软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software [doi: 10.13328/j.cnki.jos.006412]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

融入法律知识的问句匹配*

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

[1](昆明理工大学 信息工程与自动化学院, 云南 昆明 650504)
[2](昆明理工大学 云南省人工智能重点实验室, 云南 昆明 650504)
[3](模式识别国家重点实验室 (中国科学院自动化研究所), 北京 100190)
通信作者: 余正涛, E-mail: ztyu@hotmail.com

摘 要: 问句匹配是问答系统的重要任务, 当前方法通常采用神经网络建模两个句子的语义匹配程度. 但是, 在法律领域中, 问句常存在文本表征稀疏、法律词的专业性较强、句子蕴含法律知识不足等问题. 因此, 通用领域的深度学习文本匹配模型在法律问句匹配任务上效果并不好. 为了让模型更好的理解法律问句的含义、建模法律领域知识, 首先构建一个法律领域知识库, 在此基础上提出一种融合法律领域知识 (如法律词汇和法律法条) 的问句匹配模型. 具体地, 构建了合同纠纷、离婚、交通事故、劳动工伤、债务债权等 5 种法律纠纷类别下的法律词典, 并且收集了相关法律法条, 构建法律领域知识库. 在问句匹配中, 首先查询法律知识库检索问句对所对应的法律词汇和法律法条, 进而通过交叉关注模型同时建模问句、法律词汇、法律法条三者之间的关联, 最终实现更精准的问句匹配, 在多个法律类别下的实验表明提出的方法能有效提升问句匹配性能.

关键词: 法律问句匹配; 法律词典; 法律法条; 法律领域知识库
中图法分类号: TP391

中文引用格式: 刘权, 余正涛, 何世柱, 刘康, 高盛祥. 融入法律知识的问句匹配. 软件学报. http://www.jos.org.cn/1000-9825/6412.htm
英文引用格式: Liu Q, Yu ZT, He SZ, Liu K, Gao SX. Incorporating Legal Knowledge into Question Matching. Ruan Jian Xue Bao/Journal of Software (in Chinese). http://www.jos.org.cn/1000-9825/6412.htm

Incorporating Legal Knowledge into Question Matching

LIU Quan[1,2], YU Zheng-Tao[1,2], HE Shi-Zhu[3], LIU Kang[3], GAO Sheng-Xiang[1,2]

[1](Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650504, China)
[2](Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology, Kunming 650504, China)
[3](National Laboratory of Pattern Recognition (Institute of Automation, Chinese Academy of Sciences), Beijing 100190, China)

Abstract: Question matching is an important task of question answering systems. Current methods usually use neural networks to model the semantic matching degree of two sentences. However, in the field of law, questions often have some problems, such as sparse textual representation, professional legal words, and insufficient legal knowledge contained in sentences. Therefore, the general domain deep learning text matching model is not effective in the legal question matching task. In order to make the model better understand the meaning of legal questions and model the knowledge of the legal field, this study firstly constructs a knowledge base of the legal field, and then proposes a question matching model integrating the knowledge of the legal field (such as legal words and statutes). Specifically, a legal dictionary under five categories of legal disputes has been constructed, including contract dispute, divorce, traffic accident, labor injury, debt and creditor's right, and relevant legal articles have been collected to build a knowledge base in the legal field. In question matching, the legal knowledge base is first searched for the legal words and statutes corresponding to the question pair, and then the

* 基金项目: 国家重点研发计划 (2018YFC0830105, 2018YFC0830101, 2018YFC0830100); 国家自然科学基金 (61761026, 61972186, 61762056)
收稿时间: 2020-10-24; 修改时间: 2021-01-13, 2021-05-21; 采用时间: 2021-06-30; jos 在线出版时间: 2022-07-15

relationship among the question, legal words, and statutes is modeled simultaneously through the cross attention model. Finally, to achieve more accurate question matching, experiments under multiple legal categories were carried out, and the results show that the proposed method in this study can effectively improve the performance of question matching.

Key words: legal question matching; legal words; statues; legal domain knowledge base

问答系统[1–3]是信息检索的一种高级形式, 能够更加准确地理解用户用自然语言提出的问题, 并通过检索语料库、知识图谱或问答知识库返回简洁、准确的匹配答案, 而随着法制社会的发展, 问答系统在法律信息服务中的应用也逐渐广泛了起来. 相较于传统的法律咨询, 法律问答系统能更好地理解用户提问的真实意图, 进而更有效地满足用户的法律援助需求.

法律问句匹配是法律问答系统应用技术中的一项重要任务[4], 通过法律问句匹配实现检索和匹配相似法律问句的功能. 通常做法是将用户输入问句与问答库中的问题进行匹配, 检索相似度最高的问句, 并将检索后的问句答案返回给用户. 因此, 如何高效准确地匹配法律问句成为法律智能问答系统的核心问题.

问句匹配任务具体是指: 给定前提问句, 根据该去推断假设问句与前提问句的关系, 一般分为蕴含关系和矛盾关系, 蕴含关系表示从前提问句中可以推断出假设问句; 矛盾关系即假设问句与前提问句矛盾. 法律领域内的问句匹配任务是将用户的法律咨询问题当作前提问句, 库中储存的法律问题作为假设问句, 通过深度匹配模型去推断两个问句之间的关系, 从而获得准确的匹配效果. 问句匹配任务从模型层面上分为表示型[5,6]、交互型[7,8]和预训练语言模型[9]这 3 类: (1) 表示型模型对将要匹配的两个句子分别进行编码与特征提取, 最后进行相似度计算. 例如, 深度结构语义匹配模型[5]通过 3 个全连接层串行构成特征抽取层, 然后进行相似度计算. 这类模型能将文本映射为一个简洁的表示, 便于离线计算并储存, 模型匹配的计算速度快, 但缺点是这类模型只建模了句子的简单对应关系, 而实际上问句表达是有层次、有结构的, 仅分别从两个对象单独提取特征, 很难捕获问句中的结构信息. (2) 交互型模型通过将两个句子进行交互, 利用获得的特征进行进一步的匹配, 例如, 双边多视角匹配模型 (bilateral multi-perspective matching, BIMPM)[10]通过采用 4 种不同的交互方法进行匹配双向长短期记忆网络 (bi-directional long short-term memory, BiLSTM) 编码层输出的上下文向量. 这类模型能保持细粒度的匹配信息, 避免在一段文本抽象成一个表征时, 细节的匹配信息丢失, 但缺点是需要大量的有监督的文本匹配的数据训练, 且模型网络结构复杂、模型训练的时候资源消耗较大; (3) 由于预训练语言模型在多项 NLP 任务中的表现优越, 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, Transformer 的双向编码器表征 (bidirectional encoder representations from transformers, BERT)[9]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. BERT 由于其深层语言处理能力, 在文本匹配任务上具有天然优势.

