

北京航空航天大学计算机学院

硕士研究生学位论文 开题报告

论文题目：基于法律条文的判案论点挖掘

专 业：计算机科学与技术

研究方向：自然语言处理

研 究 生：罗希月

学 号：ZY2106344

指导教师：巢文涵、王丽宏

北京航空航天大学计算机学院

2022 年 11 月 23 日

目 录

1 论文研究背景与意义 1

1.1 论文选题背景 1

1.1.1 实际应用背景 1

1.1.2 学术界理论研究背景 1

1.2 研究现状概述 2

1.2.1 司法智能化的研究现状 2

1.2.2 案情关键信息识别相关研究现状 3

1.2.3 司法证据相关研究现状 4

1.3 研究目标与创新性 5

2 研究内容与技术路线 5

2.1 研究内容一：法律条文细粒度解耦 6

2.2 研究内容二：基于法律条文的案情关键信息识别与关联 6

2.3 研究内容三：基于法律条文的司法证据预测与生成 7

2.4 基于注意力机制和强化学习的判案论点挖掘 8

2.5 基于文本聚类的司法证据关联和预测 9

2.6 基于序列到序列模型的司法证据生成 9

3 论文工作安排计划 10

3.1 工作进度安排 10

3.2 论文工作基础 10

3.2.1 数据集 10

3.2.2 调研工作 11

3.2.3 现有实验结果 11

3.3 可能遇到的问题以及解决途径 11

主要参考文献 13

图 目

图 1 司法智能化主要任务示例 2

图 2 事件信息标注样例 4

图 3 实际案件中的证据描述的两种形式 5

图 4 研究内容关系示意图 6

图 5 法律条文细粒度解耦示意图 6

图 6 法律条文指导案情关键信息识别 7

图 7 司法证据预测与生成任务说明 7

图 8 基于注意力机制的判案论点挖掘 8

图 9 序列到序列模型 9

目 录

表 1 工作进度安排 10

1 论文研究背景与意义

1.1 论文选题背景

1.1.1 实际应用背景

随着我国法制制度建设的不断完善，居民法律意识也日渐增强，对司法系统的依赖程度也越来越高，这就导致案件数量激增，大量的刑事或民事案件流入司法系统，这种情况对司法系统的办案效率提出了更高的要求。

据程金华^[1]等人的研究调查显示：近年来，中国法院的诉讼案件量持续上升，先后几任最高院院长在每年全国“两会”工作报告中，都明确提到法院系统“案多人少”和法官工作负荷重的问题。改革开放以来，中国法院的诉讼案件量一直在高速增长。1993年，全国法院刑事、民事、行政和执行案件的立案总量为460多万件，到2019年增加到2652万余件，在不到30年的时间里，增加了4.75倍。一起案件的审理需要司法工作人员在仔细阅读案情描述的基础上参考大量诸如法律条文、涉案证据、历史案件等资料，并依托自身专业知识进行涉案证据的核实与整理，最终给出涉案人员罪名裁定，刑期的判决，以及给出最终的法院观点。不难看出，司法人员无论是在体能、工作压力，还是在业务能力、办案效率等方面正面临着越来越严峻的挑战。

我国法律体系是成文法系，其法律以成文法的方式存在，这也就意味着我们在进行定罪的时候，必须找到能够证明相关法条的证据。证据在司法案件定罪和量刑中都起到了极其重要的作用，目前我国法律文书中，大部分文书将同一个案件所涉及到的所有证据仅仅进行简单的罗列，这一做法导致证据的论证性以及不同证据之间的关联性缺失。因此，对司法案件中涉及到的证据进行合理的组织不仅能够提升定罪的推理逻辑性与完备性，也使得判决结果更具可解释性^[2]。

1.1.2 学术界理论研究背景

此前的工作提出了多个功能多样的法律辅助系统，如给定案情描述搜索相关案件^[3]，预测法律判决^[4]等。尽管这方面的研究成果丰富多样，但是依旧存在两个问题，首先，当前最先进的(SOTA) LJP模型面临的技术问题之一是它们未能定位能确定判断结果的关键信息^[5-6]。其次，近年的研究忽略了来对刑事案件司法证据的研究。司法证据的作用是论证用于定罪的若干子论点，而证据说明是刑事判决书的重要组成部分。

近年来，一些论文也关注到了司法决策模型没有定位到关键信息，导致缺乏可解释性的问题。但是这些论文，并没有从法律条文出发对案情描述关键信息的识别进行针对性的预测和评估。论文^[7]虽然通过关键信息进行下游任务预测，但是没有给出对案情描述关键信息的预测结果进行量化评估。论文^[8]虽然标注了一部分的案情描述关键信息位置，并且给出了定位关键信息如何影响司法预测任务

