Beijing Forest Studio
北京理工大学信息系统及安全对抗实验中心

法律文本可解释性研究

硕士研究生：董勃

导师：罗森林

2020年11月22日

内容提要

- 背景介绍
- 基本思路
- 算法原理
- 实验结果
- 未来工作
- 应用总结
- 参考文献

预期收获

- 预期收获
  - 1.了解自然语言处理可解释性目的和意义
  - 2.了解自然语言处理领域常用可解释性方法
  - 3.了解可解释性在法律领域的应用

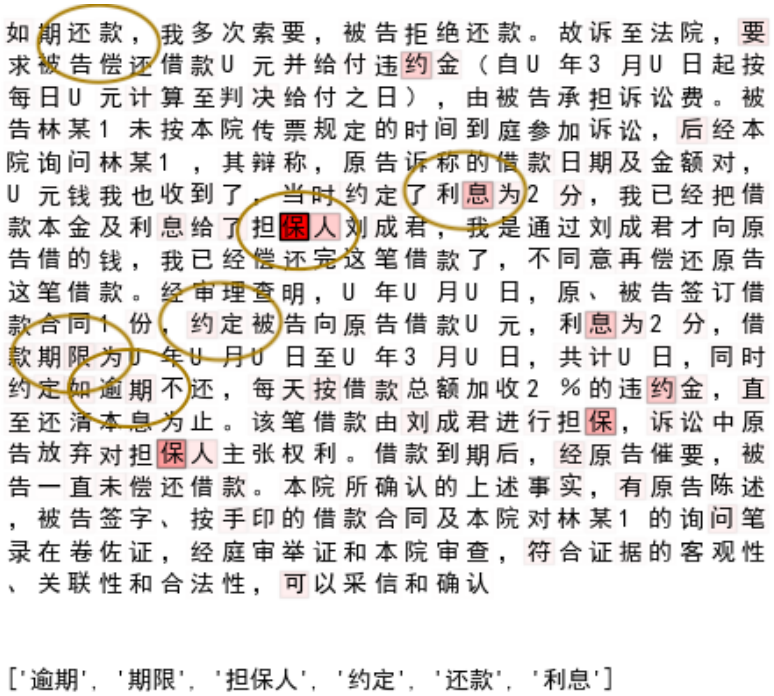
背景介绍

- 什么是可解释性?
  - NLP系统输出的结果，应该符合基本的语言学规律，符合领域知识的预期，可以用语言学的术语和业务领域的语言进行解释。
  - 让模型说人话
- 为什么需要可解释性?
  - 模型性能指标（准确率、精确率、召回率等），会随着数据集不同或时间推移而发生漂移，性能发生变化。
  - 对模型判断结果有更好的理解
  - 帮助用户信任AI
  - 向AI学习做出更好的决策

基本思路

基本思路

- NLP可解释性，从哪里来？
  - 主要集中在深度学习模型的可视化上
  - 必须基于语言特征， NLP的可解释性应该与我们的语言直觉相一致，从模型、算法的输入，到每个处理环节，再到最终的输出，都可以用基本的语言特征和语言结构来解释



基本思路

- 与领域知识有明确的因果关系
  - 可解释性和应用场景强相关。应用场景的知识，决定着NLP的需求边界、任务类型、结果预期等内容。
  - 取件延误
    - 取件的人怎么还不来?
    - 两个小时都过了还没有来取件。
  - 啤酒与尿布

基本思路

- 可解释性研究角度
  - nlp算法可解释性
    - 对算法过程的去魅，揭示算法过程中的逻辑性
    - 揭示中间过程所产生的特征的相关性、具体参数值对结果的影响。
  - 处理结果的可解释性
    - 承认算法过程是个黑盒子，对输出结果的统筹规划。

基本思路

- 可解释性方法
  - 建模之前的可解释性
    - 数据可视化
    - 探索性质的数据分析
  - 建立本身具有可解释性的模型
    - 基于规则的方法
    - 基于单个特征的方法
    - 基于实例的方法
  - 建模之后使用可解释性方法对模型作出解释
    - 隐层分析法
    - 敏感性分析法
    - 代替/替代模型法

基本思路-课题引入

- 案件文本分析
  - 罪名预测
  - 相似案例匹配
  - 链接：https://www.isclab.org.cn/2020/03/29/案件文本分析/
    - CAIL比赛说明
    - 案件文本任务分析
    - 基础模型框架介绍