上述 3 种不同类型的模型在通用领域的问句匹配任务上获得了不错的效果, 但是法律领域中问句常存在文本表征稀疏、法律词的专业性 (如表 1 所示)、蕴含法律知识等现象 (如表 2 所示), 因此通用领域的深度学习文本匹配模型在法律问句匹配任务上效果并不好. 具体来说, 相比于通用领域的文本, 法律问句具有其独特性, 法律领域的问句匹配面临如下问题:

表 1　法律问句样例

法律问句	法律类别	法律词汇
交通事故故意杀人罪怎么**判**?	交通事故	交通事故、故意杀人罪、判
男方**出轨**了, 我想**离**可他不同意? 没身份证可以办吗?	离婚	出轨、离

(1) 问句表征稀疏: 问句表征稀疏是法律问句匹配的常见问题. 通常模型分别将两个问句转换为语义向量, 然后计算两向量的相似度, 其侧重于对语义向量表示层的构建. 表征用来表示文本的高层语义特征, 但是法律文本中词汇的关系、法律问句的句法特征难以捕获, 很难判定一个表征是否能很好的表示一个法律问句. 如表 1 中, 用户询问“怎么判? ”“离”, 真实语境中分别代表“怎么判刑”“离婚”, 但如果不使用外部知识获得补充含义, 模型从句子层面上无法学习到这些深层含义.

(2) 法律词的专业性: 法律领域的词汇是区分问句是否属于法律领域的重要因素, 这类词如“故意伤害罪”等都是具有很强的法律领域性词汇. 而故意伤害罪这种罪名可认定为法律连绵词, 如 jieba 等分词工具会将其拆开为“故意”“伤害罪”, 但实际上“故意伤害罪”是不可拆分的; “可以办”这类在通用领域内适用性较强的词汇, 在法律领域内因为法律问句的语境变化, 该词的含义也得到了改变, 在这句话中, “可以办”的含义是“可以办离婚”. 因此需要引入法律领域词典解决法律问句中法律词的联绵性和适应语境性.

(3) 法律知识不足: 法律问句除了蕴含法律词汇, 还有着更深层次的法律知识. 如表 2 为法律问句对样例. 我们通过分析看出两个问句长度均较短, 且蕴含法律词汇, 但法律词汇是从句子层面上唯一能挖掘的法律知识, 因此通过提高法律词汇在问句中的权重以增强法律问句表征, 有效地增强法律含义. 但是法律词汇蕴含的知识是不足的, 因为法律问句需要权威性的法律法条来指引. 如果法律问句能结合法律法条, 既能增强语句表示, 又能解释问句匹配的原因. 表 2 所示前提问句与假设问句蕴含的法律词汇是一致的, 但是每个法律问句的背景知识是复杂的, 法律定义所有法律事实都是有法可依, 即法律事实匹配一个或者多个法条. 对于前提问句, 匹配的法条为《机动车登记规定》的第二十三条, 该法条解释了抵押机动车提交的具体材料, 其中包括了车辆登记证书; 假设问句的匹配法条为《机动车登记规定》的第二十二条, 机动车的所有人不包括句中的“别人”. 两个不同的法条都是对“车辆抵押的流程和权责”的解释, 且两个问句根据字面描述无法确认是否匹配, 所以需要通过对应法条作为桥梁进行判别. 经过专业法律人士标注, 上述问句是匹配的, 两个问句均涉及机动车登记的要素, 无论主体与行为, 都需要提交句中提到的“车辆等级证书”, 因此将法条作为依据可以认为两句话是匹配的, 所以引用法条蕴含的背景知识能正确地帮助我们匹配法律问句, 而如何融入法条知识是一个亟待解决的问题.

表 2　法律问句对样例

法律问句		法律类别	法律词汇	对应法条
前提问句	拿车做抵押, 只交了车辆登记证抵押有效吗?	债务债权	抵押、车辆登记证	申请抵押登记的, 机动车所有人应当填写申请表, 由机动车所有人和抵押权人共同申请, 并提交下列证明、凭证: (一)机动车所有人和抵押权人的身份证明; (二)机动车登记证书; (三)机动车所有人和抵押权人依法订立的主合同和抵押合同. 车辆管理所应当自受理之日起一日内, 审查提交的证明、凭证, 在机动车登记证书上签注抵押登记的内容和日期.
假设问句	别人拿着我的车辆登记证书能做抵押吗			机动车所有人将机动车作为抵押物抵押的, 应当向登记地车辆管理所申请抵押登记; 抵押权消灭的, 应当向登记地车辆管理所申请解除抵押登记

为了解决法律领域问句匹配存在的上述问题, 本文提出使用融合法律知识的方法, 从而实现更精准的问句匹配. 具体来说, 本文首先构建一个法律领域知识库, 它包含了两部分: 法律词汇知识库和法条知识库. 具体步骤如下: (1) 首先利用网络爬虫获取法律语料, 然后通过词频统计和人工校对等方法构建了法律领域词典; (2) 爬取所有的法律法条并存入 Elasticsearch 库中, 通过字符级别与词级别混合检索机制匹配法律对应法条. 构建法律知识库后, 本文提出一个基于交叉关注机制的问句匹配模型, 将法律问句与法律领域知识 (法律词汇、法条) 进行融合: 首先将法律问句对通过 BERT 的嵌入编码层, 法条对通过 BiLSTM 层构建相似矩阵, 并用 BERT 的最后一层 Transformer 的输出分别进行法律词汇的自注意力计算、与法条的相似矩阵进行注意力计算, 多个输出融合并完成匹配计算. 实验部分包含了性能对比、消融、实例预测等不同类型的实验. 多个实验结果表明: 融合法律词典和法律法条在测试集上提升了 0.66% 的准确率.

总体来说, 本文的主要贡献包括:

(1) 提出了融合法律知识的问句匹配模型, 其中对法律词汇与法律问句通过词自注意力机制进行融合, 法律法条通过交叉关注机制与法律问句表征进行融合, 提高了问句匹配的效果;

(2) 构建了包含法律词典和法律法条的法律知识库;

(3) 实验验证我们的方法相比于基线模型 BERT, 在验证集提高了 1.92% 的准确率, 测试集上提升了 0.66% 的

准确率.

1 相关工作

1.1 文本匹配

法律问句匹配可以看作为文本匹配任务. 传统的文本匹配研究主要基于人工提取的特征, 因此焦点在于如何设置合适的文本匹配学习算法来学习到最优的匹配模型. Berger 等人[11]提出使用统计机器翻译模型计算网页词和查询词间的“翻译”概率, 从而实现了同义或者近义词之间的匹配映射; Gao 等人[12]在词组一级训练统计机器翻译模型并利用用户点击数据进行模型训; Bai 等人[13]提出有监督学习语义索引模型 (supervised semantic indexing, SSI), Gao 等人[14]扩展了话题模型提出双语话题模型 (bilingual topic model), 对隐空间模型进行概率化建模.

近年来, 随着深度匹配模型的发展迅速, 占据了文本匹配模型的主导地位. 早期的研究探索将每个序列单独编码成一个向量, 然后根据这两个向量建立神经网络分类器. 在这个模式下, Bowman 等人[15], Yu 等人[16], Tan 等人[17]分别采用循环神经网络和卷积神经网络作为序列编码器, 在这些模型中, 一个序列的编码是独立于另一个序列的, 这使得最终的分类器很难建模复杂的关系. 因此, 在后期的工作中, 提出的匹配聚合框架一方面在较低层次匹配两个序列, 另一方面根据注意力机制实现聚合的自动学习. Parikh 等人[18]提出模型可分解注意力机制模型 (decomposable attention model, DecompAtt), 该模型使用一种简单的注意形式进行对齐, 并使用前馈网络聚合对齐表示; Chen 等人[19]提出模型增强序列推理模型 (enhanced sequential inference model, ESIM), 其中使用了类似的注意机制, 但采用了 BiLSTM 作为编码器和聚合器; Chen 等人[20]在 2018 年提出了将 WordNet 作为外部知识库来解决两个问句之间的逻辑关系探究.