的结果，但是这部分的关键信息标注并非在法条指导下的，而是按照事件模式对案情描述进行的标注，其标注内容无法与法律条文和司法论点相匹配，可解释性和可扩展性都不足。

论文^[9]首次研究了司法证据相关的预测，并构造了一个司法证据的数据集，但其提出的是一种论证驱动的监督学习方法来计算证据对之间的距离，并用于下一步的证据关联，证据关联是根据相应的司法论点将一组司法证据划分为几个不重叠的子集，从而提高了论点可解释性和合法性，但该任务是关联给定的已有证据，缺乏对给定案情描述的直接证据预测。据本工作调研所知，学术界关于法律领域的证据预测的研究非常有限。

1.2 研究现状概述

1.2.1 司法智能化的研究现状

司法智能化主要是指人工智能与司法的结合，让人工智能技术能够辅助法律从业人员完成相应的工作。如图 1 所示，目前主要的研究方向有罪名预测、法条推荐、刑期预测、法院观点生成等^[10-11]。

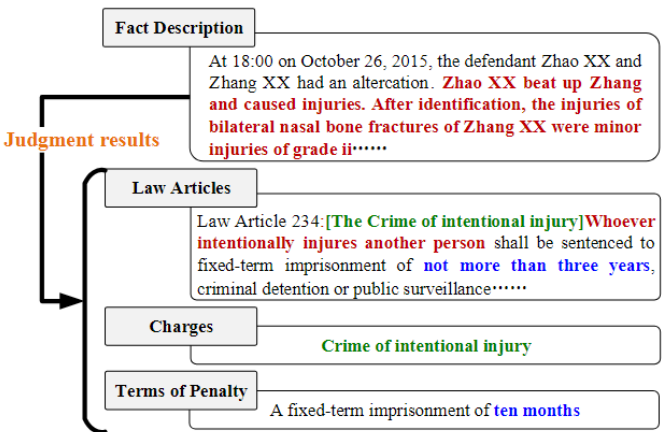


图 1 司法智能化主要任务示例

罪名预测指的是根据案件的事实描述来预测最终的判决罪名，属于法律判决预测的子任务，在智能司法中起着重要的作用，可以为法律人士提供方便的参考以提升工作效率，也可以为不熟悉法律术语以及复杂判决流程的普通民众提供法律咨询工作。

大多数的罪名预测工作是以文本分类的方法去实现。在早期，研究学者侧重于从事实描述等文本来提取有效的特征，例如 Lin 等人^[12]通过人工设计罪名的相关规则来实现罪名预测，但是这种方法需要耗费大量人力、精力去设计特征规则，并且对于设计者的法律专业素养有着较高的要求，无法推广到其他场景中。因此，之后的学者开始使用发展迅速的深度神经网络来建模法律文档，例如 Luo^[13]采用基于注意力的神经网络，对法律条文提取和罪名预测进行联合建模，通过整合法

律条文信息来实现更准确的罪名预测效果。随后, Hu 等人^[14] 考虑到司法领域中不同罪名间存在的数据不平衡问题, 一些不常见罪名涵盖的案件数量极少, 并且以往很多研究都是侧重于常见罪名的预测, 因此他们引入了 10 种额外的属性特征来辅助预测最终罪名。

而 Liu 等^[15] 通过一组固定的法条组合将多标签问题转换为多类别的分类问题, 之后 Liu^[16] 首先通过支持向量机先进行法条的粗略分类, 获得一部分可能的罪名后再通过一些基于特征的重排序来确定最终的判决罪名。Zhong 等人^[5] 和 Yang 等人^[17] 都是采取一个全局范围的思路, 模拟实际法官的行为, 综合考虑法律案件描述、适用法条、罪名预测、量刑预测等多个模块, 将几个模块串联起来共同学习训练, 从而实现较为准确的量刑参考。

针对法院观点生成, Jiang 等人^[4] 针对这一需求提出一种在给定罪名条件下自动化生成判决文书的 Seq2Seq 模型。另外, 除了面向判案流程的研究以外, 法律领域的知识问答系统也是一个备受关注的课题。这种问答系统可以为用户提供便捷的咨询服务。Wu 等人^[18] 提出了一种基于注意力和反事实的自然语言生成方法, 该方法使用注意编码器利用法条和案情描述作为输入来学习法条感知编码器, 从该编码器可以强调案情描述中的法条相关信息。反事实解码器用于消除数据中的混淆偏差, 并通过结合协同预测模型生成。Kim 和 Xu 等人^[19] 通过合并法律领域知识和文本细节特征完成对问题答案的检索。Danilo 等人^[20] 则通过融合信息抽取和知识问答提供了一套案情分析系统。

Xiao 等人^[21] 基于 Longformer 的预训练语言模型, 提出了中国法律长文件预训练语言模型, 称为 Lawformer, 用于中国法律长文档理解。Lawformer 在各种智能司法任务上进行了评估, 包括判断预测、类似案例检索、法律阅读理解和法律问答。实验结果表明, Lawformer 可以在以长文档作为输入的任务上实现部分的改进。

1.2.2 案情关键信息识别相关研究现状

由于单篇案情描述文本长度过长, 通常能达到 300-400 字的篇幅, 所以, 引导模型关注到案情描述中的关键信息, 而忽略掉一些影响较小的文本, 能够大幅提高与案情描述相关的智能司法任务的性能。一部分的论文工作注意到了这一点, 从而有了对案情关键信息识别工作。