基本思路-课题引入

- 裁决文书要素

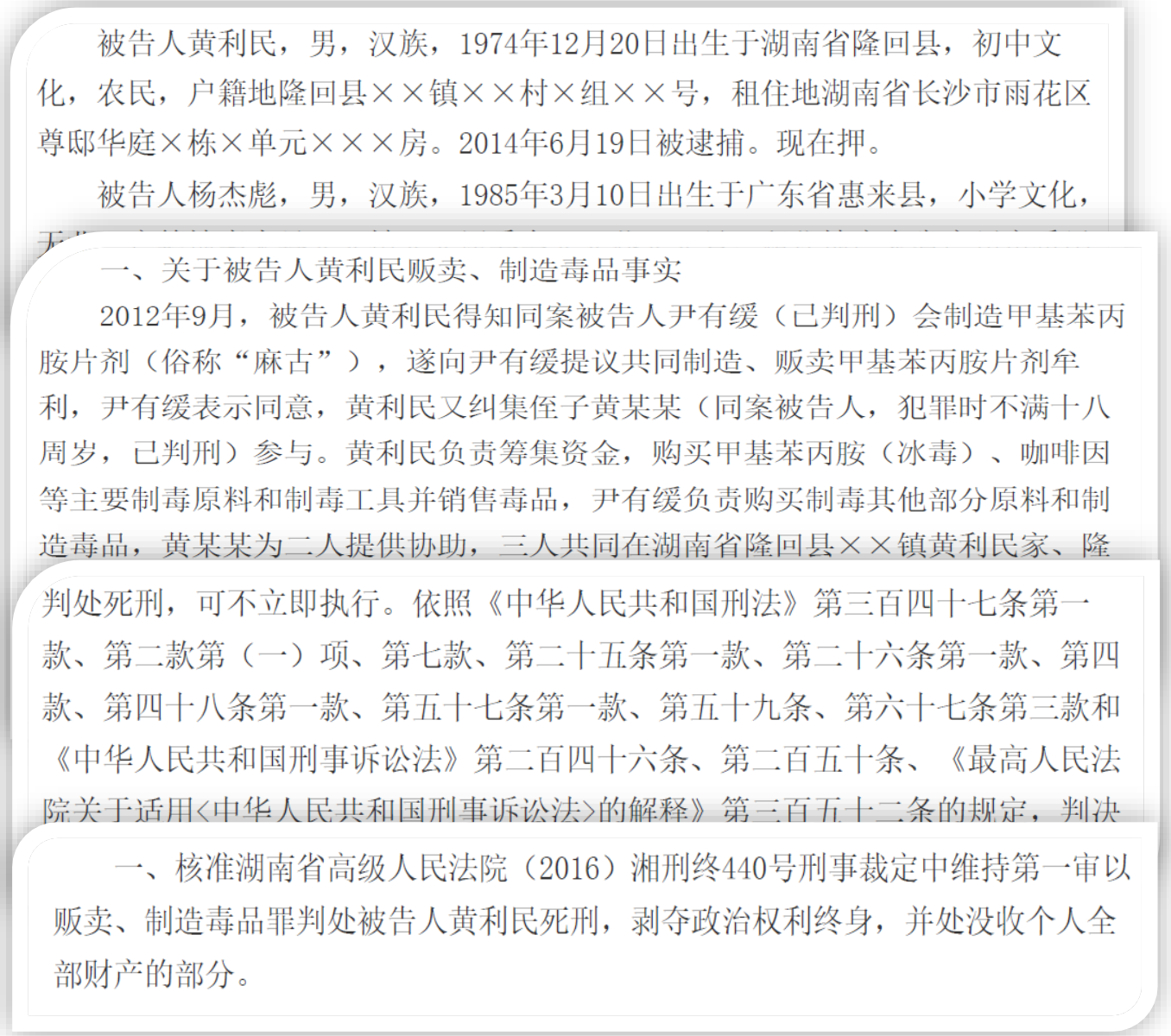
a. 原被告信息

b. 案件事实描述

c. 依据法条

d. 判决罪名

e. 刑期



被告人黄利民，男，汉族，1974年12月20日出生于湖南省隆回县，初中文化，农民，户籍地隆回县××镇××村×组××号，租住地湖南省长沙市雨花区尊邸华庭×栋×单元×××房。2014年6月19日被逮捕。现在押。

