文章编号：1003-0077（2019）00-0000-00

面向司法案件的案情知识图谱自动构建

洪文兴[1]，胡志强[1]，翁洋[2]，张恒[3]，王竹[4]，郭志新[5]

（1. 厦门大学 航空航天学院，福建 厦门 361102；2. 四川大学 数学学院，四川 成都 610065；
3. 成都星云律例科技有限责任公司，四川 成都 610036；4. 四川大学 法学院，四川 成都 610207；
5. 电子科技大学 公共管理学院，四川 成都 611731）

摘要：以法学知识为中心的认知智能是当前司法人工智能发展的重要方向。本文提出了以自然语言处理（NLP）为核心技术的司法案件案情知识图谱自动构建技术。以预训练模型为基础，对涉及的实体识别和关系抽取这两个 NLP 基本任务进行了模型研究与设计。针对实体识别任务，对比研究了两种基于预训练的实体识别模型；针对关系抽取任务，提出融合平移嵌入的多任务联合的语义关系抽取模型，同时获得了结合上下文的案情知识表示学习。在“机动车交通事故责任纠纷”案由下，和基准模型相比，实体识别的 F1 值可提升 0.36，关系抽取的 F1 值提升高达 2.37。以此为基础，本文设计了司法案件的案情知识图谱自动构建流程，实现了对数十万份判决书案情知识图谱的自动构建，为类案精准推送等司法人工智能应用提供语义支撑。

关键词：司法案件；知识图谱；实体识别；关系抽取

中图分类号：TP391 **文献标识码**：A

Automated Knowledge Graph Construction for Judicial Case Facts

HONG Wenxing[1], HU Zhiqiang[1], WENG Yang[2], ZHANG Heng[3], WANG Zhu[4], GUO Zhixin[5]

(1. School of Aerospace Engineering, Xiamen University, Xiamen, Fujian 361102, China;
2. School of Mathematics, Sichuan University, Chengdu, Sichuan 610065, China;
3. Galawxy Inc. , Chengdu, Sichuan, 610036, China;
4. School of Law, Sichuan University, Chengdu, Sichuan 610207, China;
5. School of Public Affairs and Administration, University of Electronic Science and Technology of China, Chengdu, Sichuan 611731, China)

Abstract：Legal knowledge centered cognitive intelligence is an important role for judicial artificial intelligence. This paper proposes an automated knowledge graph construction approach for judicial case facts with natural language processing (NLP) technology. Based on the pre-training model, models for two fundamental NLP tasks involving entity recognition and relation extraction are presented. For the entity recognition task, two pre-training based entity recognition models are compared. For the relation extraction task, a multi-task joint semantic relation extraction model is proposed incorporating translating embeddings. The knowledge representation learning of case facts is obtained while completing the relation extraction task. For "motor vehicle traffic accident liability dispute", compared with the baseline model, the entity recognition F1 score can be increased by 0.36, and the relation extraction F1 score is increased by 2.37. Based on the above approach, the case facts knowledge graphs are established for tremendous number of judicial documents which provide semantic computing for judicial artificial intelligence applications such as case retrieval.

Keywords：judicial case; knowledge graph; entity recognition; relation extraction

收稿日期：2019-07-31 **定稿日期**：2019-08-15

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)

0 引言

在人工智能推动下的司法改革当中，面向海量的裁判文书资源库，让机器通过一定的前沿技术认知案件，是当前人工智能司法应用的前提和薄弱之处。实现机器自动学习与认知案件将会对相似案例检索、类案精准推送、裁判文书自动生成等一系列司法应用产生重要影响。

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，实现面向司法案件的案情知识图谱自动构建，以实现机器对案件的认知。

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。随着知识图谱研究热潮的兴起，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[7]和 AceKG[8]，将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的构建及应用之处。

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。对于垂直领域的非结构化文本，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循环神经网络 RNN 的方法[9-10]（如：LSTM-CRF），基于卷积神经网络 CNN 的方法[11]（如：ID-CNN）及混合模型的方法[12-13]（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[14-15]、基于 CNN 的方法[16-18]及其混合模型的方法[19]。

上述任务一般都会利用 Word2vec[20]等词向量工具进行词向量表征，但是这种方式得到的词嵌入是静态的，无法解决一词多义的问题，随着预训练模型研究的兴起，上述问题得到了解决。比较有代表性模型的包括：基于双层双向 LSTM[21] 的模型 ELMo[22]、基于单向 Transformer[23] 的模型 GPT[24] 以及基于双向 Transformer 并融合下一句预测任务的模型 BERT[25]。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[25]模型超越以往的结果；对于关系抽取任务，采用预训练模型 GPT 并结合语言模型的多任务模型 TRE[26]达到了最好效果。

