

基于Text-CNN联合分类与匹配的合同法律智能问答系统研究

刘葛泓, 李金泽, 李卞婷, 邵南青, 窦万峰

(南京师范大学, 江苏 南京 210023)

✉1512509129@qq.com; 739036728@qq.com; libianting1999@163.com;
18001576758@163.com; douwanfeng@njnu.edu.cn



摘 要: 面向法律领域的相关问题, 需要借助专业的法律文本。利用司法领域的文本资源解决用户提出的合同法律问题, 能在很大程度上降低人工成本, 节约社会资源。为了更加智能、高效地响应用户在合同法方面的法律诉求问题, 本文设计与实现了一个合同法律智能问答系统, 并给出了一种基于文本卷积神经网络(Text-CNN)的联合分类与匹配的合同法律智能问答深度学习模式, 针对合同法领域的文本特征, 对其进行了分类。实验表明, 该模式适合于合同法领域的智能问答。

关键词: 合同法律智能问答系统; 文本卷积神经网络; 自然语言处理; 词向量

中图分类号: TP391.1 **文献标识码:** A

Research on Intelligent Contract Law Question-and-Answer System based on Text-CNN Joint Classification and Matching

LIU Gehong, LI Jinze, LI Bianting, SHAO Nanqing, DOU Wanfeng

(Nanjing Normal University, Nanjing 210023, China)

✉1512509129@qq.com; 739036728@qq.com; libianting1999@163.com;
18001576758@163.com; douwanfeng@njnu.edu.cn

Abstract: Relevant issues in the legal field require professional legal texts. The text resources in the judicial field can be used to deal with the legal questions concerning contracts raised by users, which can greatly reduce labor cost and save social resources. In order to more intelligently and efficiently respond to users' legal claims in contract law, this paper designs and implements an intelligent contract law Question and Answer System (QAS), and proposes a deep learning model of intelligent QAS focusing on contract law based on the technology of joint classification and matching of Text Convolutional Neural Network (Text-CNN), which can classify different texts of contract law according to their characteristics. Experiments show that this model is suitable for intelligent question answering in the field of contract law.

Keywords: intelligent contract law question-and-answer system; Text-CNN; natural language processing; word vector convolutional neural network; natural language processing; the word vector

1 引言(Introduction)

随着我国市场经济的发展, 合同纠纷成为人们生活中最常出现的法律问题之一。因此, 精准高效地获取法律援助成为每个合同主体的诉求。随着互联网的发展, 传统的搜索引擎已不能满足用户的需求, 智能问答系统应运而生。问答系

统的核心是文本匹配^[1]。虽然Prolo^[2]、Monroy等人^[3]在该领域有一定贡献, 但一定程度上依赖于人工标注、答案范围有限。同时, 针对合同领域问答系统的相关研究较少, 未有较好的解决方案。

针对上述问题, 本文通过词向量技术构建合同法语料

库，采用联合加分类的方式，利用文本卷积神经网络(Text-CNN^[4])模型对语料库进行特征分类，划分问题类别，缩小答案映射范围。经实验，Text-CNN文本分类模型适合于问题特征复杂且多样的合同法律问答系统，相较于LSTM^[5]模型在文本分类上拥有更高的效率。

2 系统设计与实现(System design and implementation)

针对本文提出的问答模式，建立了合同法律问题援助智能问答系统框架。该系统将合同法律科普与自动化在线咨询功能结合在一起，提供一个智能高效的法律知识咨询平台。基于系统的功能定位，可以将系统划分为三个部分：第一部分是问答系统前端APP形式的人机交互界面；第二部分是应用服务器，是系统的核心部分，主要负责处理用户的问题并反馈答案与相关信息；第三部分是存储合同法律问题相关领域知识的数据库。相应的系统框架如图1所示。

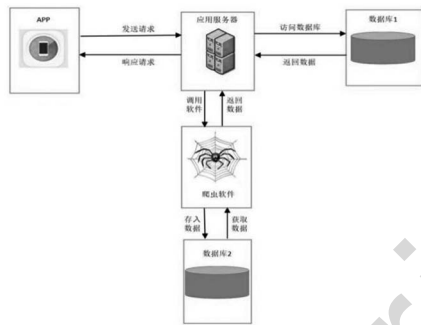


图1 App架构图

Fig.1 App architecture diagram

本系统的核心功能的实现首先需要搭建一个结构合理的系统，总体框架由三部分构成：第一部分是负责用户和系统之间信息交互的用户接口层；第二部分是应用各种算法和分词技术处理问句的业务逻辑层；第三部分是负责数据存储及使用的数据库层。其结构如图2所示。