Jiang 等人^[7] 通过从案情描述中提取出能决定罪名预测的文本片段来增加罪名预测的可解释性。这些语义片段是在罪名预测过程中提取的, 对预测结果具有决定性的影响。因此, 它们可以看作是罪名预测的解释, 提高了罪名预测的可解释性。这篇工作认为从案情描述中抽取出的好的片段必须满足三个标准: 1) 总长度应该很小; 2) 能够决定罪名; 3) 表达完整的语义。但是这篇工作中语义片段的抽取完全靠文本的注意力机制, 并没有直接的监督数据。如果存在一些情况, 法条文本和案情描述的语义上差别较大, 则无法进行抽取。

Feng 等人^[8] 则进行了语义匹配标注。他们利用从案情描述中提取事件信息的方法提取来解决 LJP 任务。如图 2 所示，他们首先定义了一个层次事件结构，并收集了一个新的带有事件注释的智能司法任务数据集。之后，提出了一个模型，该模型根据两种约束来联合学习智能司法任务和事件提取任务。虽然该模型识别出并由于判断的案情事件信息具有一定的司法原理，但是，事件的各项属性与法律条文中强调的判案维度并没有直接联系。所以，使用这一些信息来理解案情信息，是不全面且缺少司法可解释性的。

	Argument	Role
Who is the criminal?	Mike	Criminal
Who is the victim	Jessica	Victim
What happened?	robbed	Trigger-Rob
What were robbed?	gold ring	Property
What is the price of swag?	1,535 RMB	Quantity
Judgment Results: Article 263, Robbery, three-year imprisonment		

图 2 事件信息标注样例

1.2.3 司法证据相关研究现状

Teng 等人针对司法证据研究了司法证据关联相关工作。刑事案件的司法证据关联将一组司法证据划分为几个不重叠的子集，提高了司法论点的可解释性和合法性。证据关联任务是由先前关于法律辅助系统的研究推动的，特别是提高罪名预测等任务的可解释性。值得注意的是，划分为相同子集的证据通常用于论证同一个司法论点。他们提出了一种论证驱动的监督学习方法来计算案件给定证据对之间的距离，之后根据证据对之间的举例来进行司法证据关联。除此之外，他们还构建了一个真实的刑事案件证据数据集进行实验论证结果的有效性。

司法证据的作用是论证用于定罪若干子论点，而证据说明是刑事判决书的重要组成部分。司法判决书中的证据与论点等都是重要的司法元素。因为在法律判决书中，法官在撰写判决书时需要立足于事实、证据和法律，围绕案件的争议焦点、裁判结论和推理过程^[22]。如图 3 所示，一般刑事判决书中证据部分通常为以下两种表述方法，左侧利用抽象、笼统的说法或者简单罗列的方法，代替对证据的分析、论证，而右侧符合规范的证据说理形式列出了证据与对应分论点的组织结构，具有更好的可读性，然而此类规范形式在形式判决书中仅占约 5%。

Teng 等人主要使用 BERT^[23] 和 ESIM^[24] 模型来学习证据对之间的距离度量。并且发现监督方法能显著提高证据聚类结果。通过引入来自显式论点的信息，能够大大提高聚类结果。但是论文工作存在一定的问题，因为在大多数实际情况案件没有给出明确的论点，所以需要研究如何通过案件描述对判案论点进行建模，以利用显式判案论点对证据聚类进行改进。

Collection Form

证明上述事实的证据有：
接受刑事案件登记表、立案决定书、归案情况说明、
在逃人员登记表、死亡证明、户籍信息、机动车驾驶
证、行驶证、交通事故认定书、民事调解书、谅解书
，车辆检验报告、尸体检验报告，现场勘验检查笔录
，证人宋某某的证言及被告人李宇的供述和辩解。

The above facts can be proved by the following evidence:
criminal case registration form, the decision to file a
case, fugitive registration form, the victim's death certificate,
defendant's household registration information, motor
vehicle driving license, the traffic accident responsibility
certificate, paper of civil mediation, inspection report of the
accident vehicle, postmortem examination report of victim,
traffic accident scene investigation note, testimony of witness
Song and defendant Chen's confession and explanation.

Argumentation-Driven Form

证明上述事实的证据有：
1. 接受刑事案件登记表、受理道路交通事故案件登记表、立案决定书、取
保候审决定书证实：本案报案、立案及丁光富被采取强制措施的情况。
2. 常住人口信息资料证实：丁光富达到完全刑事责任年龄及身份信息。
3. 车辆检验鉴定报告书、鉴定意见告知书证实：小型客车制动性能良好。
...