被告人杨杰彪，男，汉族，1985年3月10日出生于广东省惠来县，小学文化，

一、关于被告人黄利民贩卖、制造毒品事实

2012年9月，被告人黄利民得知同案被告人尹有缓（已判刑）会制造甲基苯丙胺片剂（俗称“麻古”），遂向尹有缓提议共同制造、贩卖甲基苯丙胺片剂牟利，尹有缓表示同意，黄利民又纠集侄子黄某某（同案被告人，犯罪时不满十八周岁，已判刑）参与。黄利民负责筹集资金，购买甲基苯丙胺（冰毒）、咖啡因等主要制毒原料和制毒工具并销售毒品，尹有缓负责购买制毒其他部分原料和制造毒品，黄某某为二人提供协助，三人共同在湖南省隆回县××镇黄利民家、隆

判处死刑，可不立即执行。依照《中华人民共和国刑法》第三百四十七条第一款、第二款第（一）项、第七款、第二十五条第一款、第二十六条第一款、第四款、第四十八条第一款、第五十七条第一款、第五十九条、第六十七条第三款和《中华人民共和国刑事诉讼法》第二百四十六条、第二百五十条、《最高人民法院关于适用<中华人民共和国刑事诉讼法>的解释》第三百五十二条的规定，判决

一、核准湖南省高级人民法院（2016）湘刑终440号刑事裁定中维持第一审以贩卖、制造毒品罪判处被告人黄利民死刑，剥夺政治权利终身，并处没收个人全部财产的部分。

算法原理

算法原理

- 自省解释模型
  - 解释模型如何确定其最终输出。
  - 考虑罪名和法条的高度相关性，引入法条信息来提高罪名预测的准确性，同时法条也可以被视为对罪名预测的一种可解释性。
- 生成解释信息模型
  - 生成一些句子或词语，作为支撑系统预测结果的相关依据。
  - 根据案件的事实描述和给定的罪名标签来生成法院观点。

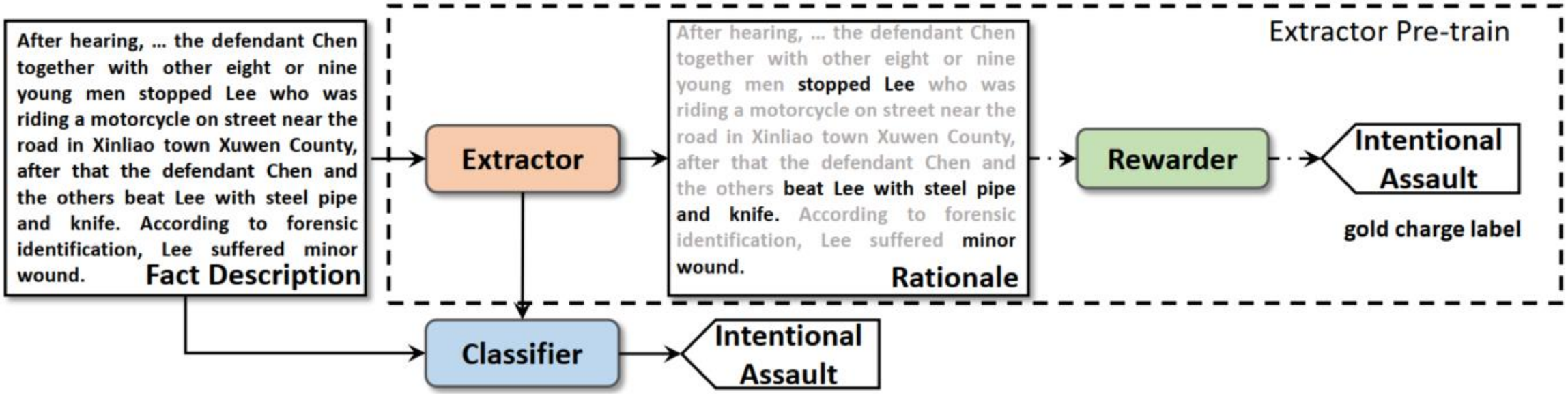
可解释的Rationable增强罪名预测系统 算法原理

T	提高罪名预测系统可解释性，自动生成法律文档
I	自行构建17万份中国裁判文书网文书
P	1. 深度强化学习方法提取事实描述中的基本原理 2. 基本原理融入分类模型 3. 分类预测
O	法律文档罪名预测+可解释性内容

P	模型自动判决只给出最终结果而不提供任何解释
C	自行构建数据集以及基本原理标注库
D	基本原理无标注语料库，且基本原理粒度难以掌握
L	COLING2018

算法原理

- 处理过程
  - 首先使用深度强化学习方法来提取事实描述中的rationales（基本原理/解释/依据），rationales意味着从输入文本中提取简短且具有决定性的文字片段。然后将rationales信息融入到分类模型中，以提高预测的准确性。
- 模型架构