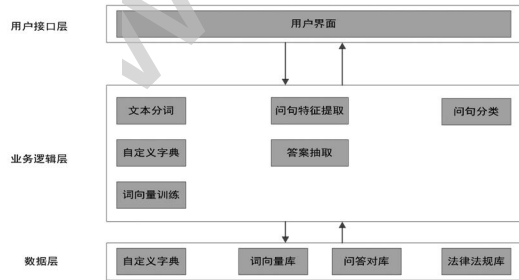


图2 App框架图

Fig.2 App architecture diagram

3 模型描述与研究方法(Model description and research methods)

本文将研究任务定义为在合同法领域语料库中查找可

能包含答案的段落(句子)^[1]，形式化表示如下：给定一个问题Q，系统将针对Q从领域语料库中查找可能的答案，语料库由分类后的问答对数据库与法规文档构成。问答对数据集为 $\{(A_1, a_1), (A_2, a_2), \dots, (A_m, a_m)\}$ ，法律法规文档集为 $\{D_1, D_2, \dots, D_n\}$ 。答案抽取首先要判断Q属于什么问题类别，在此之后的过程分为两种情况：

- (1)针对Q能够从标准问答对数据集中检索到相似问题，则返回相似问题所对应的答案集 $T=\{t_1, t_2, \dots, t_s\}$ ；
- (2)未找到相似问题，则检索法规文档集中可能包含答案的集合 $S=\{S_1, S_2, \dots, S_j\}$ 返回给用户。

问答示例如表1和表2所示，其中Q₁、Q₂为用户查询问题，T₁、T₂和T₃为已知问答对中相似问题的答案，A为结合法律法规后答案。

表1 合同问答示例1

Tab.1 Contract Q&A example 1	
形式化表示	内容
Q ₁	买卖双方签订了二手房买卖合同，购房时不知有人在房内死亡，可否主张撤销合同？
S	您可以咨询专业律师，为您找到如下法规：根据《中华人民共和国合同法》“……受害者有权请求仲裁机构变更或撤销。”

表2 合同问答示例2

Tab.2 Contract Q&A example 2	
形式化表示	内容
Q ₂	建设工程合同无约定逾期竣工的违约金怎么办
T ₁	可以与承包人协商解决，包括交付时间、逾期交付的违约金、给发包人造成的损失等
T ₂	发包人对承包人造成的逾期竣工的违约行为可以主张解除合同
T ₃	要求承包方承担违约赔偿责任

本文研究的合同法问答系统基本流程如图3所示。首先，系统搜集相关资料构建领域语料库并将其训练成词向量，通过Text-CNN模型进行问题分类。至此，准备工作完成。然后，系统接收用户输入的自然语言问句，对其进行预处理，获得查询向量。接着，使用语义相似度进行匹配检索。在问答对数据集模块，返回匹配度较高的历史问题所对应的答案。在法律法规模块，通过关键词特征对检索结果排序。最后，将在问答对数据集中匹配到的问题所对应的答案推荐给用户，对于未匹配成功的问题，返回相关的法规条款作为参考方案。

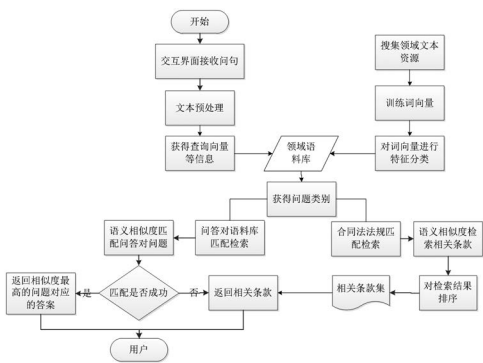


图3 系统流程图

Fig.3 System flow chart

3.1 领域语料库设计

3.1.1 数据来源

领域语料库主要分为两大板块：问答对数据库和法律法规数据库。语料库的构建主要通过网络爬虫技术^[6]。

(1)问答对数据库。在该板块语料的采集目标主要为社区问答模块中常见热门问题的问答对。以找法网为例，(这是国内一家大型法律资讯信息网站)系统共筛选出1563条合同法领域法律问答对，22万条语料内容，包含常见问题、答案、关键标签以及所属类别。

(2)法律法规数据库。通过中华人民共和国中央人民政府官网下载合同法相关法律法规，并进行处理。将每一项合同法律条款作为一行数据，通过关键词技术提取每段数据的关键词作为当前法律法规的标签保存在数据库中。