为了进一步提高性能, 采用了 3 个主要不同的角度进行改进. 首先是使用更丰富的语法或更加贴合句子特性, IM 等人使用了语法解析树, Tay 等人[21]和 Gong 等人[22]使用 POS 标识符. 第 2 种方法是增加对齐计算的复杂性, Wang 等人[23]提出模型 BIMPM[10], 采用先进的多视角匹配操作; Tan 等人[23]提出模型 MwAN, 应用多个异质性注意函数来计算对准结果. 第 3 种增强模型的方法是对对齐结果建立大量的反复处理层. Tay 等人[24]提出模型紧密堆叠残差模型 (co-stack residual affinity networks, CSRAN), 利用对齐因子分解层从对齐过程中提取额外的信息; Gong 等人[22]提出模型密集交互推理网络 (densely interactive inference network, DIIN), 采用 DenseNet 作为深度卷积特征提取器, 从对齐结果中提取信息. 如果允许重复进行序列间匹配, 则可以建立更有效的模型. Kim 等人[25]提出密集连接递归共关注网络 (densely-connected recurrent and co-attentive network, DRCN), 堆叠编码和对齐层, 它将所有之前对齐的结果串联起来, 并且必须使用自动编码器来处理激增的特征空间; Liu 等人[26]提出随机回答网络 (stochastic answer network, SAN) , 利用递归网络来组合多个对齐结果.

为了挖掘问句的隐藏知识, 不少工作提出在问句匹配中建模主题和关键词. Wu 等人[27]训练了一个线性判别分析模型来预测主题作为先验知识, 并设计了一个知识门来利用主题. Khashabi 等人[28]使用语义图提取关系并执行推理. Yang 等人[29]使用文本序列中的前 K 个单词作为关键词, 形成了注意机制中的关键词掩码. 文本方法与文献 [29] 的方法不同, 本文以 BERT[9]为基础, 加入了对法律关键词的关注层.

1.2 法律领域的自然语言处理

利用向量表征建模法律文本是常用的做法. 对于通用领域的字符级和词级别的嵌入方式, 在法律领域同样有效, 因为它能将离散文本嵌入到连续向量空间中. 文献 [30] 和文献 [31] 提出的方法对于下游任务的有效性至关重要, 因为它们能拟合文本和向量表征之间的差距. 但是并不能直接从一些法律事实描述中学习专业术语的含义. Chalkidis 等人[32]和 Nay[33]主要围绕将现有的嵌入方法 (如 Word2Vec) 应用于领域语料库. 而为了克服学习专业词汇表述的困难, 尝试在词嵌入表征中捕获语法信息和法律知识, 以完成相应的任务. 因为许多模型输出结果应该根据法律规则和知识来决定, 所以知识建模对法律 AI 至关重要.

文献 [34−36] 采用基于注意力的神经网络以及相关法律条款的补充; 一些研究人员认为适用法律条款, 收费, 罚款和处罚条款的决定逻辑关系[37], 而另一些研究人员尝试了一种混合方法来总结法律案件报告[38], 法律条款的

融入能对法律文本的表征进行增强, 对知识的不足进行补充.

2 法律领域知识库的构建

本章节介绍如何构建法律领域知识库. 目前国内已有的一些法律知识库, 通过关键字把大量的法律法条、法律文书与查询内容连接在一起, 在一定程度上提高了法律处理的效率, 而对于法律问句的短文本的法律知识查询, 现有的知识库无法给予正确的知识, 如: 问句中的法律词汇、对应的法律法条. 因此为解决法律问句的知识补给, 本文构建的法律领域知识库将法律知识系统化、形式化的保存, 并基于字符串的检索, 获得准确的法律知识.

2.1 法律领域词典的构建

由于法律领域内容广泛, 种类较多, 因此我们限定法律领域, 将法律问句类型限定于合同纠纷、离婚、交通事故、劳动工伤、债务债权等 5 个类型. 因此构建的法律领域词典是基于上述 5 个类型的法律文本, 并且以清华大学开放中文词库中的法律词汇合集中的法律词集 (以下简称 THUOCL) 作为基础词典, THUOCL (THU Open Chinese Lexicon) 是由清华大学自然语言处理与社会人文计算实验室整理推出的一套高质量的中文词库, 词表来自主流网站的社会标签、搜索热词、输入法词库等, THUOCL 包含词频统计信息 DF (document frequency) 值, 并且词库经过多轮人工筛选, 保证词库收录的准确性. 具体构建流程如图 1 所示.

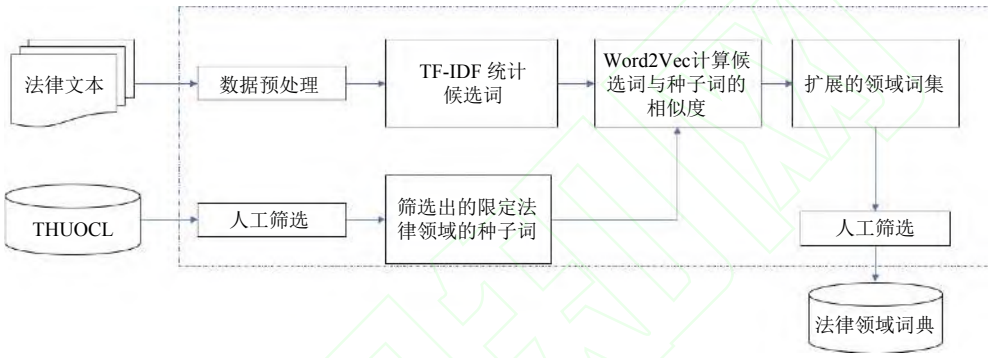

图 1 法律领域词典构建流程图

我们将 THUOCL 作为基础词集, THUOCL 包含了多个法律类别的 9896 个词条, 由于限定了法律类别, 我们人工筛选出包含合同纠纷、离婚、交通事故、劳动工伤、债务债权这 5 个类别的法律词汇, 将其作为种子词集. 首先从各大法律问答平台爬取合同纠纷、离婚、交通事故、劳动工伤、债务债权 5 种法律类别的问答对与对应的法律法条, 对爬取到的法律文本进行去重、去停用词等操作, 这样做能去除通用的词, 减少后续对法律词汇统计的干扰. 我们通过 TF-IDF[39]来对法律文本进行关键词统计. TF-IDF 算法公式如下:

$$TF\text{-}IDF_{w,D_i} = TF_{w,D_i} \times IDF_w \tag{1}$$

公式 (1) 表示关键词 w 在文档 D_i 的 TF-IDF 值. 其中 TF 代表词频, 表示关键词 w 在文档 D_i 中出现的频率.