The above facts can be proved by the following evidence:
(1) The registration form for accepting criminal cases, the registration form for
accepting cases of road traffic accidents, the decision to file a case and the
decision to obtain bail pending trial prove that the situation of the file and
the compulsory measures for Ding;
(2) The resident population information prove that Ding has reached the age
of full criminal responsibility and identity information;
(3) The vehicle inspection appraisal report and appraisal opinion notice prove
that the braking performance of small passenger cars is well;

图 3 实际案件中的证据描述的两形式

1.3 研究目标与创新性

本论文工作拟完成以下几个研究目标：

首先，针对当前司法预测工作并没有针对案情描述的文本内容与法律条文的维度关联的相关研究的不足，本工作拟提出一个案情描述文本与法条维度匹配的数据集，并将在学术界公开。此外将提出一个案情描述文本重要性预测模型，实现法律条文与案情描述文本段的对应关系预测。在此之后，也将基于从案情描述中挖掘的判案论点和抽取的关键信息，实现一些常见司法预测任务，例如罪名预测等。

其次，针对司法证据预测这个课题现有研究的不足，本工作拟提出一个司法证据数据集，并将在学术界公开。此外将提出一个司法证据预测模型，实现针对案情描述文本可解释地预测其所需的完备证据集。

最后，将上述的模型组合成完整的司法辅助系统，开发的软件系统将用于辅助司法从业者以及其他相关人员分析理解案情信息并做出有关决策。

论文将为智能司法工作，引入了案情描述文本分析任务和证据预测任务，提高了其他司法任务的可解释性，为司法从业者提供了额外的信息来辅助理解和判决案情。

2 研究内容与技术路线

本课题拟实现一个基于法条的判案论点细粒度挖掘的司法辅助系统，本论文的工作分为如下几个方面，如图 4 所示。

该系统首先细粒度解耦法律条文，根据法律条文获得分析案件所需的维度。在此基础上，以案情描述作为输入，基于细粒度解耦的法律条文，自动挖掘判案论点并且自动识别案情中的关键信息。该关键信息将被使用于最后，根据挖掘的判案论点和案情关键信息，分析论证每一个判案论点所需的证据，预测该案件判决为某罪名所需的完备证据集。上述三个功能组合成完整的司法辅助系统，从而能够辅助司法从业者分析理解与决策。

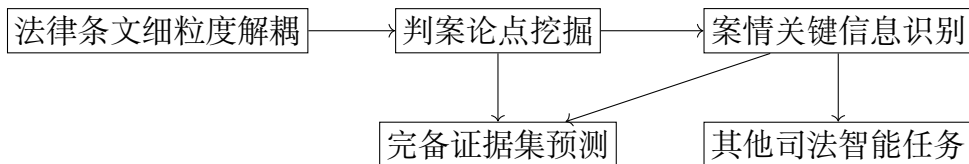


图4 研究内容关系示意图

2.1 研究内容一：法律条文细粒度解耦

根据现有的司法法律条文，我们从法律条文中细粒度解耦出分析案件所需的维度。法律条文的内容具有一定的抽象性、概括性、精确性和整体性。从学理上将，中国是典型的大陆法系国家，大陆法系要求法官遵从法律明文办理案件，这与英美法系的判例法有着明显的区别。所以，辅助国内司法从业者，就需要对法条进行细粒度解耦，并根据解耦的要素来指导进一步对案情描述的分析。

图5展示的是本工作目前对一部分法律条文实现的手动维度解耦，由图可见，本工作将多个法条的内容进行了拆解，并且组合成了法条指导下，盗窃罪所需要关注的多个维度。在之后的工作中，我们将进一步研究如何使用自动化程序，代替人工手动来实现法条解耦的功能，并且对比之间的差距。

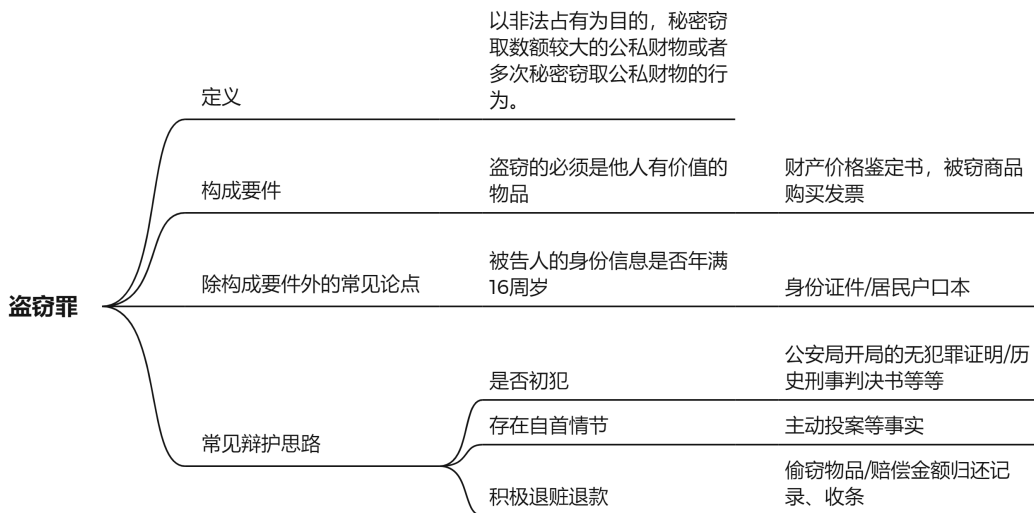


图5 法律条文细粒度解耦示意图

2.2 研究内容二：基于法律条文的案情关键信息识别与关联

对于一篇给定的案情描述，我们拟基于法律条文解耦得到的重要维度，自动总结出该案件的判案论点。之后，根据每一个判案论点，从案情描述中自动识别出对应关键信息，将法律条文与案情描述进行论点上的关联和维度上的对应。这一部分的判案论点和关键信息可以被使用于罪名预测、法条预测等司法下游任务，该种方式不仅具有更强的可解释性，而且能够更好的辅助司法从业人员理解案情、书写文书以及做出决策。