表3 语料库数据集

Tab.3 Corpus data set

数据集	数量
合同法问答对	1563条
相关语料内容	22万条
法律法规数据库	524条

3.1.2 特征表示与特征提取

为了能够利用构建好的领域语料库进行相关研究，需要将字符形式的文本通过某种编码方式让计算机理解。由于中英文差异，汉字需要先进行分词。中文字词在计算机中的表示方法通常分为两种：独热表示(One-hot Representation)^[7]和分布式表示(Distribution Representation)^[8]。独热表示通常是把每个词都表示成一个向量，该方法直观、易解释，但是不能很好地展现词与词之间的语义关系，且会造成特征空间大的缺点^[9]。而分布式表示是把每个词表示为长度相同的连续稠密词向量。分布式表示相对于独热表示，不仅降低了特征维度，令矩阵变得稠密，而且词语之间的语义关系更为明显^[9]。

本文采用分布式表示方法。word2vec模型对语料库进行训练，其基本思想是根据上下文环境中的词来预测文本中心词。word2vec^[10]是常用的词嵌入方法。该模型是根据词汇的co-occurrence信息进行编码。word2vec是Google在2013年推出的一个NLP工具，在word2vec中用到两个重要模型结构：Skip-gram结构和CBOW(Continuous Bag-of-Words Model)结构。通过该模型，系统完成了对语料库的特征表示与提取工作，将语料库训练成词向量。

3.2 问句分析

3.2.1 文本预处理

问句的预处理过程是提取重要的信息并处理成字ID或者词ID的序列^[11]，包括分词(运用jiebia技术)、去停用词、去低频词和词性标注等；然后对问句成分进行特征化，此处主要是将句子训练成词向量，通过词嵌入(Word Embedding)技术将文本数据从高纬度稀疏变为低纬度稠密的数据；同时抽取和扩展问句的关键词(使用TextRank算法)，以便后续的相似度研究。

3.2.2 问句分类处理

针对普通法律问题检索内存消耗大、效率不高的问题，本文提出了联合分类与匹配模式的智能问答模型。参照常见问题集(Frequent Asked Question, FAQ)、社区问答集(Community Question Answering, CQA)的历史数据，首先人工标注将问题分为五大类别：合同订立、合同效力、合同纠纷、合同解除和合同文本。借鉴卷积神经网络(CNN)^[11]在图像处理过程中的特点，利用针对文本分类的Text-CNN模型对合同法问题进行特征提取与分类，缩小映射范围。该模型属于文本匹配的改进方案，即先分类、再匹配。

卷积神经网络^[12]最早应用于图像方面，该模型在不同的位置都可以共享权重。比如相同的一个物体，在图片中位于不同的位置，而物体的特征不变，因此CNN可以很好地提取对象的局部特征^[12]。CNN的基本结构由五部分构成：输入层、卷积层(convolutional layer)、池化层(pooling layer)、全连接层和输出层。其中，卷积层和池化层一般为若干个，将卷积层和池化层交替设置，也即一个卷积层连接一个池化层，池化层之后再连接一个卷积层，以此类推^[12]。

CNN在图像方面取得重大成功后，人们逐渐将它运用到自然语言处理当中。不同于图像是二维的，文本信息是一维的。应用卷积之前，需要把输入的文本通过前面介绍的词向量技术进行词嵌入操作，通过卷积层、池化层提取特征，然

后对提取到的特征进行分析就可以对文本进行分类。而单层的卷积网络学习到的特征一般是局部的，因此为了提高分类效率，使学习到的特征更加全面化。本文针对合同法领域的文本特征，参考YoonKim^[4]提出的Text-CNN模型构造了如图4所示的合同法领域文本卷积神经网络模型流程图。

假设要对一些句子进行分类，其中，输入层为 $m \times n$ 的文本矩阵(m 、 n 分别表示句子长度和句中每个词的词向量维度)。Text-CNN通过不同的通道数目和卷积核大小，使用一维卷积的方式提取句子的特征。池化层从卷积后的向量中选取最大值并与其他通道的最大值拼接，组合得到这个句子的特征表示，通过(全连接层)SoftMax层进行分类。