我们将统计后的词再次进行人工筛选, 作为候选词集. 因为通过上述方法构建的法律词典有很多噪音, 所以需要将候选词集与种子词集进行一次融合, 所以采用 Python 第三方模块 gensim 中的余弦相似度计算函数, 计算每个类别的候选法律词与其对应类别的种子法律词的平均相似度, 最终保留平均相似度大于 0.5 的候选词作为其对应类别的类别领域新词. 计算公式 (5) 如下:

$$\cos\left(v_i, v_j\right) = \frac{v_i \times v_j}{\|v_i\| \times \|v_j\|} \tag{2}$$

其中, v_i 表示候选词的词向量, v_j 表示种子词的词向量. 通过上述方法得到一个扩展的限定法律类别的领域词典, 然后通过人工筛选更符合的扩展词加入种子词集, 进行增量式的迭代进而挖掘更多领域新词. 并将最终的扩展词典和基础词典整合成一个完善的法律领域词典, 得到最后的限定法律类别的法律领域词典, 并将法律领域词典存入到 Elasticsearch 数据库中, 构成词汇知识库.

2.2 法条知识库的构建

法律词汇蕴含的法律知识是不足的, 每一个法律问句根据法律规定会有一个或者多个相匹配的法条, 法律人员依据法条判断法律事实. 如何精准匹配法条, 对于法律专业人士也是存在一定难度, 目前市面上并无法条相关的知识图谱等系统框架, 为了较为准确地实现问句与法条的匹配, 我们通过字符级与词级的交叉匹配来进行法条配对.

首先我们爬取了现有发布的法律法条, 包括《中华人民共和国刑法》《中华人民共和国劳动法》等法规的全部内容, 并以每一条为单位存入到 Elasticsearch 数据库中, 将 Elasticsearch 的检索功能作为查询接口, 并且利用不同粒度的检索方式 (词级检索、字级检索) 对输入问句进行检索. 对于每个问句查询法条知识库, 词级检索与字级检索返回的都在多个法条, 考虑到通过字符串匹配的法条存在一定的误差, 因此对于两个法条合集进行取一致的法条, 并根据 Elasticsearch 库的返回优先度, 选取优先度最高的法条作为问句对应的法条. 如图 2 为法条检索流程图.

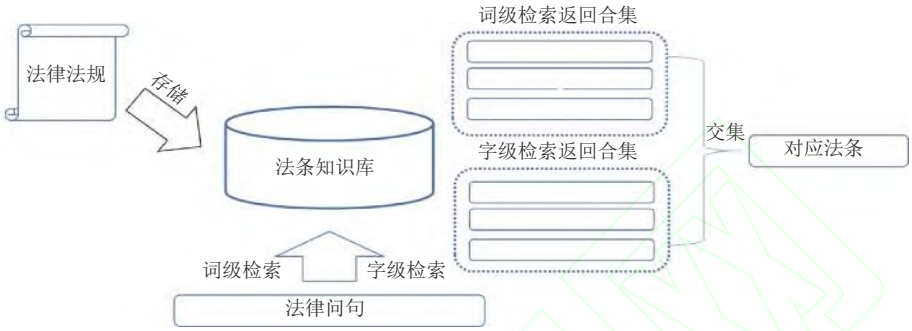

图 2 法条知识库检索流程图

2.3 法律知识库

将查询法律词汇与查询匹配法条合并, 作为法律领域知识库. 输入的法律问句可同时进行法律词汇与法条的查询, 并且可设定阈值来筛选返回法律词的数量. 如图 3 所示, 预设一个问句, 通过法律领域知识库, 可检索出“车辆登记证书”“抵押”等词汇, 判定为法律词; 返回法律词的个数是可以设定的, 如要求返回法律词个数较多的时候, 会返回不在句中的词汇, 因为 Elasticsearch 的字符串检索匹配功能是根据字符级别检索, 排序越后则词的相关性越低; 法律法条为检索出的法条, 能准确的解释有关收养关系的内容. 通过构建的法律领域知识库返回的知识, 可将这些法律知识与法律问句进行融合, 增强问句表征.

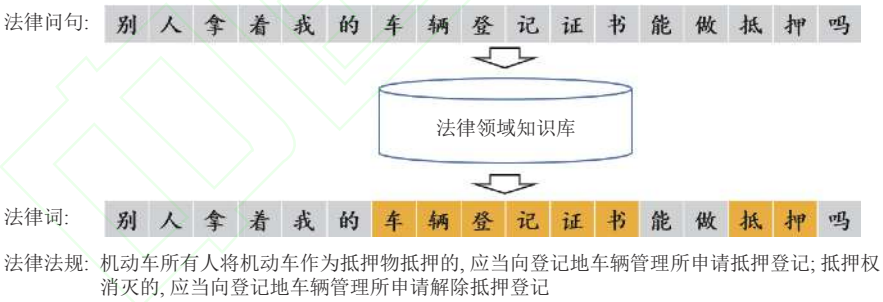

图 3 法律领域知识库查询操作

3 融入法律知识的问句匹配模型

本章节将介绍融入法律知识的问句匹配模型. 图 4 为模型的整体图.

3.1 查询知识库层

如图 4 模型的查询知识层所示, 法律问句对分别请求法律知识库, 获得匹配的法律词汇和法律法条. 将前提问

句和假设问句分别表示为 $x=\{x_1,x_2,\ldots,x_{l_x}\}$, $y=\{y_1,y_2,\ldots,y_{l_y}\}$, 其中 l_x 和 l_y 分别代表前提问句和假设问句的长度, 法条分别为 $a=\{a_1,a_2,\ldots,a_m\}$, $b=\{b_1,b_2,\ldots,b_n\}$, m 和 n 分别表示两个问句对应的法条的长度. 查询得到的法律词合集, 需要进行掩码操作的预处理, 具体操作为: 将句中的法律词表示为 1, 其他的词则表示为 0.