如图 6 所示，我们使用法律条文中解耦出的维度指导案情关键信息识别。法律条文中的维度都能在案情描述中找到一一对应的关键信息，帮助司法相关人员定位关键信息，更有效的解读案情内容。而使用这一部分基于法律条文识别出的关键信息进一步完成其他智能司法任务，也将为这些任务带来更多的可解释性。

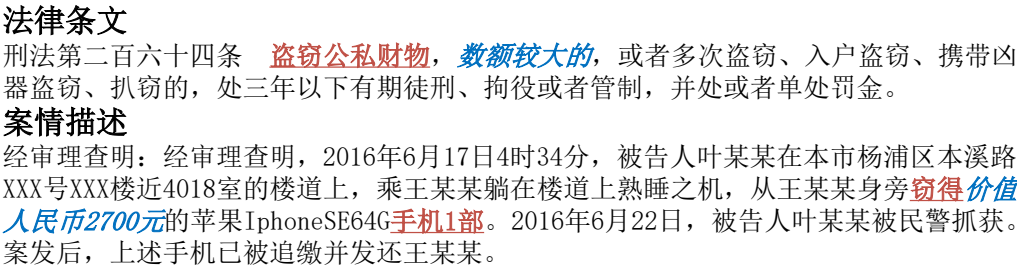


图 6 法律条文指导案情关键信息识别

2.3 研究内容三：基于法律条文的司法证据预测与生成

在司法案件中，完备的证据集至关重要，其在定罪与量刑中起着决定性的作用。所以根据一个案件的案情描述，总结出其完备的证据集，能够很好的辅助到司法从业者。我们拟基于案情描述的文本信息以及上述挖掘的判案论点，根据每一个论点预测生成对应证据，列举完备的论据可以充分论证判案论点，从而构建出完备的证据集。

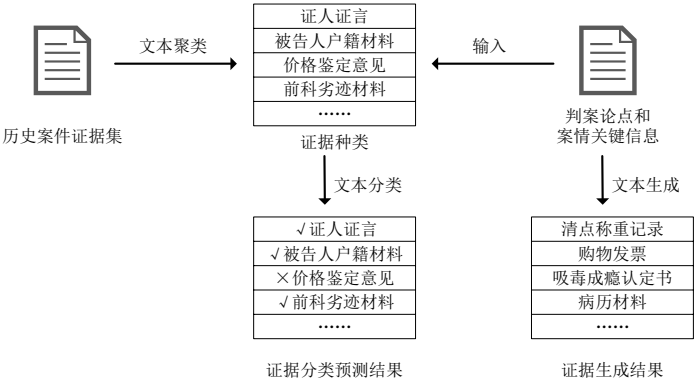


图 7 司法证据预测与生成任务说明

如图 7 所示，对于案情描述，其包含的证据一部分的作用为论证该案情所涉及的法条所解耦出的维度。另一部分的作用为证明案情中描述的某一特定事实，这一部分的证据往往不能通过聚类汇聚成一个证据类别，而是根据案情描述的具体内容进行生成。因此，本工作拟对这两种证据分别进行生成。

2.4 基于注意力机制和强化学习的判案论点挖掘

由于司法领域的独特性，模型给定的司法预测结果都需要具有可解释性。因为中国使用的是大陆法系，所以司法从业人员做出的决策都是遵从的法律明文，故而，模型在给定任何司法预测结果前，都需要基于法条内容，给出该案情描述所对应的判案论点和包含的关键信息。此处，本工作拟使用注意力机制和强化学习共同完成该任务。

在判案论点挖掘中，为了更好的体现语义高度凝练法律条文的指导作用，本工作拟通过注意力机制指导案情描述的文本进行解读。具体来说，如图 8 所示，本工作拟使用法律条文解耦出的维度信息作为查询，来指导模型关注到案情描述中的最相关的语义文本。从而，能够分辨出案情描述中的关键信息，忽略掉在法条指导下不相关、不重要的语义信息，实现判案论点挖掘。此时挖掘的判案论点和注意力机制下所关注到的文本片段，可以辅助完成其他司法任务，并且提供可解释性。