司法裁判文书记载了人民法院的审理过程和结果，相比网络百科，新闻资讯文本，裁判文书的特点主要包括：文书制作的合法性，文本必须依法制作，这是基本前提；形式的程序性，表现在结构固定化和用语成文化；语言的精确性，表现在语义表达单一，准确精当。相比裁定书，由于判决书的数量占比大，案件事实与裁判说理记录更加详实，对于司法的技术研究更有价值。

司法判决书主要包括类结构化的案件基本信息和非结构化的文本。类结构化的案件基本信息反映了案件发生的主体，这也是案情事实的基础。非结构化文本类型主要包括案件当事各方的陈述、法院认定事实、法院说理及裁判结果这三种段落类型，案件当事各方的陈述虽描述了一定的客观事实，但由于都带有一定的主观性，可能会存在陈述矛盾的事实；法院说理及裁判结果集中在依据法律规范进行裁判主文的论证；法院认定事实，基于案件审理中举证质证情况，描述了影响案件裁判结果的事实。因此，以案件基本信息为基础，围绕法院认定事实文本进行案情知识图谱构建是非常有必要与合理的。

本文以“机动车交通事故责任纠纷”案由下的司法判决书为研究对象，研究目标是为每一份文书自动构建形成一份案情知识图谱。本文的主要贡献如下：

(1) 我们对比研究了两种基于 BERT 的实体识别模型，实验表明解码输出层采用 CRF 可使得

实体识别效果进一步提升 0.36。

(2) 我们提出了一种融合平移嵌入的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 值提升高达 2.37。

(3) 我们设计了一个融合类结构化文本和非结构化文本的案件案情知识图谱自动构建流程，结果验证了该流程的可行性与有效性，并构建一个大规模司法案件的案情知识图谱，为类案精准推送等下游任务提供了语义支撑。

1 实体识别模型

1.1 基准模型

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。图 1 展示了实体识别的基准模型。对于文本输入片段的 token 序列“[CLS] 原 告 李 家 书 [SEP]”，其中，特殊标识符[CLS]和[SEP]用于标识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，则对应的正确预测标签序列应为“O O O B-NP I-NP I-NP O”，其中 NP 代表自然人主体实体类别。

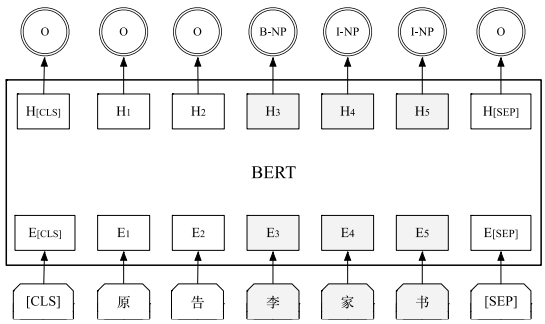

图 1 实体识别模型 BERT-Softmax

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（词）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 的综合嵌入表示，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 产生一个隐含状态H_i。在各项任务表现优秀的 BERT 预训练模型得益于 Transfomer[23]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[23]）网络、前馈神经网络、残差网络（Residual network[27]）以及层标准化（Layer normalization[28]）模块组成。

解码输出层实现对每个 token 的隐含状态进行标签预测。假设给定特征抽取层输出的隐含状态 H 及预测输出标签序列 Y：

$$H = (H_{[\mathrm{CLS}]}, H_1, \dots, H_n, H_{[\mathrm{SEP}]})$$

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

H 经过线性投影，得到大小为$(n+2)\times m$ 的得分矩阵 P，其中，$(n+2)$为输入序列的总长度，m 为不同标签的数量，P_{ij}对应该句中第 i 个 token 在第 j 个标签上的得分。对于输出得分矩阵 P，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标签。其中，第 i 个 token 在第 j 个标签上的概率计算公式为：

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

1.2 修正模型

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，实体的输出标签序列前后存在一定的依赖关系。例如：输出标签 I-MV 只能跟随 B-MV，而不能跟随标签 B-NP，其中，MV 代表机动车实体，NP 代表自然人主体实体。因此，我们采用条件随机场（Conditional random field[29]，CRF）作为解码输出层，以此来解决这个问题。