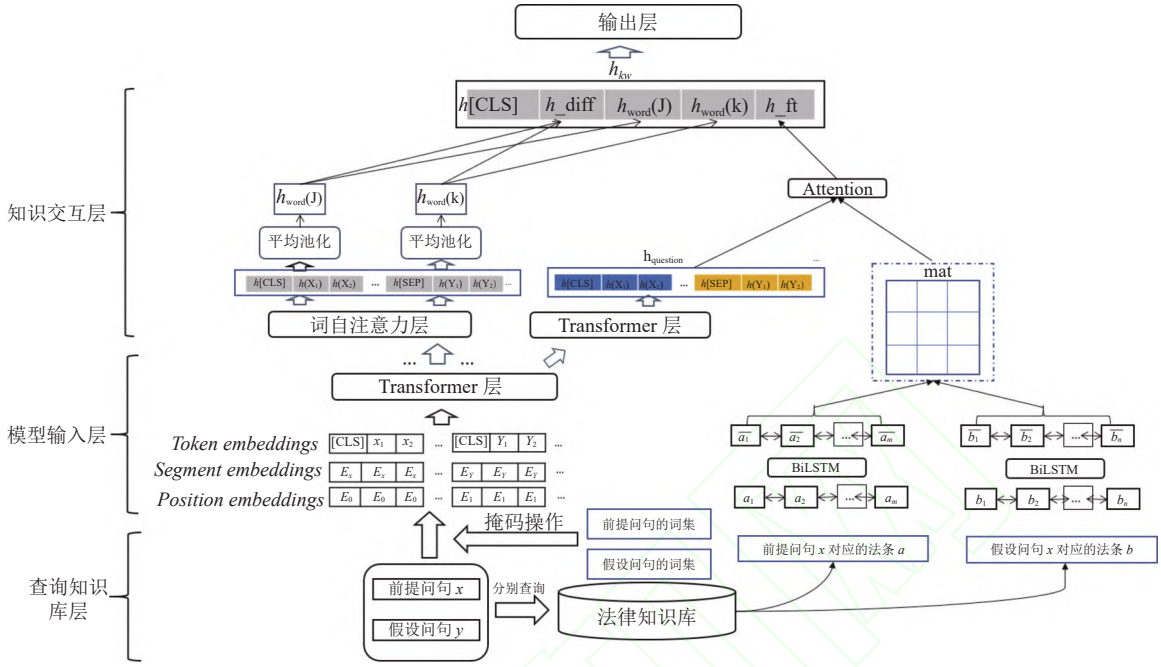

图 4 融入法律知识的问句匹配模型

3.2 模型输入层

对于法律问句输入, 仿照 BERT 在语义相似度任务的做法, 将两个法律问句用字符 [SEP] 衔接, 然后将两个衔接序列同时输入到词嵌入编码层. 其中词嵌入编码分为字符嵌入层 (*TokenEmbeddings*)、段嵌入层 (*Segment-Embeddings*)、位置嵌入层 (*PositionEmbeddings*). 其中字符嵌入层是通过建立字向量表将每个字转换为一个一维向量, 作为模型输入, 输入文本在送入字符嵌入层之前要先进行按字符级分词处理, 且两个特殊的标志 [CLS] 和 [SEP] 会分别插入在文本开头和结尾; 段嵌入层是为了使模型区分拼接的句子对是两个句子; 位置嵌入层是将单词的位置信息编码成特征向量, 位置嵌入是向模型中引入单词位置关系的至关重要的一环. 最后这 3 个向量拼接起来的输入到模型中.

法条编码层为通用的词嵌入编码层, 通过 BiLSTM 层分别学习前提问句对应的法条 a 和假设问句对应的法条 b 中的词与上下文信息, 得到新的向量表示 $\overline{a_i}$、$\overline{b_j}$, 分别代表两个法条经过编码后的隐状态表示:

$$\overline{a_i}=BiLSTM(a,i),\forall i\in[1,\ldots,m] \tag{3}$$

$$\overline{b_j}=BiLSTM(b,j),\forall j\in[1,\ldots,n] \tag{4}$$

构建二维相似性矩阵 mat, 通过矩阵 mat 获得两个不同的法条直接的相关性, 如公式 (7):

$$mat=\overline{a_i}^{\mathrm{T}}\overline{b_j} \tag{5}$$

3.3 知识交互层

对于模型左边为法律问句对的输入学习过程, 法律问句在查询法律领域知识库后获得法律词汇, 然后通过在 BERT 的最后一层并行地堆叠一个附加的法律词汇融入层来融入法律词. 使用注意力机制在问句对的语义匹配任务中是非常重要的做法, 但是由于监督信号不足, 深度模型可能无法准确捕获法律问句的关键信息, 以进行有效的

相似度判别. 为此, 我们提出了词自注意力机制, 词自注意力机制对法律问句给予了对句中的法律词汇更多的关注. 具体如图 5 所示, 前提问句中的每个字只关注假设问句中的法律词 (高亮词), 反之亦然. 词自注意力机制将通过前提问句和假设问句之间的法律词的差异来强化模型, 以了解它们的差别. 通过词自注意力层, 法律字信息被注入到更接近输出目标的位置, 而不是在 BERT 的原始序列输入中. 这种机制可以通过操纵 Transformer 层中的自注意力机制 (self-attention) 的掩码来实现.

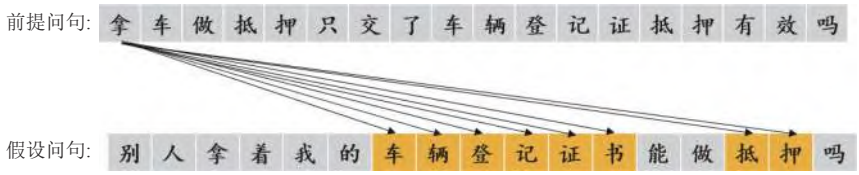

图 5 词自注意力机制

通过词自注意力层后, 将两个句子的输出结果 (排除了 [CLS] 和 [SEP]) 用平均池化进行操作, 形成句子对的两个不同视角的表征 $h_{\text{word}}(J)$ 和 $h_{\text{word}}(K)$, 其中 $h_{\text{word}}(J)$ 是前提问句 x 中的每一个字与假设问句 y 中的法律词, 通过词注意力机制输出的矩阵通过平均池化归结为向量; $h_{\text{word}}(K)$ 则是假设问句 y 中的每一个字与前提问句 x 中的法律词, 使用词注意力机制输出的矩阵通过平均池化归结为向量. 为了抽象化表示这两种表征之间的差异, 我们引入了词差异向量 h_diff, 如下所示:

$$h\_diff = (h_{\text{word}}(J) - h_{\text{word}}(K)) \oplus (h_{\text{word}}(K) - h_{\text{word}}(J)) \tag{6}$$

从 BERT 的最后一层的输出后的两个法律问句对的表征 $h\_question$, 我们将其与已经构建好的法条相似矩阵进行注意力机算, 对法律问句对的表征与法条信息进行交互, 得到表征 h_{ft}, 它能学习到匹配法条的相关性与不匹配法条的差异性, 公式如下所示:

$$h_{ft} = \sum_{n=1}^{m} h_{\text{question}} \times Softmax(mat) \tag{7}$$

最后, 我们将法条对与 Transformer 层的输出经过注意力机制得到的表征 h_{ft}, 并且从原始的 Transformer 层的输出提取 h[CLS], 与法律词自注意力层的输出、词差异向量 h_diff 进行拼接, 用于后续分类, 如下所示:

$$h_{kw} = h[\text{CLS}]\, h_{\text{word}}(J) \oplus h_{\text{word}}(K) \oplus h\_diff \oplus h_{ft} \tag{8}$$

4 实 验

4.1 数据集

法律领域的问句匹配数据集是稀缺的, 而且人工标注成本巨大, 所以我们通过半监督标记的方式构建法律领域的问句匹配数据集. 首先我们对法律咨询平台 (找法网、度小法) 进行数据爬取, 总共爬取了 20 万条问答对, 由于问题的答案为法律网站的专业人士回答, 具有一定的专业性, 因此可将其作为问句的专业解释. 如果两个问句的答案是一致的, 可以认定两个法律问句是指向同一事实, 因此在构成法律问句对的数据时, 本文采用了一种无监督数据挖掘, 并通过人工标注的策略:

(1) 针对 5 个类别的法律问题, 我们对每个类别内的问句之间进行相似度计算, 并设定阈值, 对满足条件的问句进行去重, 从字符级层面上保证语义相似;

(2) 由于第一步筛选出来的问句对通常不是描述同一个法律事实, 无法用统一答案回答, 所以为了保证前提问句与假设问句是在描述同一个法律事实, 我们用问题的答案作为监督信号, 来指导问句的语义相似性构建, 筛选出最终的法律问句, 最后通过人工校对, 确保了问句对在结构与语义的一致.