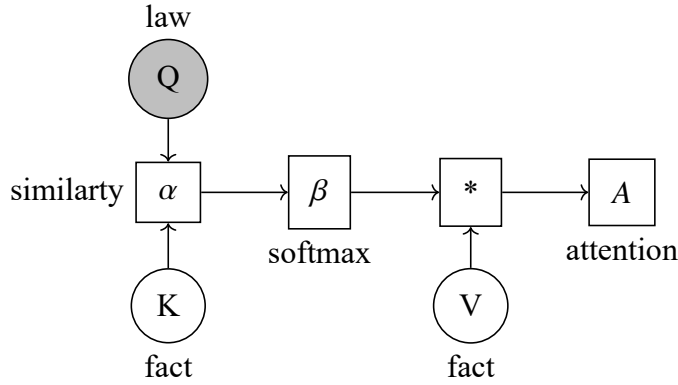


图 8 基于注意力机制的判案论点挖掘

具体的计算过程如下。其中 A 即为基于法律条文计算出的案情描述的注意力值。

$$\alpha = Q_{law} K_{fact}^T$$

$$\beta = softmax(\frac{\alpha}{\sqrt{d_k}})$$

$$A = \beta * V_{fact}$$

对于案情描述中的关键信息抽取，本工作期待通过注意力机制得到的文本是可读的。因此，本工作引入了强化学习机制^[25]，以罪名预测等司法任务的结果作为激励，最终让模型能够在每一个案情描述的文本片段上做出是否为关键信息的二元决策。强化学习是受到生物能够有效适应环境的启发，以试错的机制与环境进行交互，通过最大化累积奖赏的方式来学习到最优策略。目前常用的强化学习方法包括蒙特卡罗、Q 学习、SARSA 学习、TD 学习、策略梯度和自适应动态规划等。

2.5 基于文本聚类的司法证据关联和预测

针对司法证据，本工作首先构造了一个司法证据的数据集，数据集包括案情描述、该案件证据集、罪名、所用法条和处罚结果。本工作拟使用多种方法实现关联组织来自不同案件、论证不同判案论点的分散证据。

首先，本工作选择使用 InfoMap 算法^[26]，将证据视为短文本，进行文本的无监督聚类。本工作使用 SimBERTv2 模型，将证据文本向量化。之后 InfoMap 算法将 SimBERTv2 模型转换为的文本向量，通过最小化熵来寻求最优的证据文本聚类结果。之后，将聚类结果中出现频次较多的类别归堆，关联后视为同一种证据，在之后，基于案情描述预测证据集的任务时，将会使用关联总结后的证据类别进行预测。

其次，将使用证据对的论证信息来计算它们之间的潜在距离。Nogueira 等人^[27]将相关查询段落对之间的相关性定义为 0，将不相关的查询段落对定义为 1。类似地，我们假设支持相同司法论点的两个证据之间的距离较小，将支持相同司法论点的证据对之间的距离被标记为 0。证明不同司法论点的同一刑事案件所涉及的证据对之间的距离被标记为 1。从而作为标记数据，计算司法证据对之间的潜在距离，并使有潜在距离进行聚类。

2.6 基于序列到序列模型的司法证据生成

在一部分的案情描述中，有一部分的证据来自于案情描述的具体需求。例如：案情描述中提及了“嫌疑人通话”的内容，那么本案件就需要“通话记录”作为证据之一。所以，这一部分的司法证据，则需要根据案情描述的具体文本进行生成。

因为，本工作使用序列到序列模型，来完成这部分的司法证据生成。该序列到序列模型的输入为，案情描述的片段，输出则为证明该片段所需的证据。如果该段案情描述无需额外特殊证据证明，则不进行输出。

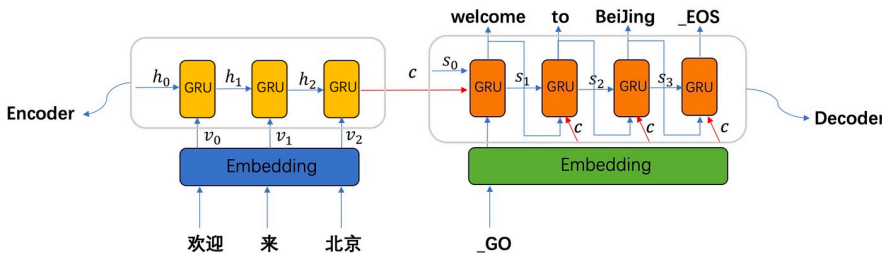


图 9 序列到序列模型

序列到序列学习 (Seq2Seq) 指的是训练模型学习从一个序列到另一个序列的映射函数。Seq2Seq 模型通常都是基于编码器-解码器的架构，分为编码器 (Encoder) 和解码器 (Decoder) 两部分，其结构如图 9 所示。其中 Encoder 需要接收、处理输入序列，并且编码生成对应的上下文向量，作为下一步的语境；而 Decoder 会接收

上一步产生的上下文向量，逐步预测目标序列的下一个字符，从而完成全部目标序列的生成。

3 论文工作安排计划

3.1 工作进度安排

日期	计划与安排
2022.12-2023.03	标注、整理、构建数据集，阅读相关论文，根据现有方法和模型构建基线模型。
2023.04-2023.08	设计并实现模型，完成对应实验，分析实验结果。
2023.08-2023.10	设计实现可视化展示部分，构建完整系统。
2023.10-2023.11	整理总结研究成果，撰写毕业论文。

