面向司法判决书的个案案情知识图谱构建

洪文兴[1]，胡志强[1]，翁 洋[2]

(1. 厦门大学航空航天学院，厦门 361005；2. 四川大学数学学院，成都 610064)

摘 要：在人工智能推动下的司法改革中，让机器通过前沿技术认知个案，是当前司法应用的前提和薄弱之处，实现机器的个案认知将会对相似案例检索、案例智能推送等各项司法应用带来新的思路与变革。为解决这个问题，我们提出了以深度学习方法为核心技术驱动的知识图谱解决方案，进行了面向海量的司法判决书的个案案情知识图谱构建的研究与实践。对于知识图谱构建涉及的两个重要的 NLP 任务，我们以预训练模型为基础进行了模型设计与优化。针对实体识别任务，我们对基准模型的解码输出层进行改进，可提升 0.36 的 F1；针对关系抽取任务，我们提出了一种多任务联合的语义关系抽取模型，相比基准模型，F1 提升高达 2.37；我们设计了个案案情知识图谱构建流程，并进行了个案案情图谱实验研究，验证了本文构建方案及流程的可行性和有效性，在“机动车交通事故责任纠纷”案由下，我们构建了一个大规模个案案情知识图谱，为下一步的司法应用提供了语义数据支撑。

关键词：个案案情；知识图谱；实体识别；关系抽取

Knowledge Graph Construction of Case for Legal Judgments

Hong Wenxing[1]，Hu Zhiqiang[1]，Weng Yang[2]

（1. School of Aerospace Engineering，Xiamen University，Xiamen 361005，China;

2. School of Mathematics，Sichuan University，Chengdu 610064，China）

Abstract:

作者简介：洪文兴（1980— ），男，博士，副教授，hwx@xmu.edu.cn.

通信作者：翁 洋，wengyang@scu.edu.cn.

基金项目：国家重点研发计划资助项目（2018YFC0830300）；福建省科技计划资助项目（2018H0035）；厦门市科技计划资助项目（3502Z20183011）；掌数金融科技研发基金资助项目.

Supported by the National Key R&D Program of China（No.2018YFC0830300），the Science and Technology Program of Fujian，China（No.2018H0035），the Science and Technology Program of Xiamen，China（No.3502Z20183011），the Fund of XMU-ZhangShu Fintech Joint Lab.

1 引言

在人工智能推动下的智慧法院建设中，面向海量的裁判文书资源库，让机器通过一定的技术手段学习与认知个案，是当前司法应用的前提和薄弱之处。实现机器自动学习与认知个案将会对相似案例检索、案例智能推送、裁判文书自动生成等一系列司法应用将产生重要影响。

当前以连接主义代表性的深度学习技术、以符号主义代表性的知识图谱技术正在得到广泛而深刻的研究，也将会对各个行业领域带来深刻的影响和变革。为此，我们以深度学习作为驱动技术，以知识图谱作为知识载体，实现面向司法判决书的个案案情知识图谱构建，以实现机器对个案的学习与认知。

知识图谱的概念由谷歌公司于 2012 年正式提出，谷歌以此技术为基础构建下一代智能化搜索引擎。现有具有代表性的大规模知识库包括：Freebase[1]、Wikidata[2]、DBpedia[3]、YAGO[4]、Zhishi.me[5]、CN-DBpedia[6]等，其中后两个属于专门的中文知识库。上述知识库数据基本都来源于开放社区或开放域的数据，属于通用知识图谱，对实际垂直领域应用的意义并不大。随着知识图谱研究的兴起，领域知识图谱的研究也逐渐得到重视，例如目前两个大型开放学术知识图谱 OAG[1]和 AceKG[2]，有益于对学术数据挖掘的研究和开发，此外，医疗、金融等领域也可见到知识图谱构建及应用之处。

目前面向垂直领域的知识图谱，数据来源主要还是（类）结构化的文本数据，面向非结构化文本的知识图谱构建研究的并不广泛。对于垂直领域非结构化的文本，采用开放信息抽取的方法并不可行，为此我们设计了有监督的实体识别-关系抽取串联的管道模型。针对实体识别任务，目前有效的方法还是基于深度学习的方法，此前主流的方法可分为基于 RNN 的方法[7][8]（如：LSTM-CRF），基于 CNN 的方法[9]以（如：ID-CNN）及混合模型的方法[10][11]（如：LSTM-CNN-CRF）。针对关系抽取任务，此前有效且主流的方法仍可分为基于 RNN 的方法[12][13]、基于 CNN 的方法[14][15][16]及其混合模型的方法[17]。

上述任务一般都会利用 Word2vec[18]等词向量工具训练好的词向量，但是这种词嵌入是静态的，无法解决一词多义的问题，随着预训练模型的研究的兴起，上述题得到解决。比较有代表性模型的包括：基于双层双向 LSTM 的 ELMo[20]、基于单向 Transformer 的 GPT[21]及基于双向 Transformer 并融合下一句任务的 BERT[22]。基于大规模文本进行无监督预训练可以充分学习其中蕴含的语义信息，通常都能直接提升现有的各项 NLP 任务。对于实体识别任务，谷歌的 BERT-Softmax[22]模型超越以往结果；对于关系抽取任务，采用预训练模型 GPT 并结合语言模型多任务的模型 TRE[23]达到了最好效果。

---

[1] https://www.openacademic.ai/oag/

[2] https://www.acemap.info/app/AceKG/

本文以海量的司法判决书为研究对象，研究目标是为每一份文书构建形成一份案例案情知识图谱。本文的主要贡献如下：

- 我们对实体识别基准模型进行改进获得 BERT-CRF 模型，使得实体识别效果 F1 进一步提升 0.36。
- 我们提出了一种多任务联合的语义关系抽取模型 BERT-Multitask，相比基准模型，关系抽取结果 F1 提升高达 2.37。
- 我们选取民事案由“机动车交通事故责任纠纷”，设计一个融合类结构化文本和非结构化文本的个案案情知识图谱构建方案，实验结果证明了该流程的可行性和有效性，并构建一个大规模的个案案情图谱，为类案检索、推送等下游任务提供了语义数据支撑。