实验结果

CASE 1 $[Official\ Embezzlement]_{charge}$

…PP 利用 其 担任 [公司 业务员 的 职务 便利]$_{key\ point}$， 从 公司 仓库 提 走 多部 手机， 后 将 手机 卖掉， 货款 挥霍 。 …

…Using his [position as a company salesman]$_{key\ point}$, PP took phones from the company's warehouse, sold the phones, and squandered the money…

CASE 2 $[Larceny]_{charge}$

…PP_1 [趁 PP_2 家中 无 人 之 机]$_{key\ point}$， 进入 到 PP_2 家 卧室 内 伺机 盗窃 。 被 PP_2 回家 后 发现， PP_1 翻 墙 逃跑…

…[When PP_2 was not at home]$_{key\ point}$, PP_1 went to PP_2's bedroom to steal. When PP_2 came home, PP_1 fled the wall and ran…

CASE 3 $[Negligently\ Causing\ Fire]_{charge}$

…在 焚烧 耕地 上 的 杂草 时， [不慎]$_{key\ point}$ 引发 山林 火灾 。案 发 后， PP 积极 救火， 主动 向 上级 说明 失火 情况…

… When burning weeds on land, PP [inadvertently]$_{key\ point}$ ignited the mountain fire. PP actively doused the fire and reported the fire situation …

CASE 4 $[Arson]_{charge}$

…PP_1 因 生意 竞争 与 PP_2 产生 积怨 。PP_1 酒后 [萌生 放火 烧 PP_2 手机店 的 念头]$_{key\ point}$， 进入 PP_2 的 店内 将 纸箱 点燃…

…PP_1 hates PP_2 for business competition. After drinking, PP_1 [wanted to burn PP_2's shop]$_{key\ point}$. PP_1 entered the shop and lighted the carton…

CASE 5 $[Negligent\ Homicide]_{charge}$

… PP_1 驾驶 货车 在 倒车 过程 中， [因 疏忽大意]$_{key\ point}$ 将 负责 指挥 倒车 的 PP_2 挤 伤， 后 PP_2 抢救 无效 死亡…

…When reserving the truck, PP_1 [inadvertently]$_{key\ point}$ injured PP_2, who was in charge of commanding PP_1. PP_2 died later. …

CASE 6 $[Intentional\ Homicide]_{charge}$

…PP_1 从 家中 携带 匕首 出门 寻找 PP_2 [进行 报复]$_{key\ point}$， 将 PP_2 捅倒 后， 在 颈 部 来回 割， 致 PP_2 当场 死亡…

…PP_1 took the dagger and looked for PP_2 [for revenge]$_{key\ point}$. He stabbed PP_2 and cut the neck back and forth, causing PP_2 to die on the spot…

引入法律属性的相似案例匹配 算法原理

T	从候选文书中匹配到与目标文书更相似的文书+解释性说明
I	CAIL2019 相似案例匹配数据集
P	1. 原始数据的规则匹配 2. 原始数据要素提取 3. 数据拼接 4. 结果预测
O	相似文书+解释性说明

P	模型预测只给出最终结果而不提供任何解释，且现有算法准确率较低
C	匹配规则自行制定，要素提取引入外部数据进行模型预训练，输入数据为三元组（A,B,C）
D	规则匹配正则表达式构建，法律文书长度过长
L	IJCNN2020

规则匹配

- 人工标注方案

任务描述
请仔细阅读说明文档，之后根据文章内容回答问题

原告：赵香娟，女，1968年5月31日出生，汉族，万荣县。 被告：张青霞，女，1963年4月30日出生，汉族，万荣县。 被告：黄建业，男，1962年5月10日出生，汉族，万荣县。