图 2 展示了修正后的模型 BERT-CRF，和基准模型的区别在于解码输出层的不同。结合上述特征抽取得到的隐含状态输出 H，定义综合得分函数：

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

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

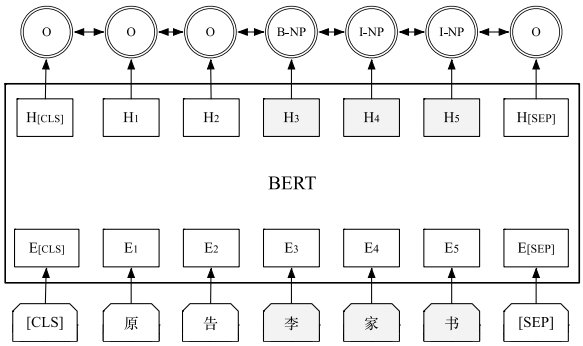

图 2 实体识别模型 BERT-CRF

对输入隐含状态H的所有可能的输出标签序列应用 Softmax，得到预测标签序列 Y 的概率：

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

我们需要使得综合得分最大化，一般取预测输出标签序列概率的对数：

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

其中，Y_X代表输入隐含状态对应所有可能的输出标签序列空间。

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。

$$Y^* = \text{argmax}_{\tilde{Y}\in Y_X} f\big(H,\tilde{Y}\big)$$

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[30]）进行解码。

2 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 模型，引入语言模型作为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 Embeddings[31]，TransE）任务作为辅助优化目标。

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”，目标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。其中位于句子起始的特殊标识符[CLS]用于对整个句子做表征，这里不再需要对句子末尾进行标识。

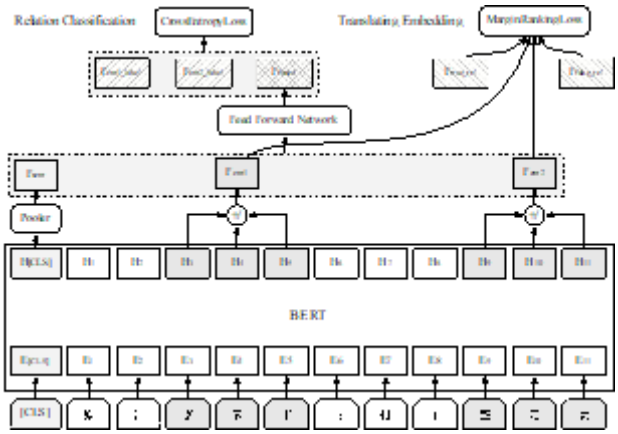

图 3 一种多任务联合的语义关系抽取模型

这里的输入嵌入层与 BERT 特征抽取层和 1.1 节中实体识别任务中介绍的基本一致。对于输入的每个 token，进行初始嵌入和特征抽取得到更高层次，更加丰富语义信息的隐含状态 H。

$$H = (H_{[\text{CLS}]}, H_1, \dots, H_n)$$

在得到隐含状态 H 的基础上，进一步的映射和处理得到句子的特征、实体 1 的特征以及实体 2 的特征，为下一步特征融合和多任务联合学习作准备。

对于句子特征的获取，取第一个 token（即[CLS]）的隐含状态$H_{[\text{CLS}]}$，输入 Pooler 层，即可得到句子的特征表示，用公式可表示为：

$$F_{sent} = \tanh(H_{[\text{CLS}]} W_{sent})$$

其中，W_{sent}为权重参数，进行相同维度的映射。

对于实体特征的获取，我们对组成一个实体的所有的 token 序列的隐含状态输出的平均值作为该实体的特征表示，公式表示为：

$$F_{ent} = \frac{1}{N}\sum_{k}^{N+k-1} H_i$$

其中，N 为该实体序列的长度，k 为实体序列起始的位置索引。

我们对句子特征、实体 1 特征、实体 2 特征进行连接，输入到前馈神经网络，进行特征融合，以获得融合特征F_{fused}，计算公式为：

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

其中，W_{fused}为权重参数，b_{fused}为偏置参数。

对于关系分类任务，考虑到特定的关系一般只会发生在特定的两个实体类别之间，我们将实体 1 的类别特征、实体 2 的类别特征与上述得到的混合特征进行连接得到最终的特征F_{final} =