通过上述方法自动构建训练集, 并人工标注验证集与训练集. 本文数据集包括 5 种法律类别: 合同纠纷 (数据量 5610 条), 离婚 (数据量 3342 条), 交通事故 (数据量 5578 条) 及债务债权 (数据量 3702 条). 训练集为上述方法自动构建, 验证集与测试集均为人工标注, 正负比例均趋于 1:1, 其中训练集 21502 条, 验证集 300 条, 测试集 500 条, 且包含的法律类别均衡.

4.2 实验设置

ESIM、RE2、BIMPM 为通用的深度匹配模型, 参数设置一致. 词向量维度为 100 维, 序列最大长度设置为 50, 训练轮次设置为 100, 学习率设置为 0.000 5, Dropout 设置为 0.2, 提前截至的容忍度设置为 5, 连续 5 个训练轮次的验证集损失没有下降, 则自动中止程序. 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. 本文模型以 BERT 为基线模型进行改进, 将法律词的个数设置为 2, 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, 法条的最大序列长度也设置为 64.

4.3 基准模型

BIMPM[10]: 采用了一种先进的多视角匹配的思想, 融合最大匹配、最大池化匹配、注意力匹配、最大注意力匹配等视角来进行语义相似度机算.

RE2[40]: 一种可插拔的匹配模型框架, 包含了原始点对齐特性、先前对齐特性和上下文特性等 3 个特性, 同时简化所有剩余组件. 参数少, 速度快.

ESIM[19]: 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或手工设计的对齐特性作为对齐层的额外输入.

BERT[9]: 预训练语言模型, 本文将 BERT 作为文本特征提取器来学习法律文本的特征.

4.4 实验分析

4.4.1 模型有效性实验

为了验证本文模型的有效性, 我们通过将本文模型与 4 个基准模型在自行构建的法律领域问句的数据集上进行了对比实验, 具体实验结果如表 3 所示.

表 3 有效性实验结果 (%)

模型	Acc (dev)	Rec (dev)	Pre (dev)	Acc (test)	Rec (test)	Pre (test)
ESIM	88.83	87.65	86.42	87.32	86.23	86.98
BiMPM	87.04	86.21	85.12	86.86	85.58	85.87
RE2	88.39	87.12	86.03	86.00	85.12	85.49
BERT	91.40	90.56	90.85	90.61	89.52	90.23
本文模型	**93.32**	92.38	92.45	**91.27**	90.94	91.02

根据表 3 实验结果看出: (1) 本文模型在验证集和测试集的各项指标上均达到了最优, 在验证集上本文模型相比 BERT 在 Acc 提升 1.92%、Rec 提升 1.82%、Pre 提升了 1.60%, 在测试集上本文模型相比 BERT 在 Acc 提升 0.66%、Rec 提升 1.42%、Pre 提升了 0.79%. 与传统的 BERT 相比, 本文模型具有更强的法律问句匹配的能力, 通过加入词自注意力层能显示指示句中法律词的位置, 增加模型识别法律词位置的能力, 使模型在句间关系判断中可以识别到更多细粒度的特征, 提高模型的领域适配能力; 又通过融入问句对应的法条, 让问句对的表征与法条的相似矩阵进行注意力机制交互, 能让问句既能学习到匹配法条与问句的相似之处, 增强表征能力, 也可以让问句与不相匹配的法条进行差异性放大, 尽管匹配法条的规则是依赖于字符串匹配, 但从字面层面上能保证有多个相似的词是与问句相同, 能有效保证法律问句的表征不稀疏, 且提升模型的参数量. (2) BERT 的实验效果展现出了预训练语言模型的优越性, 相比于其他的深度匹配基线模型, 在法律问句匹配的任务上效果更好, 是一个更强的基线模型. (3) RE2 作为最新的交互文本匹配模型, 相比于 ESIM, 并未表现的更好, 但是模型的参数量小, 训练速度更快. 除了预训练模型的使用以外, 大幅度融入外部知识来增强短文本表征是一个有效、实用的方法.

4.4.2 模型模块消融实验

本章节为考究各个部分是否对问句匹配的效果都有提升, 因此我们将 BERT, BERT+词自注意力机制, BERT+词自注意力机制+法条知识 3 个模型在构建的法律问句对的数据集上进行对比实验, 实验结果如表 4 所示.

从表 4 看出, 相比于原始的 BERT, BERT+词自注意力机制在验证集上的评测结果显示, 在验证集上 BERT+词自注意力机制模型相比 BERT 在 Acc 提升 1.50%、Rec 提升 0.89%、Pre 提升了 0.38%, 在测试集上 BERT+词

自注意力机制模型相比 BERT 在 Acc 提升 0.61%、Rec 提升 0.85%、Pre 提升了 0.80%, 添加法律词汇信息能提高 BERT 对法律问句对的表征能力; BERT+词自注意力机制在验证集上的评测结果显示, 在测试集上 BERT+词自注意力机制模型相比 BERT 在 Acc 提升 0.28%、Rec 提升 0.26%、Pre 提升了 0.33%, 融入法条知识提升效果并不明显, 因为在匹配法条的规则是根据字符串匹配, 有可能导致不正确的知识的引入, 而效果有所提升应该是由于模型的参数得到了增加, 后续可通过建模正确的匹配机制提高效果. 添加法律词汇信息能提高 BERT 对法律问句对的表征能力 BERT+词自注意力机制+法条知识模型相比于 BERT+词自注意力机制在验证集上 Acc 提升 0.42%、Rec 上提升 0.93%、Pre 上提升了 1.22%, 在测试集上 Acc 提升 0.05%、Rec 提升 0.57%、Pre 提升了 0.43%. 添加了法条对的输入能增强法律问句的背景知识, 建立了问句之间的法律关系, 模型在加入法条知识后参数量也得到了提升, 且法律问句对应的法条能从语义中增强法律问句的知识表征.

表 4 各模块的对比实验 (%)

模型	Acc (dev)	Rec (dev)	Pre (dev)	Acc (test)	Rec (test)	Pre (test)
BERT	91.40	90.56	90.85	90.61	89.52	90.23
BERT+词	92.90	91.45	91.23	91.22	90.37	91.03
BERT+法条	91.56	91.03	90.72	90.89	89.78	90.56
BERT+词+法条	**93.32**	92.38	92.45	**91.27**	90.94	91.46

4.4.3 法律词汇合集个数的探究实验

法律词汇的个数也是影响实验效果的一大因素, 尽管在实验的过程中将句子中的法律词个数设置为 2, 但是否增大法律词汇数量对实验效果有提升? 因此我们对法律问句中的法律词的个数进行了探究, 采用本文 BERT+WORD 的模型设置不同的词汇的个数进行实验, 实验效果如表 5 所示.

表 5 词个数对比实验 (%)

法律词数量 (n)	Acc (dev)	Rec (dev)	Pre (dev)
0	91.40	90.56	90.85
1	92.13	91.45	91.98
2	**92.90**	**91.45**	**91.23**
3	92.42	91.65	91.19

分析表 5, 当法律词的个数为 2 效果最好. 因为法律问句的文本长度较短, 蕴含的法律词汇有限, 检索法律领域知识库的词是根据相关度返回, 当返回的词的个数上升时, 返回的词的相关性则越低, 反而会下降句子对真正的法律词的关注度.