表 1 工作进度安排

本论文的工作进度安排如表 1 所示。

3.2 论文工作基础

3.2.1 数据集

实验目前所用数据集，来自三个方面。

- 1、 本工作构建了一个刑事判决文书的证据数据集。证据数据集中，每份判决书由案情描述、证据集、罪名、相关法条、判决结果构成。数据集来自于中国裁判文书网公开的原始刑事一审判决文书。由于判决文书的书写规范性、严谨性，本工作通过一定的规则，从原始文书中抽取所需的信息，从而构建证据数据集（之后会考虑聘请具有法律背景的专家来进行人工标注）。
- 2、 本工作构建了一个案情描述关键信息数据集。在自动抽取构建的证据数据集中，本工作随机选择了一部分的判决书，人工标注了基于法条指导下的案情描述中的关键信息。具体来说，基于法条解耦出的维度，本工作标注了每一个维度对应的案情描述中的语句，本工作认为这一部分的信息是案情中的“关键信息”，并且作为任务预测目标。
- 3、 本工作使用了司法决策任务相关论文 [5, 8, 28-29] 常用的数据集，来自法研杯 2018 年的 CAIL 数据集^[30]。CAIL 数据集中的每份判决书由案情描述和判决结果（适用法律条文、罪名、刑期）组成。CAIL 有两个子数据集：CAIL-big 和 CAILsmall。本工作按照现有工作^[6, 17]对 CAIL 数据集进行预处理。具体而言，论文过滤掉了具有多个适用罪名和法律条文的案例样本。此外，本工

作仅保留不少于 100 个相应案例样本的罪名和法律条文，并且删除了所有的二审案例样本。

3.2.2 调研工作

目前已经完成的调研工作有：

- 1、智能司法相关任务的调研工作。任务包括罪名预测、相关法条预测、刑期预测、法庭观点生成等任务。调研了近年的文献和综述，复现了多篇论文的工作。
- 2、证据关联的调研工作。调研了与司法证据相关的论文工作，这一部分的相关论文和工作较少，之后考虑复现相关论文。
- 3、智能司法辅助系统的调研工作。调研了市面上主流的一些司法辅助系统，了解了目前智能司法的主要需求，这对之后的论文工作和司法辅助系统的构建，均有指导意义。

3.2.3 现有实验结果

待补充

3.3 可能遇到的问题以及解决途径

本文所设涉及的工作的主要的技术难点有：

- 1、对于司法文书的案情描述，其长度并不固定，有较多案情描述的长度超出 BERT 模型所能接纳的最大长度，这使得直接使用 BERT 模型来理解案情描述的语义信息产生问题。所以，本工作所需要处理长文本建模这一技术难点。本工作拟对案情描述进行分段的语义处理的方式，综合多段的语义结合成长文本的案情描述语义信息。
- 2、在案情描述的关键信息抽取中，无论是基于强化学习还是注意力机制的机理解释，其结果的连贯性和可读性都是重要的问题。关键信息抽取若以词为单位，抽取出的关键事实很可能是不可读的词语碎片，而若基于句子，则粒度过于粗糙。因此如何构造关键信息数据集，并且基于此进行文本分类实验，提高模型给定关键信息的可读性，会在之后的实验中逐步探索。
- 3、当从案情描述中总结出判案论点和关键信息之后，如何使用这一部分的语义信息来指导后续的司法任务以及证据预测任务，并且体现可解释性是难点之一。在本工作中，其指导和可解释证据预测体现在判案论点和证据集的对应关系上，其指导和可解释其他司法任务体现在，本工作使用抽取的部分关键信息和挖掘的判案论点，替代原有的案情描述全文，交给下游司法任务，由此论证。但是如何明确对应关系，以及只使用关键信息，而非全文信息导致了一定量的信息丢失，是否能给下游司法任务带来提升，都是本工作在开展

实验的过程中需要继续面对的问题。

- 4、 虽然司法判决文书的写作具有规范性和严谨性，但是由于全国不同地域不同级别的法院对法律文书的书写习惯不同，司法判决文书的行文结构会存在一定的差异。本工作基于规则从判决文书中抽取出的证据集会参杂部分噪声，从而影响证据集的聚类以及预测工作。因此，本工作通过方法抽取出证据后，仍需要大量的数据清洗工作和部分人工过滤与处理得到较高质量的数据集。
- 5、 在证据生成的实验中，可能会存在某一个案情描述片段没用对应任何证据，或者某一个案情描述片段对应了多个证据，但是判决书中罗列的证据彼此之间没有先后关系。因此，这种情况对序列到序列模型提出了更高的要求，在之后的模型构建中，会针对这种特殊的场景设计与之相对应的模型结构。