2 实体识别模型

实体在知识三元组中以节点的形式呈现，是构成知识图谱的主体和基础。这部分展示了基于预训练模型 BERT[22]的基准模型 BERT-Softmax[22]及其联合 CRF 的改进后的 BERT-CRF 模型。

2.1 基准模型

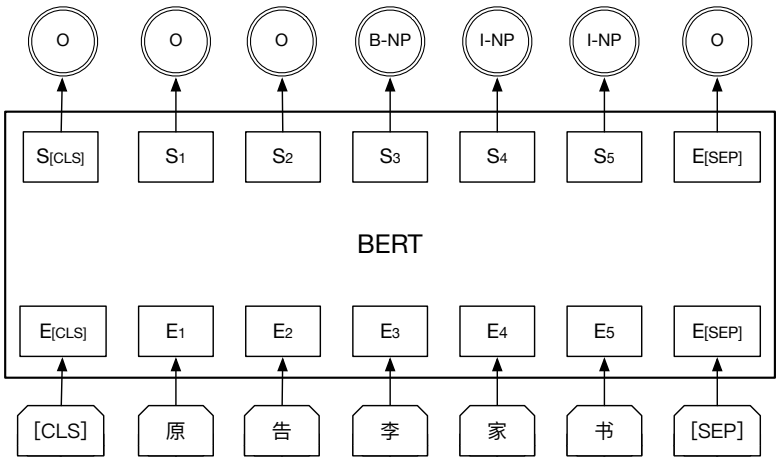

图 1：实体识别模型 BERT-Softmax

图 1 展示了实体识别的基准模型。对于文本输入片段“[CLS] 原 告 李 家 书 [SEP]”，其中，特殊标识符[CLS]和[SEP]分别作为句子序列的开始和结束，“李家书”作为一个自然人书主体，采用 BIO 表示法，则正确对应的预测标签序列应为“O O O B-NP I-NP I-NP O”，其中 NP 代表自然人主体实体类别。

模型整体可划分为三大网络层，分别是输入嵌入层、特征抽取层以及解码输出层。

输入嵌入层对输入的 token 序列进行向量空间的嵌入表示。每个 token 的空间嵌入表示组成包括对应的字符嵌入，位置嵌入及句子嵌入，公式表示为$E_t = E_c + E_p + E_s$。其中，E_t为该 token 的语义嵌入表示，E_c为该 token 的字符嵌入表示，E_p为该 token 所处的位置的嵌入表示，E_s为该 token 所处的句子的嵌入表示。

特征抽取层在 token 的向量空间嵌入的基础上实现更高层次的语义特征的抽取与表示，为每个 token 产生一个序列输出S_i。在各项任务表现优秀的 BERT 预训练模型得益于 Transfomer[25]模型的强大的特征抽取与编码能力。特征抽取层主要使用到 Transfomer 模型的 6 层的编码层，每个编码层主要由多头部注意力（Multi-head attention[25]）网络、前馈神经网络、残差网络（Residual network[26]）以及层标准化（Layer normalization[27]）模块组成。

解码输出层实现对每个 token 的序列输出进行标签预测。假设给定特征抽取层的序列输出 S:

$$S = (s_{[\mathrm{CLS}]}, s_1, \ldots, s_n, s_{[\mathrm{SEP}]})$$

S 经过线性投影，可以得到大小为$n \times m$的得分矩阵 P，其中，n 为输入序列的长度，m 为不同标签的数量，$P_{i,j}$对应该句中第 i 个 token 在第 j 个标签上的得分。对于序列中每个 token，可以应用 Softmax 得到该 token 预测标签的概率分布，取概率分布最大值所在的索引对应的标签作为预测的标签。其中，第 i 个 token 在第 j 个标签上的概率计算公式为（首尾 token 都会标记为 O）:

$$p(y_j|t_i) = e^{P_{iy_j}} \Big/ \sum_{j=0}^{m} e^{P_{iy_j}}$$

2.2 改进模型

上述基准模型存在一个问题：上层解码预测输出层采用 Softmax，序列的预测输出标签彼此独立，实际上，实体的输出标签是存在一定的约束关系的。例如：输出标签 I-MV 不能跟随 B-NP，其中，MV 代表机动车实体类别，NP 代表机动车实体类别。因此，我们采用条件随机场（Conditional random field[28]，CRF）作为解码输出层，以此来解决这个问题。

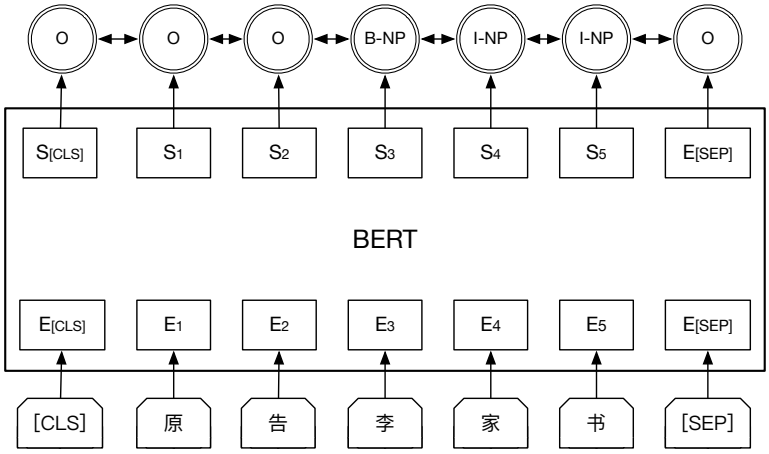

图 2：实体识别模型 BERT-CRF

图 2 展示了改进后的网络模型 BERT-CRF，和基准模型的唯一区别就在于解码输出层的不同。假设给定输入序列和预测输出标签序列 Y:

$$Y = (y_{[\mathrm{CLS}]}, y_1, \cdots, y_n, y_{[\mathrm{SEP}]})$$

结合上述特征抽取得到的序列输出 S，可以定义综合得分函数：

$$f(S,Y) = \sum_{i=1}^{n} A_{y_i,y_{i+1}} + \sum_{i=1}^{n} P_{i,y_i}$$

其中，A 是输出标签之间的转移得分矩阵，其中，$A_{i,j}$对应标签 i 到标签 j 的得分。

对输入序列所有可能的输出标签应用 Softmax，得到预测标签序列 Y 的概率：

$$p(Y|S) = e^{f(S,Y)} / \sum_{\tilde{Y} \in Y_X} e^{f(S,\tilde{Y})}$$

我们需要使得综合得分最大化，一般取真值标签序列的对数概率：

$$\log\big(p(Y|S)\big) = f(S,Y) - \log\Big(\sum_{\tilde{Y} \in Y_X} e^{f(S,\tilde{Y})}\Big)$$

其中，Y_X代表输入序列对应所有可能的标签序列。

根据上述公式，最大得分对应的输出标签序列即为最优的预测标签序列。

$$Y^* = \underset{\tilde{Y} \in Y_X}{\operatorname{argmax}} f\big(S,\tilde{Y}\big)$$

一般仅考虑任意两个标签之间的转移关系，上述最优解可以采用动态规划进行求得，我们采用维特比算法（Viterbi algorithm[29]）进行求解。

3 关系抽取模型

关系在知识三元组中以边的形式呈现，二元关系是一个三元组的语义核心。启发于 GPT[30]模型，引入语言模型作为一个辅助目标可以提高模型泛化性能并加快收敛，我们引入知识三元组的平移嵌入（Translating Embedding[31]）的任务作为辅助优化任务。

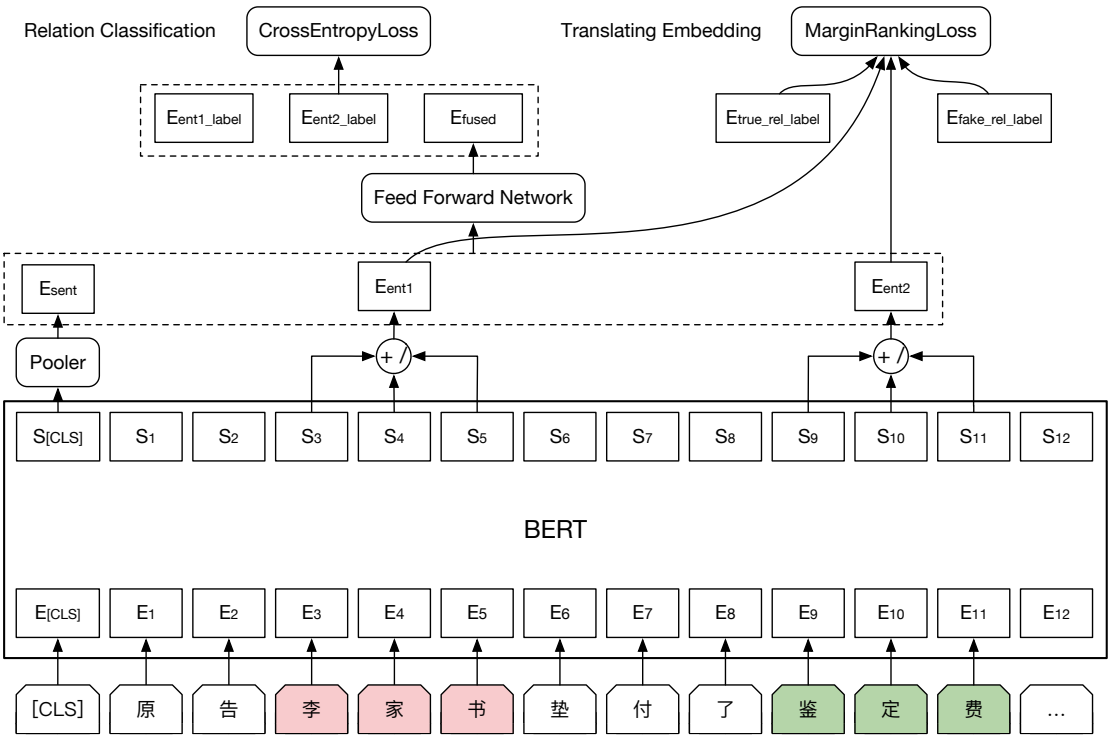

图 3：一种多任务联合的语义关系抽取模型

图 3 展示了一种融合关系分类和平移嵌入两种任务联合的语义关系抽取模型 BERT-Multitask。对于给定的自然人主体类实体 1“李家书”、人身损害赔偿项目类实体 2“鉴定费”以及这两个实体出现的句子“原告李家书垫付了医疗费和…”，目标任务是判断这两个实体存在预定义关系中的哪一类，结合上下文语境，正确的预测标签应为“遭受”关系类别。其中位于句子起始特殊标识符[CLS]用于对整个句子做嵌入，这里不再需要对句子末尾进行标识。