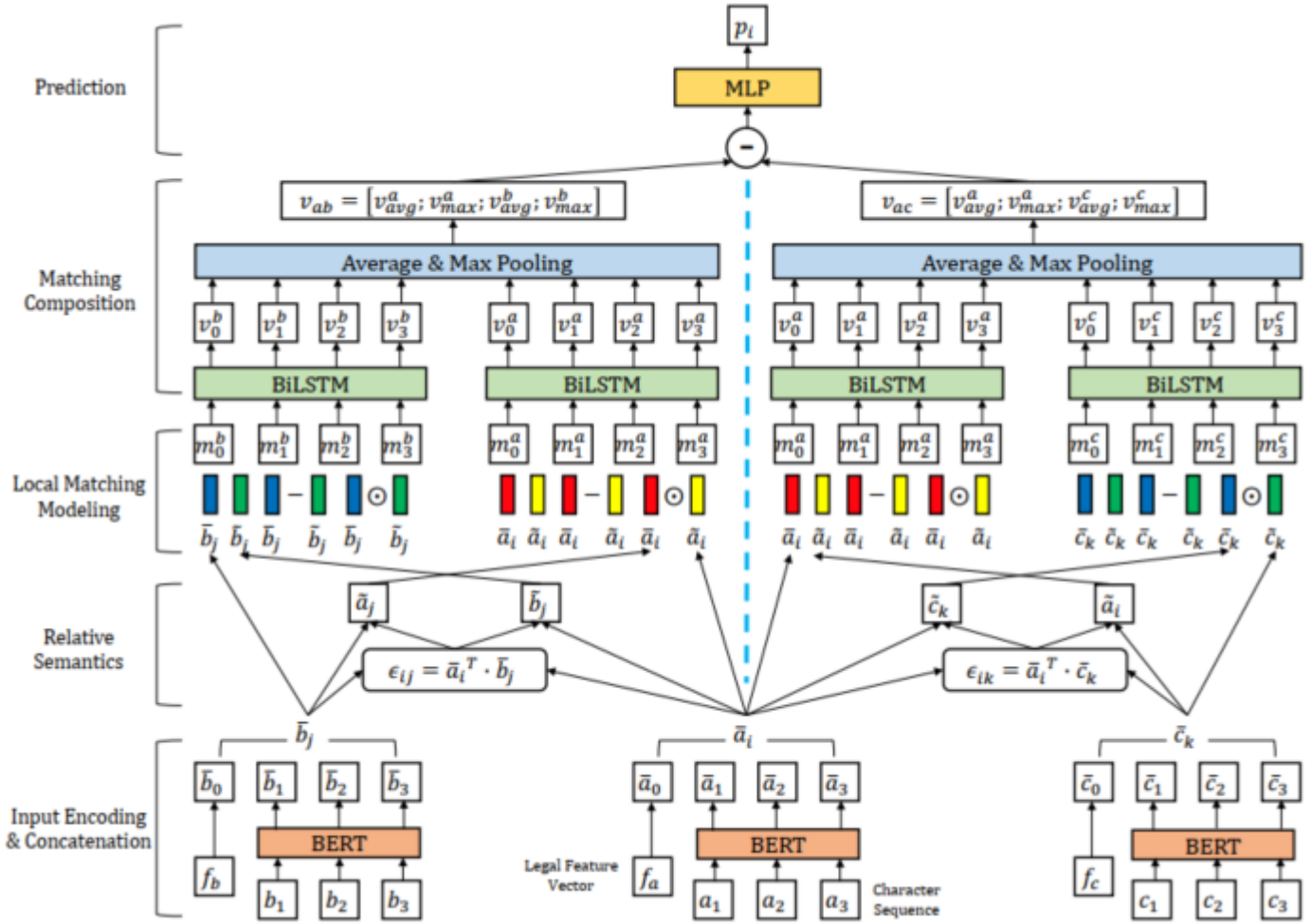
原告向本院提出诉讼请求：1、二被告立即偿还我200000元本金及其利息（利息按借条注明的利率从借款之日算至还清之日）；2、二被告承担诉讼费用。事实与理由：二被告系夫妻关系。其二人因家庭经营需要资金为由，先后分四次向我借款200000元。借款到期后，我多次催要，二被告至今未予归还，请求法院判如所请。 二被告未到庭参加诉讼，亦未提交答辩意见。 原告为证实其主张，提交了四张借条。1、被告张青霞于2013年4月24日借走原告现金20000元，借条未注明利息；2、被告张青霞于2013年5月24日借走原告现金30000元，借条注明月息为1.5分；3、被告张青霞、黄建业于2015年11月24日借走原告现金90000元，借条注明月息为1.5分；4、被告张青霞、黄建业于2016年4月1日借走原告现金60000元，借条注明月息为1.8分。上述借条，拟证明二被告共向原告借款200000元之事实。 原告称，上述被告张青霞于2013年4月24日所借的20000元，借条虽未注明利息，但二被告实际按月息2.5分向其结息，原告未就此进行举证。 原告称，2016年6月30，经结算二被告尚欠原告利息4000元。休庭后，原告考虑到她与二被告系亲戚关系，故向本院书面声明愿放弃4000元欠息，并主张四笔借款的利息应自2016年7月1日起，统一按月息1.5分计算。 二被告未提交任何证据。