4.4.4 法律问句测试实例分析

为了具体说明本文提出的方法在法律问句匹配任务中为何有效性, 本文选取了测试集中的一个问题对进行模型预测后, 展示出预测得分. 首先将法律问句查询法律知识库, 获取相关法律知识 (法律词汇、法律法条), 表 6 为两个测试问句中的法律词和匹配的法条.

表 6 测试问句查询法律知识库的结果

法律问句	法律词	法条
在试用期5个多月, 公司还没有**签合同**, 主动**辞职**可以要求赔双倍工资吗?	签合同、辞职	劳动合同期限三个月以上不满一年的, 试用期不得超过一个月; 劳动合同期限一年以上不满三年的, 试用期不得超过二个月; 三年以上固定期限和无固定期限的劳动合同, 试用期不得超过六个月. 同一用人单位与同一劳动者只能约定一次试用期. 以完成一定工作任务为期限的劳动合同或者劳动合同期限不满三个月的, 不得约定试用期. 试用期包含在劳动合同期限内. 劳动合同仅约定试用期的, 试用期不成立, 该期限为劳动合同期限.
试用期没有**签合同**, 2个月后我主动**离职**可以要求赔双倍工资吗?	签合同、离职	

表 6 为测试问句分别查询法律知识库的返回结果, 可以发现两个问句中都含有“签合同”这个词, 并且在返回的法律词中有出现, 且两个问句匹配的法条是同一条, 证明两个法律问句蕴含的法律知识是一致的. 将测试问句分

别输入到 ESIM、BERT 和本文模型中, 通过获得 Acc (准确率) 的得分来判定模型对两个问句是否匹配的效果, 在已知两个问句的标签的情况下, 得分越高, 则代表模型预测的更准确. 表 7 为 3 个不同的模型对该问句对的评分:

表 7　不同模型预测结果

模型	预测结果
ESIM	0.821
BERT	0.876
本文模型	**0.901**

分析表 7, 3 个模型预测值都超过了 0.5, 我们采样的法律问句对在数据集中的标签是匹配的, 因此证明在判断测试问句是否匹配, 模型均判断正确, 且本文模型相比于 BERT 提高了 0.025. 因为通过查询法律知识库, 在融入法律词汇与法条知识后, 增强了两个问句的表征, 词自注意力机制能提高问句中法律词汇的权重, 而法条知识的融入则更好的解释了法律问句的语义, 同时也增加了模型参数量.

5 结束语

本文针对法律问句匹配任务, 提出了一种融合法律知识 (法律词汇、法律法条) 增强问句表征的问句匹配方法, 通过查询构建的法律知识库获得的法律词和法律法条, 通过交叉关注模型同时建模问句、法律词汇、法律法条三者之间的关联, 有效地增强了问句表征. 相比于基线模型, 对于法律领域的问句匹配任务有一定的提升效果. 在下一步的工作中, 拟更多地去探索法律问句之间的关系, 尝试挖掘更多的潜在关系, 包括融入问句答案信息或者拓展其他法律领域任务中.

References:

[1] Buscaldi D, Rosso P, Gómez-Soriano JM, Sanchis E. Answering questions with an *n*-gram based passage retrieval engine. Journal of Intelligent Information Systems, 2010, 34(2): 113–134. [doi: 10.1007/s10844-009-0082-y]

[2] Taniguchi R, Kano Y. Legal yes/no question answering system using case-role analysis. In: Proc. of the JSAI Int'l Symp. on Artificial Intelligence. Kanagawa: Springer, 2017. 284–298. [doi: 10.1007/978-3-319-61572-1_19]

[3] Fawei B, Wyner A, Pan JZ, Kollingbaum M. Using legal ontologies with rules for legal textual entailment. In: Proc. of the AICOL Int'l Workshops on AI Approaches to the Complexity of Legal Systems. Springer, 2018. 317–324. [doi: 10.1007/978-3-030-00178-0_21]]

[4] Xian J, Mo XL, Xi JQ. A question answering system based on question pattern match. Journal of Shandong University, 2006, 41(3): 99–103 (in Chinese with English abstract). [doi: 10.3969/j.issn.1671-9352.2006.03.024]

[5] Huang PS, He XD, Gao JF, Deng L, Acero A, Heck L. Learning deep structured semantic models for web search using clickthrough data. In: Proc. of the 22nd ACM Int'l Conf. on Information & Knowledge Management. San Francisco: Association for Computing Machinery, 2013. 2333–2338. [doi: 10.1145/2505515.2505665]

[6] Shen YL, He XD, Gao JF, Deng L, Mesnil G. A latent semantic model with convolutional-pooling structure for information retrieval. In: Proc. of the 23rd ACM Int'l Conf. on Conf. on Information and Knowledge Management. Shanghai: Association for Computing Machinery, 2014. 101–110. [doi: 10.1145/2661829.2661935]

[7] Severyn A, Moschitti A. Learning to rank short text pairs with convolutional deep neural networks. In: Proc. of the 38th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Santiago: Association for Computing Machinery, 2015. 373–382. [doi: 10.1145/2766462.2767738]

[8] Yin WP, Schütze H, Xiang B, Zhou BW. ABCNN: Attention-based convolutional neural network for modeling sentence pairs. Trans. of the Association for Computational Linguistics, 2016, 4: 259–272. [doi: 10.1162/tacl_a_00097]

[9] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis: Association for Computational Linguistics, 2019. 4171–4186. [doi: 10.18653/v1/N19-1423]

[10] Wang ZG, Hamza W, Florian R. Bilateral multi-perspective matching for natural language sentences. In: Proc. of the 26th Int'l Joint Conf. on Artificial Intelligence. Melbourne: IJCAI. org, 2017. 4144–4150. [doi: 10.24963/ijcai.2017/579]

[11] Berger A, Lafferty J. Information retrieval as statistical translation. ACM SIGIR Forum, 2017, 51(2): 219–226. [doi: 10.1145/3130348.3130371]

[12] Gao JF, He XD, Nie JY. Clickthrough-based translation models for web search: From word models to phrase models. In: Proc. of the 19th ACM Int'l Conf. on Information and Knowledge Management. Toronto: Association for Computing Machinery, 2010. 1139–1148. [doi: 10.1145/1871437.1871582]

[13] Bai B, Weston J, Collobert R, Grangier D. Supervised semantic indexing. In: The 31th European Conf. on Advances in Information Retrieval. Toulouse: Springer, 2009. 761–765. [doi: 10.1007/978-3-642-00958-7_81]

[14] Gao JF, Toutanova K, Yih WT. Clickthrough-based latent semantic models for web search. In: Proc. of the 34th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Beijing: Association for Computing Machinery, 2011. 675–684. [doi: 10.1145/2009916.2010007]

[15] Bowman SR, Angeli G, Potts C, Manning CD. A large annotated corpus for learning natural language inference. In: Proc. of the 2015 Conf. on Empirical Methods in Natural Language Processing. Lisbon: Association for Computational Linguistics, 2015. 632–642. [doi: 10.18653/v1/D15-1075]

[16] Yu L, Hermann KM, Blunsom P, Pulman S. Deep learning for answer sentence selection. arXiv: 1412.1632, 2014.