主要参考文献

- [1] 程金华. 法院”案多人少”的跨国比较——对美国、日本和中国的实证分析[J]. 社会科学辑刊, 2022(5):72-84.
- [2] Bibal A, Lognoul M, De Streel A, et al. Legal requirements on explainability in machine learning[J]. Artificial Intelligence and Law, 2021, 29(2):149-169.
- [3] Chen Y L, Liu Y H, Ho W L. A text mining approach to assist the general public in the retrieval of legal documents[J]. Journal of the American Society for Information Science and Technology, 2013, 64(2):280-290.
- [4] Ye H, Jiang X, Luo Z, et al. Interpretable charge predictions for criminal cases: Learning to generate court views from fact descriptions[J]. arXiv preprint arXiv:1802.08504, 2018.
- [5] Zhong H, Guo Z, Tu C, et al. Legal judgment prediction via topological learning [C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018: 3540-3549.
- [6] Xu N, Wang P, Chen L, et al. Distinguish confusing law articles for legal judgment prediction[J]. arXiv preprint arXiv:2004.02557, 2020.
- [7] Chao W, Jiang X, Luo Z, et al. Interpretable charge prediction for criminal cases with dynamic rationale attention[J]. Journal of Artificial Intelligence Research, 2019, 66: 743-764.
- [8] Feng Y, Li C, Ng V. Legal judgment prediction via event extraction with constraints [C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022: 648-664.
- [9] Teng Y, Chao W. Argumentation-driven evidence association in criminal cases[C]// Findings of the Association for Computational Linguistics: EMNLP 2021. 2021: 2997-3001.
- [10] Feng Y, Li C, Ng V. Legal judgment prediction: A survey of the state of the art[C]// Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22. 2022: 5461-5469.
- [11] Cui J, Shen X, Nie F, et al. A survey on legal judgment prediction: Datasets, metrics, models and challenges[J]. arXiv preprint arXiv:2204.04859, 2022.
- [12] Lin W C, Kuo T T, Chang T J, et al. Exploiting machine learning models for chinese legal documents labeling, case classification, and sentencing prediction[J]. Proceedings of ROCLING, 2012, 17(4):140.
- [13] Luo B, Feng Y, Xu J, et al. Learning to predict charges for criminal cases with legal basis[J]. arXiv preprint arXiv:1707.09168, 2017.
- [14] 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.
- [15] Liu C L, Liao T M. Classifying criminal charges in chinese for web-based legal services[C]//Asia-pacific web conference. Springer, 2005: 64-75.
- [16] Liu Y H, Chen Y L, Ho W L. Predicting associated statutes for legal problems[J]. Information Processing & Management, 2015, 51(1):194-211.
- [17] Yang W, Jia W, Zhou X, et al. Legal judgment prediction via multi-perspective bi-feedback network[J]. arXiv preprint arXiv:1905.03969, 2019.
- [18] Wu Y, Kuang K, Zhang Y, et al. De-biased court' s view generation with causality [C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 763-780.
- [19] Kim M Y, Xu Y, Goebel R. Legal question answering using ranking svm and syntactic/semantic similarity[C]//JSAI International Symposium on Artificial Intelligence. Springer, 2014: 244-258.
- [20] Carvalho D S, Nguyen M T, Tran C X, et al. Lexical-morphological modeling for legal text analysis[C]//JSAI International Symposium on Artificial Intelligence. Springer, 2015: 295-311.
- [21] Xiao C, Hu X, Liu Z, et al. Lawformer: A pre-trained language model for chinese legal long documents[J]. AI Open, 2021, 2:79-84.
- [22] 潘自强, 邵新. 裁判文书说理: 内涵界定与原则遵循[J]. 法治研究, 2018(4):99-107.
- [23] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.
- [24] Chen Q, Zhu X, Ling Z, et al. Enhanced lstm for natural language inference[J]. arXiv preprint arXiv:1609.06038, 2016.
- [25] Mnih V, Kavukcuoglu K, Silver D, et al. Playing atari with deep reinforcement learning[J]. arXiv preprint arXiv:1312.5602, 2013.
- [26] Rosvall M, Bergstrom C T. Maps of random walks on complex networks reveal community structure[J]. Proceedings of the National Academy of Sciences, 2008, 105(4):1118-1123.
- [27] Reimers N, Schiller B, Beck T, et al. Classification and clustering of arguments with contextualized word embeddings[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 567-578.
- [28] Liu L, Zhang W, Liu J, et al. Interpretable charge prediction for legal cases based on interdependent legal information[C]//2021 International Joint Conference on Neural Networks (IJCNN). IEEE, 2021: 1-8.
- [29] Yue L, Liu Q, Jin B, et al. Neurjudge: a circumstance-aware neural framework for legal judgment prediction[C]//Proceedings of the 44th International ACM SIGIR

- Conference on Research and Development in Information Retrieval. 2021: 973-982.
- [30] Xiao C, Zhong H, Guo Z, et al. Cail2018: A large-scale legal dataset for judgment prediction[J]. arXiv preprint arXiv:1807.02478, 2018.
- [31] Wang P, Fan Y, Niu S, et al. Hierarchical matching network for crime classification [C]//proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval. 2019: 325-334.