规则匹配

- 规则匹配方案
  - 原被告数量和类型：法定代表人、自然人
  - 担保类型：无担保、抵押、担保
  - 年利率：0、24%、36%
  - 支付方式：微信转账、微信支付、支付宝、银行转账、手机银行、现金
  - 还款情况：已偿还、未偿还、部分偿还
  - 借款证据：合同、协议、收据、流水、担保、抵押、微信、短信、聊天

算法原理

- 算法框架


Prediction
Matching Composition
Local Matching Modeling
Relative Semantics
Input Encoding & Concatenation
MLP
Average & Max Pooling
Average & Max Pooling
BiLSTM
BiLSTM
BiLSTM
BiLSTM
BERT
BERT
BERT
Legal Feature Vector
Character Sequence

算法原理

- 注意力权重e_{ij}

$$e_{ij} = \bar{a}_i^T \cdot \bar{b}_j$$

- 语义相关矩阵：Softmax函数归一化权重，将语义相关矩阵映射到（0, 1）

$$\alpha_{ij}^a = \frac{\exp(\epsilon_{ij})}{\sum_{k=0}^{l_b} \exp(\epsilon_{ik})} \qquad \alpha_{ij}^b = \frac{\exp(\epsilon_{ij})}{\sum_{k=0}^{l_a} \exp(\epsilon_{kj})}$$

- 计算语义相关性（bj的内容与a的相关性）

$$\tilde{a}_i = \sum_{j=0}^{l_b} \alpha_{ij}^a \bar{b}_j, \forall \mathrm{i} \in [0, \ldots, l_a] \qquad \tilde{b}_j = \sum_{i=0}^{l_a} \alpha_{ij}^b \bar{a}_j, \forall \mathrm{j} \in [0, \ldots, l_b]$$

- 拼接矩阵：差与积

$$m^a = [\bar{a}; \tilde{a}; \bar{a} - \tilde{a}; \bar{a} \odot \tilde{a}] \qquad m^b = [\bar{b}; \tilde{b}; \bar{b} - \tilde{b}; \bar{b} \odot \tilde{b}]$$

实验结果

- 模型优化

	Method	Valid	Test
Baseline	BERT LSTM CNN	61.93 62.00 62.27	67.32 68.00 69.53
Our Baseline	BERT LSTM CNN	64.53 64.33 64.73	65.59 66.34 67.25
Best Score	11.2yuan Backward AlphaCourt	66.73 67.73 70.07	72.07 71.81 72.66
Our Method	LFESM	70.01	74.15

缺陷与不足

- 获取更多的法律属性信息
  - 由于正则表达式的限制，现有法律属性难以完全且准确的匹配到全部关键信息；
  - 某些信息难以形成正则表达式；
- 法律文书长度限制
  - BERT模型只能处理512字符以下的文档

改进方法

- 规则匹配扩充及修改
- 关键要素识别模块
  - 应用其他领域数据集和模型，对该任务的数据集输入进行扩充
  - 与规则匹配进行特征互补
- 交互层的修改
  - 特征交互层更换更深层次的特征交互模块

改进方法

- 关键要素
  - 债权人转让债权，借款金额x万元，有借贷证明，贷款人系金融机构，返还借款，公司|单位|其他组织借款，连带保证，催告还款，支付利息
  - 订立保证合同，有书面还款承诺，担保合同无效|撤销|解除
  - 拒绝履行偿还，免除保证人保证责任，保证人不承担保证责任，质押人系公
  - 贷款人未按照约定的日期|数额提供借款，多人借款，债务人转让债务，约定利率不明。

应用总结

应用总结

- 算法的应用领域
  - 智慧法庭，辅助判案
  - 法律检索，类案匹配
- 未来工作
  - 更大的数据集，更多的文书种类
  - 实现类案推送等落地应用

参考文献

[1] Hong Z, Zhou Q, Zhang R, et al. Legal Feature Enhanced Semantic Matching Network for Similar Case Matching[C]//2020 International Joint Conference on Neural Networks (IJCNN). IEEE, 2020: 1-8.

[2] Jiang X, Ye H, Luo Z, et al. Interpretable rationale augmented charge prediction system[C]//Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations. 2018: 146-151.

[3] 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.

道德经

知人者智，自知者明。胜人者有力，自胜者强。知足者富。强行者有志。不失其所者久。死而不亡者，寿。

谢谢！