这里的输入嵌入层和 BERT 特征抽取层和 2.1 节实体识别任务中介绍的基本一致，对于输入的每个 token，我们可以得到更高层次，更加丰富的语义空间的嵌入表示，同样可以得到输入序列的包含语义特征的序列输出。

$$S = (s_{[\mathrm{CLS}]}, s_1, \dots, s_n)$$

在得到序列输出 S 的基础上，进一步的映射和处理可以得到句子的嵌入、实体 1 的嵌入以及实体 2 的嵌入，为下一步特征融合和多任务联合学习作准备。

对于句子嵌入的获取，取第一个 token（即[CLS]）的嵌入表示，输入 Pooler 层，即可得到句子的嵌入表示E_{sent}，用公式可表示为：

$$E_{sent} = \tanh\left(S_{sent} W_{sent}\right)$$

其中，W_{sent}为权重参数，进行相同维度的映射。

对于实体嵌入的获取，我们对组成一个实体的所有的 token 序列输出取平均值作为该实体的嵌入表示，公式表示为：

$$E_{ent} = \frac{1}{N} \sum_{k}^{N+k-1} s_i$$

其中，N 为实体序列的长度，k 为实体序列起始的索引。

我们对句子嵌入E_{sent}、实体 1 嵌入E_{ent1}、实体 2 嵌入E_{ent2}进行拼接，输入到前馈神经网络，进行特征融合，可以获得融合特征E_{fused}，计算公式为：

$$E_{fused} = \mathrm{gelu}((E_{sent} \oplus E_{ent1} \oplus E_{ent2}) W_{fused} + b_{fused})$$

其中，W_{fused}为权重参数，b_{fused}为偏置参数。

对于关系分类任务，考虑到特定的关系只会发生在特定的两个实体类别之间，我们将实体 1 的类别标签嵌入E_{ent1_label}、实体 2 的类别标签嵌入E_{ent2_label}与混合特征进行拼接得到最终的嵌入特征$E_{final} = E_{ent1_label} \oplus E_{ent1_label} \oplus E_{sent}$，输入到输出层 Softmax，以进行输出标签 y 的预测。

$$P\left(y \middle| E_{final}\right) = \mathrm{softmax}(E_{final} W_{final})$$

分类任务中，我们需要使得下列目标最小化：

$$\mathcal{L}_1 = -\sum\nolimits_{E_{final} \in E_X} \log P\left(y \middle| E_{final}\right)$$

知识三元组平移模型 TransE 的基本思想是：对于给定三元组(h, r, t)的集合 Set，由

两个实体$h,r \in E$（实体集合）和一条关系$r \in R$（关系集合）组成，当(h,r,t)成立时，有$h+r \approx t$，否则$h+r$与 t 相离应尽可能的远，使用$d(h,r,t)$来对它们之间的距离进行度量，可以采用L_1范数或L_2范数。

对于一条训练数据，我们总有实体 1、实体 2、它们所同时所在的句子以及它们之间真实的关系标签，考虑到这样得到的关系三元组都是正样例，为解决没有负样例的问题，我们为每条数据中随机从关系类别中随机选择一个非真实关系类别和实体对组成一个负样例。假定真实关系标签的嵌入为$E_{true_rel_label}$，假关系标签的嵌入为$E_{fake_rel_label}$，采用基于距离的排序方法得到最小化的优化目标：

$$\mathcal{L}_2 = \sum_{X \in D} [\gamma + d(E_{ent1} + E_{true\_rel\_label}, E_{ent2}) - d(E_{ent1} + E_{fake\_rel\_label}, E_{ent2})]_+$$

其中，$d\left(E_{ent1} + E_{rel\_label}, E_{ent2}\right) = |E_{ent1} + E_{rel_{label}} - E_{ent2}|$，$[x]_+ = \max(x, 0)$，$\gamma > 0$为待优化的间隔超参数。

综合考虑关系分类任务和知识三元组平移嵌入任务，我们可以得到总的优化目标损失函数为：

$$\mathcal{L} = \mathcal{L}_1 + \lambda * \mathcal{L}_2$$

其中，$\lambda > 0$为平移嵌入任务损失的权重。

4 实验

本节将在实际的司法场景中，通过一系列实验来验证上述模型的有效性及整个技术方案的可行性。第 4.1 节介绍了研究数据的准备工作，第 4.2 节展示了实体识别的实验结果及分析，第 4.3 节展示了关系抽取的实验结果及分析，第 4.4 节介绍了案件案情知识图谱自动构建的结果。

4.1 数据准备

4.1.1 预定义实体和关系

我们以民事案件“机动车交通事故责任纠纷”案由下的一审判决书为研究对象，组织高校及相关司法企业的法律专家参与研究和讨论，并结合实际的判决书和现行的法律法条，确定该案由下司法判决书中普遍存在并有重要意义的实体和关系。

最终，我们在该案由下预定义了 20 类实体类型如表 1 所示，考虑到违法行为种类过于繁多，而且大多数违法行为在文书中出现频次较低，我们选择了比较常见的 9 类违法行为进行研究，编号对应为 12-20。

表 1：预定义实体类型

#	实体类别	#	实体类别
1	自然人主体	11	财产损失赔偿项目
2	非自然人主体	12	未取得驾驶资格

3	机动车	13	饮酒后驾驶
4	非机动车	14	醉酒驾驶
5	保险类别	15	超载
6	责任认定	16	超速
7	一般人身损害	17	违反道路交通信号灯
8	伤残	18	违法变更车道
9	死亡	19	不避让行人
10	人身损害赔偿项目	20	行人未走人行横道或过街设施

我们预定义了 9 类关系类型如表 2 所示，一个关系类型可能会对应多个关系类型对，例如对于编号 1 的“驾驶”关系，就会存在（自然人主体 驾驶 机动车）和（自然人主体 驾驶 非机动车）这 2 条概念层知识三元组，合计得到 30 条概念层知识三元组（这里不考虑“其他”类关系类型）。

表 2：预定义关系类型

#	关系类别	关系对应的实体对	三元组数量
1	驾驶	自然人主体→[机动车|非机动车]	2
2	所有	自然人主体→[机动车|非机动车] 非自然人主体→[机动车|非机动车]	4
3	搭乘	[机动车|非机动车]→自然人主体	2
4	投保	机动车→保险类别	1
5	实施	自然人主体→[常见 9 类违法行为]	9
6	发生事故	机动车→[机动车|非机动车|自然人主体] 非机动车→[机动车|非机动车|自然人主体]	6
7	承担	自然人主体→责任认定	1
8	遭受	自然人主体→[一般人身损害|伤残|死亡|人身损害赔偿项目|财产损失赔偿项目]	5
9	其他	—	—

