

北京航空航天大学计算机学院

博士研究生学位论文 开题报告

论文题目：基于法条的判案论点细粒度挖掘

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 17 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 学术界理论研究背景 1

1.1.2 项目研究背景 1

1.1.3 实际应用背景 1

1.2 研究现状概述 1

1.3 研究目标与创新性 1

2 研究内容与技术路线 2

2.1 研究内容一：法律条文细粒度解耦 2

2.2 研究内容二：判案论点挖掘 2

2.3 研究内容三：司法证据预测与生成 2

2.4 基于注意力机制和强化学习的判案论点挖掘 3

2.5 基于文本聚类的司法证据关联和预测 4

2.6 基于序列到序列模型的司法证据生成 4

3 论文工作安排计划 5

3.1 工作进度安排 5

3.2 论文工作基础 5

3.2.1 数据集 5

3.2.2 调研工作 6

3.2.3 现有实验结果 6

3.3 可能遇到的问题以及解决途径 6

主要参考文献 8

图 目

图 1 [研究内容关系示意图](#) 2

图 2 [司法证据预测与生成任务说明](#) 3

图 3 [基于注意力机制的判案论点挖掘](#) 4

图 4 [序列到序列模型](#) 5

目 录

表 1 工作进度安排 5

1 论文研究背景与意义

1.1 论文选题背景

参考文献引用如下^[1-5]

(可以从如下几个方面进行论述：1、学术界理论研究背景，2、项目研究背景，3、实际应用背景)

1.1.1 学术界理论研究背景

1.1.2 项目研究背景

1.1.3 实际应用背景

1.2 研究现状概述

针对本论文遇到的问题，XXX 等人的方法存在 XXXX 问题。(用 2 页左右的篇幅，对文献综述所罗列的研究现状进行总结和分析，并列举与论文密切相关的几项工作)

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 1 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描述中的关键信息。最后，根据判案论点和案情描述中的关键信息，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。

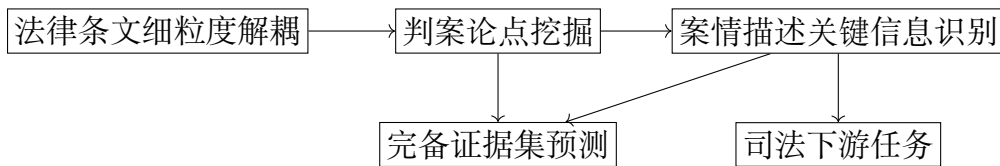


图 1 研究内容关系示意图

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。

[此处贴法律条文细粒度解耦的举例图](#)

2.2 研究内容二：判案论点挖掘

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。

[此处贴判案论点挖掘的举例图](#)

2.3 研究内容三：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。

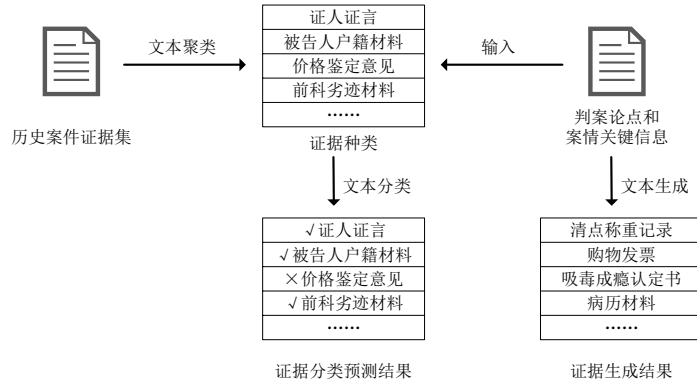


图 2 司法证据预测与生成任务说明

如图 2 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 3 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。