[17] Tan M, dos Santos C, Xiang B, Zhou BW. Improved representation learning for question answer matching. In: Proc. of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin: Association for Computational Linguistics, 2016. 464–473. [doi: 10.18653/v1/P16-1044]

[18] Parikh A, Täckström O, Das D, Uszkoreit J. A decomposable attention model for natural language inference. In: Proc. of the 2016 Conf. on Empirical Methods in Natural Language Processing. Austin: Association for Computational Linguistics, 2016. 2249–2255. [doi: 10.18653/v1/D16-1244]

[19] Chen Q, Zhu XD, Ling ZH, Wei S, Jiang H, Inkpen D. Enhanced LSTM for natural language inference. In: Proc. of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver: Association for Computational Linguistics, 2017. 1657–1668. [doi: 10.18653/v1/P17-1152]

[20] Chen Q, Zhu XD, Ling ZH, Inkpen D, Wei S. Neural natural language inference models enhanced with external knowledge. In: Proc. of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne: Association for Computational Linguistics, 2018. 2406–2417. [doi: 10.18653/v1/P18-1224]

[21] Tay Y, Luu AT, Hui SC. Compare, compress and propagate: Enhancing neural architectures with alignment factorization for natural language inference. In: Proc. of the 2018 Conf. on Empirical Methods in Natural Language Processing. Brussels: Association for Computational Linguistics, 2018. 1565–1575. [doi: 10.18653/v1/D18-1185]

[22] Gong YC, Luo H, Zhang J. Natural language inference over interaction space. In: Proc. of the 6th Int 'l Conf. on Learning Representations. Vancouver: OpenReview.net, 2018.

[23] Tan CQ, Wei FR, Wang WH, Lv WF, Zhou M. Multiway attention networks for modeling sentence pairs. In: Proc. of the 27th Int'l Joint Conf. on Artificial Intelligence. Stockholm: IJCAI. org, 2018. 4411–4417. [doi: 10.24963/ijcai.2018/613]

[24] Tay Y, Luu AT, Hui SC. Co-stack residual affinity networks with multi-level attention refinement for matching text sequences. In: Proc. of the 2018 Conf. on Empirical Methods in Natural Language Processing. Brussels: Association for Computational Linguistics, 2018. 4492–4502. [doi: 10.18653/v1/D18-1479]

[25] Kim S, Kang I, Kwak N. Semantic sentence matching with densely-connected recurrent and co-attentive information. In: Proc. of the 33rd AAAI Conf. on Artificial Intelligence. Honolulu: AAAI, 2019. 6586–6593. [doi: 10.1609/aaai.v33i01.33016586]

[26] Liu XD, Duh K, Gao JF. Stochastic answer networks for natural language inference. arXiv: 1804.07888, 2018.

[27] Wu Y, Wu W, Xu C, Li ZJ. Knowledge enhanced hybrid neural network for text matching. In: Proc. of the 32nd AAAI Conf. on Artificial Intelligence. New Orleans: AAAI, 2018. 5586–5593.

[28] Khashabi D, Khot T, Sabharwal A, Roth D. Question answering as global reasoning over semantic abstractions. In: Proc. of the 32nd AAAI Conf. on Artificial Intelligence. New Orleans: AAAI, 2018. 1905–1914.

[29] Yang JX, Rong WG, Shi LB, Xiong Z. Sequential attention with keyword mask model for community-based question answering. In: Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis: Association for Computational Linguistics, 2019. 2201–2211. [doi: 10.18653/v1/N19-1228]

[30] Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. In: Proc. of the 1st Int'l Conf. on Learning Representations. Scottsdale, 2013.

[31] Joulin A, Grave E, Bojanowski P, Douze M, Jégou H, Mikolov T. FastText. zip: Compressing text classification models. arXiv:

1612.03651, 2016.

[32] Chalkidis I, Kampas D. Deep learning in law: Early adaptation and legal word embeddings trained on large corpora. Artificial Intelligence and Law, 2019, 27(2): 171–198. [doi: 10.1007/s10506-018-9238-9]

[33] Nay JJ. Gov2Vec: Learning distributed representations of institutions and their legal text. In: Proc. of the 1st Workshop on NLP and Computational Social Science. Austin: Association for Computational Linguistics, 2016. 49–54. [doi: 10.18653/v1/W16-5607]

[34] Luo BF, Feng YS, Xu JB, Xu JB, Zhang X, Zhao DY. Learning to predict charges for criminal cases with legal basis. In: Proc. of the 2017 Conf. on Empirical Methods in Natural Language Processing. Copenhagen: Association for Computational Linguistics, 2017. 2727–2736. [doi: 10.18653/v1/D17-1289]

[35] Wu L, Wang Y, Gao JB, Li X. Where-and-when to look: Deep siamese attention networks for video-based person re-identification. IEEE Trans. on Multimedia, 2019, 21(6): 1412–1424. [doi: 10.1109/TMM.2018.2877886]

[36] Wu L, Wang Y, Li X, Gao JB. Deep attention-based spatially recursive networks for fine-grained visual recognition. IEEE Trans. on Cybernetics, 2019, 49(5): 1791–1802. [doi: 10.1109/TCYB.2018.2813971]

[37] Zhong HX, Guo ZP, Tu CC, Xiao CJ, Liu ZY, Sun MS. Legal judgment prediction via topological learning. In: Proc. of the 2018 Conf. on Empirical Methods in Natural Language Processing. Brussels: Association for Computational Linguistics, 2018. 3540–3549. [doi: 10.18653/v1/D18-1390]

[38] Galgani F, Compton P, Hoffmann A. Combining different summarization techniques for legal text. In: Proc. of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data. Avignon: Association for Computational Linguistics, 2012. 115–123.

[39] Seki Y. Sentence extraction by tf/idf and position weighting from newspaper. In: Proc. of the 3rd NTCIR Workshop on Research in Information Retrieval. Tokyo: National Institute of Informatics, 2002.

[40] Yang RQ, Zhang JH, Gao X, Ji F, Chen HQ. Simple and effective text matching with richer alignment features. In: Proc. of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019. 4699–4709. [doi: 10.18653/v1/P19-1465]

附中文参考文献:

[4] 冼健, 莫玄朗, 奚建清. 基于问题模式匹配的智能答疑系统原型. 山东大学学报(理学版), 2006, 41(3): 99–103. [doi: 10.3969/j.issn.1671-9352.2006.03.024]

刘权(1995－), 男, 硕士, 主要研究领域为自然语言处理, 文本匹配.

刘康(1981－), 男, 博士, 研究员, CCF 高级会员, 主要研究领域为自然语言处理.

余正涛(1970－), 男, 博士, 教授, CCF 高级会员, 主要研究领域为自然语言处理, 机器翻译, 信息检索.

高盛祥(1977－), 女, 博士, 副教授, CCF 专业会员, 主要研究领域为自然语言处理, 信息检索, 机器翻译.

何世柱(1987－), 男, 博士, 副研究员, CCF 专业会员, 主要研究领域为问答系统, 知识推理, 知识图谱, 自然语言处理.