4.1.2 数据预处理及标注划分

考虑到中国东、中、西三大区域经济及技术发展水平的差异，司法能力水平和判决书写作规范也会存在一些区别。我们在东部选取“江苏省”、“浙江省”，中部选取“河南省”、“湖北省”，西部选取“四川省”、“云南省”以解决地域差异性带来的影响。每个省份随机选取 100 份判决书，合计获取 600 份判决书作为研究的原始文书数据。

对上述选取的原始判决书进行数据预处理。首先需要对文书进行段落类型标记，考虑到文书的用语表达和写作格式的规范性，我们采用基于规则的方法对文书的段落类型进行标记。本文知识图谱构建来源的文本段落类型包括“当事人信息”类型的类结构化文本和“法院认定事实”类型的非结构化文本，我们随机选择 500 篇判决书进行规则预标注

并交由人工审核，评估得到基于规则的这两种段落类型的分段效果的 F1 值分别为 99.85 和 90.34。

基于提取的类结构化文本，我们利用基于规则的方法对涉及到的民事主体进行提取，以获取案件的基本信息及用于案情事实中的“原告”和“被告”的补全处理。

我们采用开源的标注工具 brat[32]进行部署与配置，以实现多人在线进行实体和关系的标注，将法院认定事实文本以句号进行句子划分并导入标注系统，标注示例截图如图 4 所示。

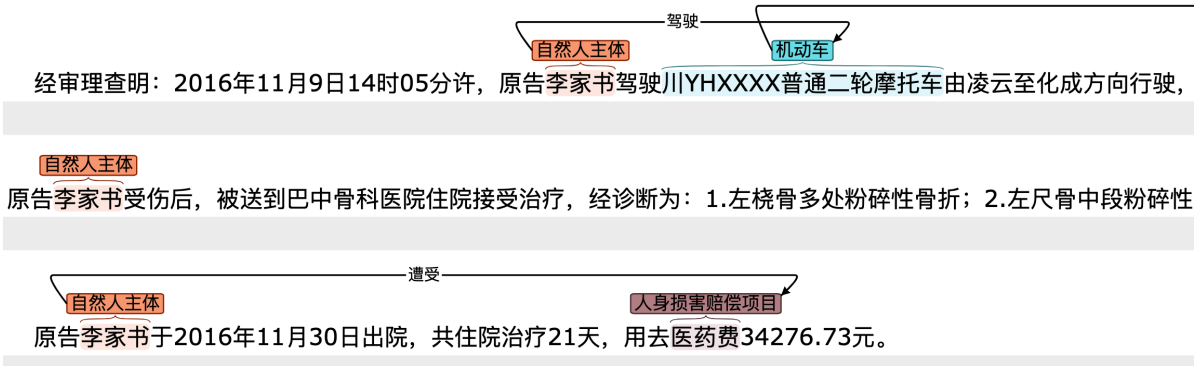

图 4：brat 在线标注示例截图

经过人工标注及审核，去除长度过短及质量很差的案例，最终获得 585 份案例的法院认定事实文本。为避免句子层级划分数据影响客观评价，我们在案例层级进行数据集的划分，数据集划分情况如表 3 所示，实体任务和关系任务数据量的统计都是针对句子层级。在建立关系任务数据集时，对于不存在关系且在关系实体对定义内的的两个实体，我们以 0.5 的概率选择“其他”关系标签组成一条“负样例”。

表 3：标注案例数据集划分

数据集	案例数量	实体任务数据量	关系任务数据量
训练集	430	5681	9756
验证集	55	744	1115
测试集	100	1313	2314

4.2 实体识别

本实验环境在 Tesla T4 16GB GPU 环境下进行，使用 PyTorch 框架进行开发，在基本不损失精度的前提下，为了减少 GPU 的内存开销，加快训练速度，我们在程序设计中融入了一项称之为 apex[3]的混合精度训练（Mixed precision training[33]）技术，BERT 模型使用的是 PyTorch 的实现[4]（关系抽取任务与此一致）。

表 4 展示了本实验涉及到的一些重要的超参数设置，主要是根据先前的工作及实际经验调试，并没有进行严格的网格搜索。BERT-Softmax 和 BERT-CRF 模型的参数除了

[3] https://github.com/NVIDIA/apex
[4] https://github.com/huggingface/pytorch-pretrained-BERT

初始学习率不一致，前者为 2e-5，后者为 1e-5，其余参数设置相同。句子的最大长度设为 400，超过此长度的以标点切分两段处理；在权重参数添加系数为 0.01 的 L2 正则化项（偏置项不添加），且在顶层的线型层的 dropout[34]设置为 0.1，以避免过拟合；小批量大小设为 16；最大梯度设为 2.0。

表 4：实体识别超参数设置

参数	取值	参数	取值
max length	400	batch size	16
learning rate	2e-5/1e-5	gradient clipping	2.0
weight decay	0.01	linear dropout	0.1

表 5：不同模型在实体识别任务中的表现

模型	准确率	召回率	**F1**
BERT-Softmax	93.89	94.85	94.37
BERT-CRF	93.62	95.86	**94.73**

如表 5 所示，在测试集上，改进模型 BERT-CRF 相比基准模型 BERT-Softmax，准确率下降 0.27，召回率上升 1.01，F1 值有 0.36 的提升。综合来看，BERT-CRF 模型还是要优于 BERT-Softmax，表现出利用标签之间的转移约束关系可以使得实体识别的效果进一步得到微提升。