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。

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

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

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

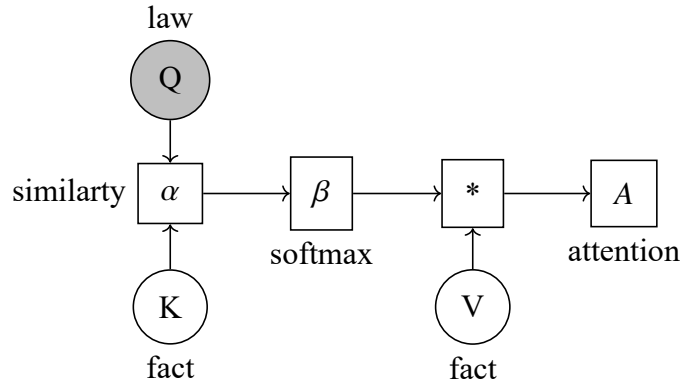


图3 基于注意力机制的判案论点挖掘

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，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。

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。本工作为关联组织来自不同案件、论证不同判案论点的分散证据，选择使用 InfoMap 算法^[6]，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。

此处贴 InfoMap 算法的简单原理公式和图

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。

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 4 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

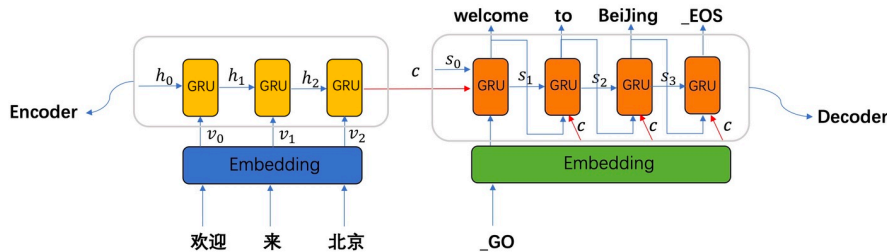


图 4 序列到序列模型

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

（在以下几个方面选择几个方面进行说明：1）收集或者准备的数据集、2）完成或者正在进行的调研工作；3）已完成或者正在进行的理论推导、4）已经完成或者正在进行的开发系统或软件模块、5）正在进行或者完成的实验与实验结果）

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【引用】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，从原始文书中抽取出所需的信息，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文书，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【引用】常用的数据集，来自法研杯 2018 年的 CAIL 数据集^[7]。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^[8-9]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

3.2.3 现有实验结果

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。

（描述论文可能遇到理论证明、工程开发、核心算法不足、实验数据等方面困难以及应对措施）

主要参考文献

- [1] Yan S, Xiong Y, Lin D. Spatial temporal graph convolutional networks for skeleton-based action recognition[C]//Thirty-second AAAI conference on artificial intelligence. 2018.
- [2] Wu Z, Pan S, Chen F, et al. A comprehensive survey on graph neural networks[J]. IEEE transactions on neural networks and learning systems, 2020, 32(1):4-24.
- [3] 江载芳, 申昆玲, 沈颖. 诸福棠实用儿科学 (上册)[M]. 8 版. 人民卫生出版社, 2015.
- [4] 胡坚, 李崇巍, 胡秀芬, 等. 幼年皮炎炎诊治建议[J]. 中华儿科杂志, 2012, 50(8): 617-621.
- [5] Lovell D J, Lindsley C B, Rennebohm R M, et al. Development of validated disease activity and damage indices for the juvenile idiopathic inflammatory myopathies: Ii. the childhood myositis assessment scale (cmas): a quantitative tool for the evaluation of muscle function[J]. Arthritis & Rheumatism: Official Journal of the American College of Rheumatology, 1999, 42(10):2213-2219.
- [6] 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.
- [7] Xiao C, Zhong H, Guo Z, et al. Cail2018: A large-scale legal dataset for judgment prediction[J]. arXiv preprint arXiv:1807.02478, 2018.
- [8] Xu N, Wang P, Chen L, et al. Distinguish confusing law articles for legal judgment prediction[J]. arXiv preprint arXiv:2004.02557, 2020.
- [9] Yang W, Jia W, Zhou X, et al. Legal judgment prediction via multi-perspective bi-feedback network[J]. arXiv preprint arXiv:1905.03969, 2019.