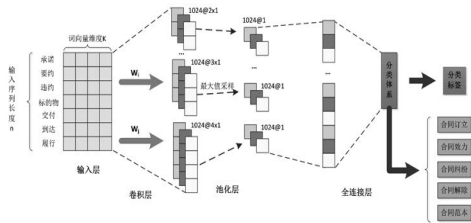


图4 合同法领域Text-CNN流程图

Fig.4 Text-CNN flow chart in the field of contract law

3.3 相似度计算

词向量技术作为语义相似度计算方法之一，可以把文本映射成为连续空间中的向量^[13]。这样一来，文本之间的语义相似度就可以通过词向量之间的余弦相似度表示^[13]。在获得文本对应的词向量后，本文采用词袋模型^[13]得到每个句子的句向量表示，即对句子中的每个词向量通过式(1)进行平均值计算。

$$S = \frac{1}{|S|} \sum_{i=1}^{|S|} s_i \quad (1)$$

其中， S 是求得的句向量， s_i 为每个词的词向量表示。由此就可以计算出每个句子的句向量，然后利用余弦相似度比较两个句子之间的语义相似度。对于每个文本组合 (q, a) ^[13]，词向量余弦相似度计算方法如式(2)所示。

$$\text{emb}_{\text{score}}(q, a) = \frac{q_{\text{emb}} \cdot a_{\text{emb}}}{|q_{\text{emb}}| \times |a_{\text{emb}}|} \quad (2)$$

式(2)的分子部分为向量的内积，它是用于计算距离的方式。一般情况下，内积计算的缺点是向量长度的过大或过小会影响后面的度量结果，但通过余弦相似度的方式除以他们的长度或者先对向量进行归一化处理之后，再计算内积，可以规避该问题。

4 实验结果与分析(Experimental results and analysis)

4.1 数据集向量化处理

在前面，我们针对合同法领域建立了专门的领域语料

库，建立分词结果的词汇表并赋予索引，设置滑动窗口大小为5，即考虑一句话中当前词与上下文词的最大距离(单侧词窗)为5；词向量维度，即训练时隐变量的维度为200。词向量模型的输入为已搜集到的经停用词处理后的语料集，输出为模型文件`_model`和词向量`_vector`文件。

实验训练结果部分如图5所示，以词汇表中不重复的词作为中心词，取规定窗口内的单词(此处上下文各取5个)记作 w ，预测当上下文出现 w 时输出该中心词的概率。下图给出了在某个问题中以“的”和“合同”为中心词的词向量训练结果。

```
0.09928781], '办事机构': [0.010012993, 0.027931659, -0.01914956, 0.016601967,
0.019914482, 0.03333435, -0.0073872744, 0.005399299, -0.012709207,
0.0030585509, -0.0085344, -0.000865503, -0.0054758876, -0.04725379,
9.4841234e-05, 0.024302216, 0.022591814, -0.017802933, -0.041542977,
0.009066959, 0.024969092, 0.004293297, 0.015783038, 0.014815913, -0.019944614,
-0.027754715, 0.020683808, -0.015583595, -0.016291937, 0.03000783,
0.063426115, 0.010435746, 0.014239427, 0.011605136, 0.01741648, -0.027236147,
```

图5 语料库训练结果集

Fig.5 Corpus training result set

4.2 问题分类

4.2.1 模型对比

在短文本分类领域，卷积神经网络(Convolutional Neural Network, CNN)和长短期记忆神经网络(Long Short-Term Memory neural network, LSTM)是两个主流的神经网络模型^[14]。

为了测试本文提出的Text-CNN模型对合同法领域的文本进行分类的优越性，这里使用相同语料集分别测试基于LSTM^[5]和Text-CNN两种模型的分类型性能。测试内容分为三块：模型训练时长、训练集准确率和测试集准确率。其中，准确率定义公式如(3)所示。

$$\frac{\sum_{i=1}^N \frac{\sum_{j=1}^n A_i \cdot B_j}{\sqrt{\sum_{j=1}^n (A_j)^2 \cdot \sum_{j=1}^n (B_j)^2}}}{N} \quad (3)$$

其中， N 为问答集数量， A 、 B 分别表示分类结果与对比答案的词向量矩阵，下标 i 表示每一组的词向量。

(1)模型构建。在本次实验所构建的Text-CNN模型中，借助训练好的词向量依次获取用户输入问题和分类结果的句向量表示(其中未登录词用左右词的词向量平均值表示)。最后计算两者的向量化表示的余弦值，并将其作为两者的相似度表示。