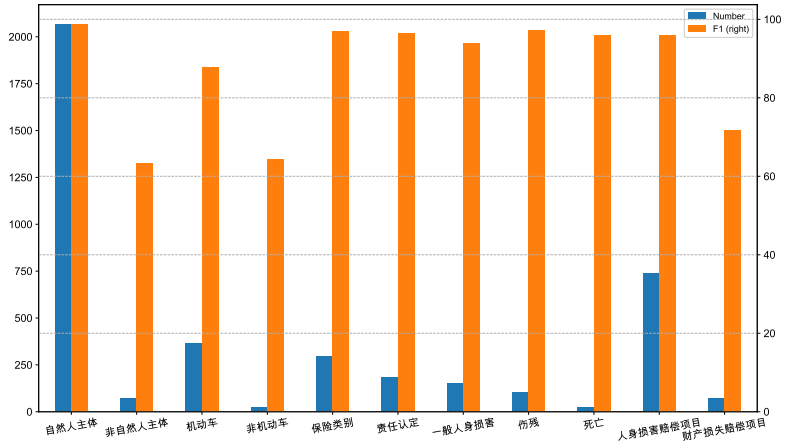

图 5：各实体类别的数量及对应的 F1 值（Number≥20）

基于较优的 BERT-CRF 模型，测试集上各实体类别的实体数量及对应 F1 值如图 5 所示，由于 9 类常见违法行为实体类别出现的数量均低于 20，故不参与统计。统计结果表明：在实体数量大于等于 20 个的 11 类实体上，F1 值在 95 以上的有 6 类，90 以上的有 7 类，85 以上的有 8 类，整体效果表现良好。但是对于非自然人主体、非机动车及财产损失赔偿项目实体类的表现就稍差一些，分析原因，首先这三类实体数量比较少，数据的不充分导致模型欠学习，死亡类实体数量虽然很少，但由于表达很固定，如“死亡”及“致死”等，因而也能够获得不错的学习效果；第二个原因在于实体表达的多样性，比如非机动车和财产损失赔偿项目表达的形式很多样，例如财产损失可能会涉及到各种物

品损失的表达，导致模型学习难度较大。而对于非自然人主体类实体，目前只是标注一些跟机动车相关的租赁公司、运输公司等，文本中还会出现许多汽车相关但是却不会产生关系的服务公司、维修公司等，这样导致其他类公司或组织对模型影响就比较大。

4.3 关系抽取

图 5 展示了关系抽取任务中一些重要的超参数设置，一些同名参数的意义和实体任务中参数介绍一致。特别地，实体标签的嵌入维度设为 128，关系标签的嵌入维度设为 768，平移嵌入任务的损失权重取值为 1e-5。

表 5：关系抽取超参数设置

参数	取值	参数	取值
max length	400	batch size	16
learning rate	2e-5	entity emb size	128
weight decay	0.01	rel emb size	768
batch size	16	λ	1e-5

参照 SemEval-2010 Task 8[35]多关系分类任务官方评测标准，我们取宏平均（Macro-averaged）的准确率、召回率及 F1 进行效果评估，唯一区别的是，学术研究标准任务评测未考虑“其他”关系类别，考虑到模型要投入实际应用与更加客观的评价，我们将“其他”关系类别也一并考虑。实验表明，“其他”类关系抽取的效果往往要低于整体评估的效果。

如表 6 所示，在测试集上，改进后的模型 BERT-Multitask 模型相比基准模型 BERT-Softmax，准确率、召回率及 F1 都获得了全面提升，综合指标 F1 提升高达 2.37，表明融入平移嵌入任务的多任务联合的关系抽取模型能够明显改善关系抽取的效果。在模型 BERT-Multitask 训练完成后，利用该模型可以得到一件非常有价值的副产物，即一种融合了上下文环境及知识三元组语义关系的实体和关系的向量嵌入表示。

表 6：不同模型在关系识别任务中的表现

模型	准确率	召回率	**F1**
BERT-Base	88.57	90.98	89.27
BERT-Multitask	**91.53**	**92.51**	**91.64**

基于表现更好的 BERT-Multitask 模型，测试集上各个关系类别的数量及表现如图 6 所示。9 类关系类别中，F1 在 95 以上的有 5 类，90 以上的有 6 类，85 以上的有 7 类，综合表现良好。负样例“其他”关系类的 F1 值为 91.05，低于综合表现 91.64，结合先前的一些工作，表明负样例“其他”关系类效果往往要低于平均水平。而对于“搭乘”和“发生事故”这两类关系的抽取效果则要表现较差一些，经过分析，发现这两类关系的实体相隔距离往往较远，而且这两类的数据量相对较少。

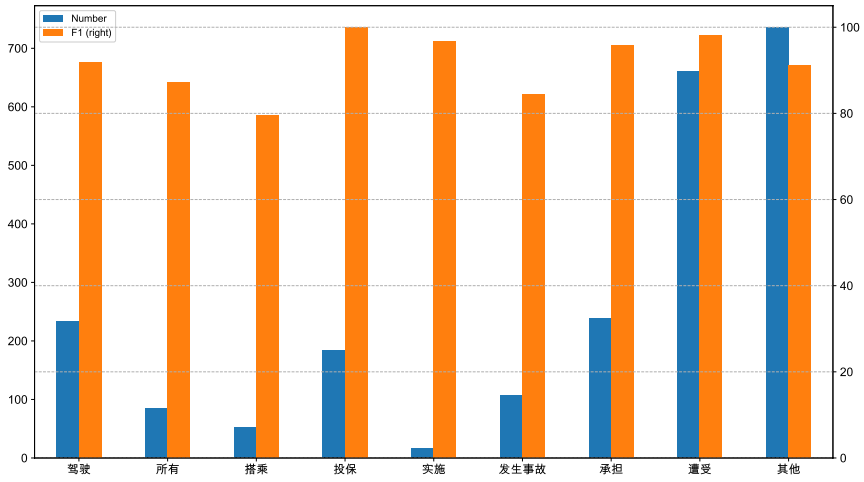

图 6：各关系类别数量及对应的 F1 值

4.4 个案案情知识图谱构建

个案案情知识图谱构建的文本类型包括类结构化文本以及非结构化文本，类结构的文本涉及的段落类型包括：“文书标题”、“案号”、“受理法院”及“当事人信息”，非结构化的文本涉及的段落类型只涉及“法院认定事实”。个案案情知识图谱构建流程图如图 7 所示。

