, 抽取效果见表 3。

表 3 不同模型的合同要素抽取效果

Tab. 3 Contract elements extraction effect of different models %

模型	准确率	召回率	F_1 值
CRF	69.21	65.96	67.46
LSTM	70.73	76.30	73.41
Flat	70.34	84.37	76.72
LatticeLSTM	71.27	86.68	78.13
BiLSTM-CRF	72.10	88.00	79.26

由表 4 可知, 相较于其他要素抽取模型, 本文提出的 BiLSTM-CRF 模型的抽取效果最好。与判别式模型 CRF 相比, 本文模型的准确率、召回率和 F_1 值分别提高了 2.89%、22.04% 和 11.8%; 与 Lattice LSTM 模型相比, 本文模型的准确率、召回率和 F_1 值分别提高了 0.83%、1.32% 和 1.13%。综合来看, 本文提出的基于 BiLSTM-CRF 的合同抽取模型在进行合同要素抽取的任务上, 明显优于其他模型。

5 结束语

本文根据《民法典》、合同文本及要素的特点, 构建了合同文本的语料库, 提出了基于深度学习的合同分类及要素抽取模型, 实现了对合同文本标题分类、文本分类和要素抽取, 并实验验证了本文提出模型的有效性。

在合同文本分类中从标题分类与文本分类两方面展开研究, 提出了一种基于注意力机制的 BiLSTM 模型, 采用词向量表示合同标题, 利用 BiLSTM 模型获取合同标题的特征, 实现了对合同标题的分类任务; 提出了一种基于改进的 HAN 深度学习模型, 经过词序列编码器、词级 Attention 层、句子序列编码器和句子级 Attention 层实现了对合同文本分类。提出了一种基于改进的 HAN 深度学习模型, 将合同要素拆成单个的字, 使用包含前向和后向的 BiLSTM 神经网络层对输入的合同要素文本序列进行建模, 利用 CRF 层作为模型的输出层生成对应的类别标签序列, 实现合同要素抽取, 并实验证实了合同要素抽取模型的有效性。

在未来工作中, 将尝试实现更多不同类型合同分类和抽取任务, 进一步提高合同文本分类与要素抽取的性能。

参考文献

- [1] 中华人民共和国民法典[N]. 人民日报, 2020-06-02(001).
- [2] SEBASTIANI F. Text categorization[M]//Encyclopedia of Database Technologies and Applications. IGI Global, 2005: 683-687.
- [3] 李保利, 陈玉忠, 俞士汶. 信息抽取研究综述[J]. 计算机工程与应用, 2003(10): 1-5, 66.
- [4] 于游, 付钰, 吴晓平. 中文文本分类方法综述[J]. 网络与信息安全学报, 2019, 5(05): 1-8.
- [5] 李明浩, 刘忠, 姚远哲. 基于 LSTM-CRF 的中医医案症状术语识别[J]. 计算机应用, 2018, 38(S2): 42-46.