(2)模型细节。在训练之前，问答集均已经过分词、去停用词等预处理工作。Text-CNN模型使用基于一维卷积层的

序列分类，训练时以100个样本为一个batch进行迭代，利用200维词向量处理问答并作为模型输入。

(3)训练结果分析。语料集为问答集qa_corpus中的1563条数据，其中80%作为训练集，20%作为测试集。实验结果如表4所示，由表我们可以发现针对合同法领域的文本分类处理Text-CNN无论是精度还是速度都优于LSTM模型。

表4 模型实验结果

Tab.4 Model experimental results

模型	模型训练时长	训练集准确率	测试集准确率
Text-CNN	9s	0.984	0.959
LSTM	28s	0.838	0.717

4.2.2 系统测试

本文利用卷积神经网络对合同法领域文本进行分类，将问题分为五大类：合同订立、合同效力、合同纠纷、合同解除与合同文本。在问题分类之后，我们对所构建的问答系统进行了测试。如图6所示，用户输入一个有关合同法的问题“建设工程合同无约定逾期竣工的违约金怎么办”，系统给出回答，该问题属于“合同纠纷”。经法律专业人士判断，该分类正确。



图6 问题分类测试结果图

Fig.6 Problem classification test result graph

4.3 答案抽取

智能问答系统的最后一个步骤为答案抽取。由于本文的研究重点在于针对合同法领域的问题分类与匹配，故答案排序的实现借助ElasticSearch。以问题“建设工程合同无约定逾期竣工的违约金怎么办”为例，图7是得到答案的过程。

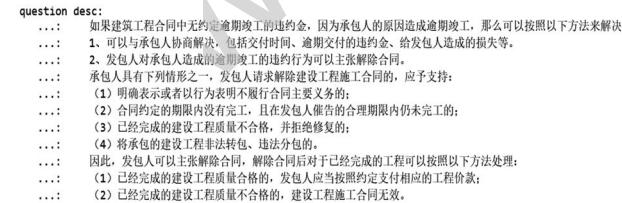


图7 答案抽取结果图(检索问答对)

Fig.7 Answer extraction result graph (retrieval of question and answer pairs)

如果不存在相似的问题，则将输入问题的关键词作为索引查询法规数据库。以问题“买卖双方签订了二手房买卖合同，购房时不知有人在房内死亡，可否主张撤销合同”为

例，如图8所示为查询之后的答案。

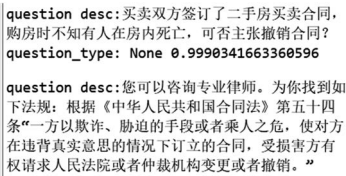


图8 答案抽取结果图(检索法律法规数据库)

Fig.8 Answer extraction result graph (retrieve laws and regulations database)

5 结论(Conclusion)

本文研究的合同法领域问答系统实现模式，主要贡献有三点。(1)根据合同法领域的本文特征，本文提出一种结合法律法规与常见问题问答对的两阶段答案抽取方法，在进行了句子预处理后，将问句先与问答对数据库进行匹配，若匹配成功，则返回相似度最高的历史问题对应的答案，否则，在法律法规中查找最优答案，提高了获得答案的效率。(2)针对合同法领域问题的特点，本文利用Text-CNN文本分类模型对合同法问题进行了分类，从而在答案匹配时，缩小了映射范围。经实验证明，该方法优于传统的LSTM模型，提高了答案检索效率。(3)根据专业数据资源创建的合同法领域问答测试集，并利用本文的方法进行了测试，实验结果也证明了该模式的有效性。

在之后的研究中，我们将进行如下工作：(1)构建适用于合同法领域的排序学习算法，提高答案选择的效率。(2)在更深的层次研究和分析合同法领域的文本特征，研究特征选择方法^[15]，并对其进行比较和评估，以更好地优化合同法问答系统。

参考文献(References)

[1] 仇瑜,程力.特定领域问答系统中基于语义检索的非事实型问题研究[J].北京大学学报(自然科学版),2019,55(1):55-64.

[2] Prolo C, Quaresma P, Rodrigues I, et al. A Question-answering System for Portuguese Knowledge andReasoning for Answering Questions[C].Workshop Associated with IJCAI05. Edinburgh, 2005: 45-48.

[3] Monroy A, Calvo H, Gelbukh A. NLP for shallow question answering of legal documents using graphs[C].International Conference on Intelligent Text Processing and Computational Linguistics. Mexico City, 2009: 498-508.

[4] Yoon Kim. Convolutional Neural Networks for Sentence Yoon Kim[D]. Computer Science: Computation and Language, 2014: 1-6.

(下转第4页)