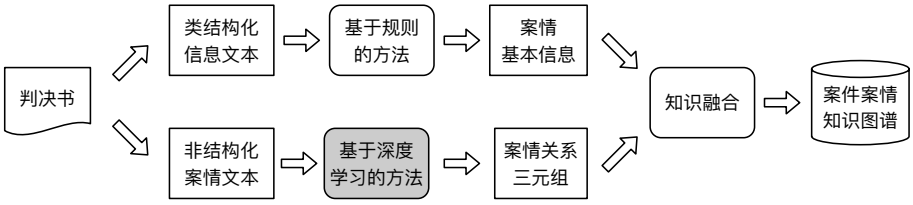

图 7：个案案情知识图谱构建流程图

类结构化文本在形式上类似于维基百科中的信息框（Infobox），考虑到文书写作格式的规范性，我们采用基于规则的方法进行抽取；利用前面学习到的实体识别模型和关系抽取模型组成一个串联的管道模型，以完成非结构化案情事实文本的知识三元组的提取。整个案情知识图谱构建步骤如下：

Step 1：分段标记。给定一篇司法判决书 *Doc*，采用基于规则的方法进行分段标记，识别出上述定义的类结构化文本$Text_1$和非结构化案情事实文本$Text_2$；

Step 2：类结构化信息抽取。基于规则对类结构化文本进行抽取，获得“文书标题”、“案号”、“受理法院”作为“案件”实体的属性信息，从“当事人信息”类文本中抽取民事主体基本信息 *Info*，涉及名称及其委托代理人信息等；

Step 3：数据预处理。对$Text_2$进行文本预处理，涉及原被告的指代补全及分句处理等，获得句子列表$List_1$；

Step 4：实体识别。基于学习的实体识别模型 BERT-CRF，对$List_1$逐一进行实体识别，获得实体数据列表$List_2$，每条数据包含：句子及实体（包含类别）集合；

Step 5：关系抽取。对$List_2$每一条数据所含实体在预定义关系实体对范围内进行组

合形成关系数据列表$List_3$，每条数据包含：实体 1 及其类别，实体 2 及其类别，所在句子。利用学习的关系抽取模型 BERT-Multitask 对$List_3$逐一进行关系抽取，最终获得案情事实三元组 *Triples*；

Step 6：知识融合。由于关系类别是标准的预定义，知识主要在于实体对齐。主要是基于一些规则的方法：例如利用两实体之间的固定表达制定规则，“如下简称”、“简写为”等类似表达；根据实体自身的特点，“川 A×××××号小轿车”与“川 A×××××号”在同类实体类别“机动车约束”下及客车货车约束，可根据车牌号进行对齐处理，基于类似的规则处理得到案情知识 *Knowledge*。

Step 7：知识入库与可视化。将 *Knowledge* 写入图数据库存储并进行可视化展示。

新输入“机动车交通事故责任纠纷”案由下的一份判决书，通过上述流程自动生成的案例案情知识图谱如图 8 所示。

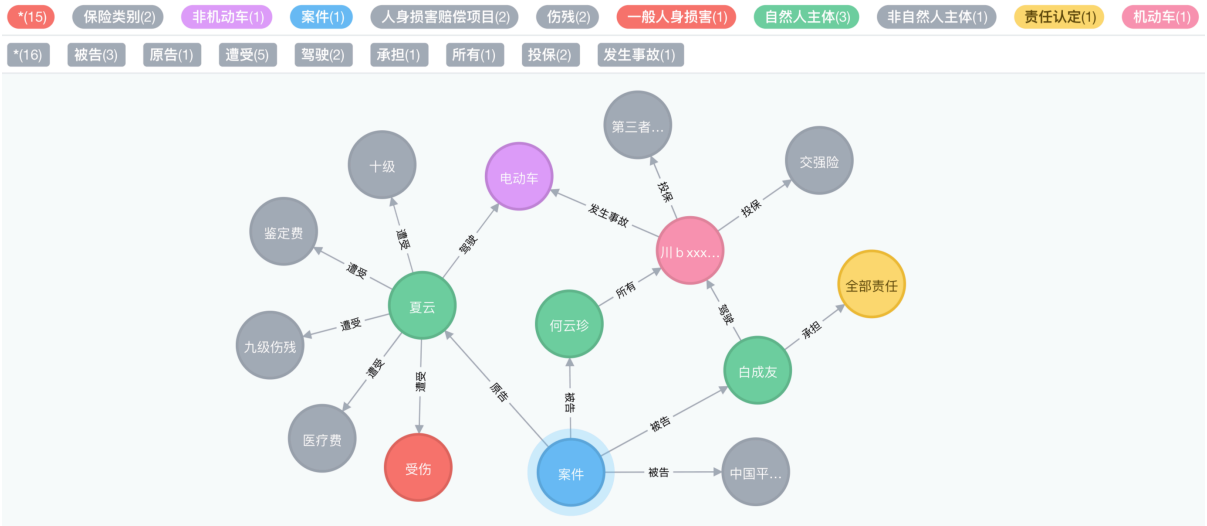

图 8：个案案情知识图谱自动构建示例

5. 结语

本文致力于面向司法判决书的个案案情知识图谱构建的研究。对于知识图谱构建涉及的两个重要的自然语言处理问题进行了重点研究，针对实体识别任务，对基准模型进行改进获得 BERT-CRF 模型，使得实体识别效果进一步获得微提升；针对关系抽取任务，我们提出了一种多任务联合的语义关系抽取模型 BERT-Multitask，明显改善了关系抽取的效果。最后，我们设计了一个融合了类结构化文本和非结构化文本（管道模型）的知识抽取流程，并通过实验对技术流程的可行性和有效性进行了论证，并构建了一个大规模的个案案情知识图谱。

本文研究的司法判决书案情图谱构建的一个重要的前期工作是段落类型标记任务，目前主要是基于规则的方法，非结构化案情事实文本的提取效果相对较差，下一步将结合一些机器学习的方法提升该任务的效果；为进一步提升与评价案情图谱构建质量，下一步将继续加大数据标注规模，算法与构建流程中应充分考虑法律知识并结合文书自身

特点，建立合理的个案案情知识图谱构建质量评价体系。接下来将基于已构建的个案案情司法知识图谱，进行相似案例检索或推荐的应用研究。

参考文献

[1] Bollacker K, Evans C, Paritosh P, et al. Freebase: a collaboratively created graph database for structuring human knowledge[C]//Proceedings of the 2008 ACM SIGMOD international conference on Management of data. AcM, 2008: 1247-1250.

[2] WMF. Wikidata[EB/OL]. [2019-05-23]. https://www.wikidata.org/wiki/Wikidata:Main_Page.

[3] Lehmann J, Isele R, Jakob M, et al. DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia[J]. Semantic Web, 2015, 6(2): 167-195.

[4] Suchanek F M, Kasneci G, Weikum G. YAGO: A core of semantic knowledge unifying WordNet and Wikipedia[C]//Proceedings of WWW. 2007, 7: 697-706.

[5] Niu X, Sun X, Wang H, et al. Zhishi. me-weaving chinese linking open data[C]//International Semantic Web Conference. Springer, Berlin, Heidelberg, 2011: 205-220.

[6] Xu B, Xu Y, Liang J, et al. CN-DBpedia: a never-ending Chinese knowledge extraction system[C]//International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems. Springer, Cham, 2017: 428-438.

[7] Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv preprint arXiv:1508.01991, 2015.

[8] Lample G, Ballesteros M, Subramanian S, et al. Neural architectures for named entity recognition[J]. arXiv preprint arXiv:1603.01360, 2016.

[9] Strubell E, Verga P, Belanger D, et al. Fast and accurate entity recognition with iterated dilated convolutions[J]. arXiv preprint arXiv:1702.02098, 2017.

[10] Ma X, Hovy E. End-to-end sequence labeling via bi-directional lstm-cnns-crf[J]. arXiv preprint arXiv:1603.01354, 2016.

[11] Chiu J P C, Nichols E. Named entity recognition with bidirectional LSTM-CNNs[J]. Transactions of the Association for Computational Linguistics, 2016, 4: 357-370.

[12] Zhang S, Zheng D, Hu X, et al. Bidirectional long short-term memory networks for relation classification[C]//Proceedings of the 29th Pacific Asia conference on language, information and computation. 2015: 73-78.

[13] Zhou P, Shi W, Tian J, et al. Attention-based bidirectional long short-term memory networks for relation classification[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2016, 2: 207-212.

[14] Zeng D, Liu K, Lai S, et al. Relation classification via convolutional deep neural network[J]. 2014.

[15] Huang X. Attention-based convolutional neural network for semantic relation extraction[C]//Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. 2016: 2526-2536.

[16] Wang L, Cao Z, De Melo G, et al. Relation classification via multi-level attention cnns[J]. 2016.

[17] Zhang X, Chen F, Huang R. A Combination of RNN and CNN for Attention-based Relation Classification[J]. Procedia computer science, 2018, 131: 911-917.

[18] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]//Advances in neural information processing systems. 2013: 3111-3119.

[19] Peters M E, Ammar W, Bhagavatula C, et al. Semi-supervised sequence tagging with bidirectional language models[J]. arXiv preprint arXiv:1705.00108, 2017.

[20] Peters M E, Neumann M, Iyyer M, et al. Deep contextualized word representations[J]. arXiv preprint arXiv:1802.05365, 2018.

[21] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training[J]. URL https://s3-us-west-2. amazonaws. com/openai-assets/research-covers/languageunsupervised/language understanding paper. pdf, 2018.

[22] 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.

[23] Alt C, Hübner M, Hennig L. Improving Relation Extraction by Pre-trained Language Representations[J]. 2018.

[24] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training[J]. URL https://s3-us-west-2. amazonaws. com/openai-assets/research-covers/languageunsupervised/language understanding paper. pdf, 2018.

[25] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in neural information processing systems. 2017: 5998-6008.

[26] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[27] Lei Ba J, Kiros J R, Hinton G E. Layer normalization[J]. arXiv preprint arXiv:1607.06450, 2016.

[28] Lafferty J, McCallum A, Pereira F C N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[J]. 2001.

[29] Forney G D. The viterbi algorithm[J]. Proceedings of the IEEE, 1973, 61(3): 268-278.

[30] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training[J]. URL https://s3-us-west-2. amazonaws. com/openai-assets/research-covers/languageunsupervised/language understanding paper. pdf, 2018.

[31] Bordes A, Usunier N, Garcia-Duran A, et al. Translating embeddings for modeling multi-relational data[C]//Advances in neural information processing systems. 2013: 2787-2795.

[32] Stenetorp P, Pyysalo S, Topić G, et al. BRAT: a web-based tool for NLP-assisted text annotation[C]//Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2012: 102-107.

[33] Micikevicius P, Narang S, Alben J, et al. Mixed precision training[J]. arXiv preprint arXiv:1710.03740,

2017.

[34] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. The Journal of Machine Learning Research, 2014, 15(1): 1929-1958.

[35] Hendrickx I, Kim S N, Kozareva Z, et al. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals[C]//Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions. Association for Computational Linguistics, 2009: 94-99.