$F_{ent1\_label} \oplus F_{ent2\_label} \oplus F_{sent}$，输入到输出层Softmax，以进行输出标签 y 的预测。

$$P(y|F_{final}) = \text{softmax}(F_{final}W_{final})$$

关系分类任务中，我们需要使得下列目标最小化：

$$\mathcal{L}_1 = -\sum_{F_{final} \in F_X} \log P(y|F_{final})$$

其中，X 为输入空间，F_X为最终的特征空间。

知识三元组平移嵌入模型 TransE[31]的基本思想是：对于给定的知识三元组集合(h,r,t)，由头实体和尾实体$h,r \in E$（实体集合）和一条关系（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、所在的句子以及对应的真实关系标签，考虑到这样得到的知识三元组都是正样例，为解决没有负样例的问题，我们从关系类别集合中随机选择一个非真实关系类别组成一条负样例。假定真关系标签的特征为F_{true_rel}，假关系标签的特征为F_{fake_rel}，采用基于距离排序的方法得到平移嵌入任务的最小化优化目标：

$$\mathcal{L}_2 = \sum_X [\gamma + d(F_{ent1} + F_{true\_rel}, F_{ent2}) - d(F_{ent1} + F_{fake\_rel}, F_{ent2})]_+$$

其中，距离d的度量采用L_1范数，为$d(F_{ent1} + F_{rel}, F_{ent2}) = ||F_{ent1} + F_{rel} - F_{ent2}||$，$[x]_+ = \max(x, 0)$，$\gamma > 0$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，得到综合损失函数为：

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

其中，$\lambda > 0$为平移嵌入任务损失的权重系数。

3 实验

3.1 数据准备

3.1.1 预定义实体和关系

我们以民事案由“机动车交通事故责任纠纷”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，并结合实际文书和现行的法律规范，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，编号对应为 12-20。

表 1 预定义实体类型

#	实体类型	#	实体类型
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

预定义的 9 种关系类型如表 2 所示，一个关系类型可能会对应多个实体类型对，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）这两条概念层知识三元组，合计可得到 30 条概念层知识三元组（这里不计“其他”关系类）。

表 2 预定义关系类型

#	关系类别	关系对应的实体类型对
1	驾驶	自然人主体→[机动车|非机动车]
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]
3	搭乘	[机动车|非机动车]→自然人主体
4	投保	机动车→保险类别
5	实施	自然人主体→[常见 9 类违法行为]
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]
7	承担	自然人主体→责任认定
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]
9	其他	—

3.1.2 数据预处理及标注划分

考虑到中国东、中、西三大区域的经济及技术发展水平的差异，司法能力水平和文书写作规范也会存在一些区别。我们在东部选取“江苏省”和“浙江省”，中部选取“河南省”和“湖北省”，西部选取“四川省”和“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，考虑到文书的结构规范性和用语成文化，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”等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文本类型的分段效果的 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，并用于后续实体对齐。

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进行分句处理并导入标注系统，在线标注示例截图如图 4 所示。

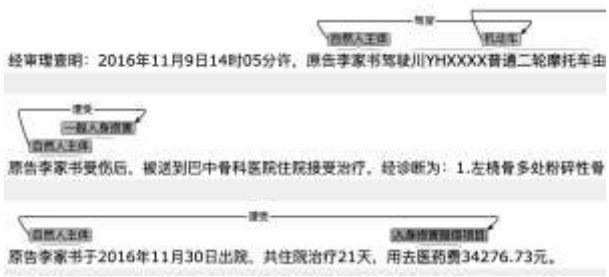

图 4 brat 在线标注示例截图

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标注文本。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的数据量统计针对句子层级。在建立关系任务数据集时，对于不存在关系且在关系实体类型对定义内的两个实体，选择“其他”关系标签并以 0.5 的保留概率作为“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

3.2 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[1] 的混合精度训练（Mixed precision training[33]）技术，BERT 模型使用的是 PyTorch 的实现 [2]（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，并未进行严格的网格搜索。BERT-Softmax 和 BERT-CRF 模型的参数除了初始学习率不一致，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（偏置项不做处理），且在顶层的线性层的 dropout[34]设置为 0.1，以避免过拟合；小批量大小设为 16；梯度裁剪的梯度阈值设为 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	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**

[1] https://github.com/NVIDIA/apex

[2] https://github.com/huggingface/pytorch-pretrained-BERT

如表 5 所示，在测试集上，修正模型 BERT-CRF 相比基准模型 BERT-Softmax，准确率下降 0.27，召回率上升 1.01，综合指标 F1 有 0.36 的提升。综合来看，修正模型优于基准模型，表现出融入输出标签之间的转移约束关系可以使得实体识别的效果得到进一步提升。

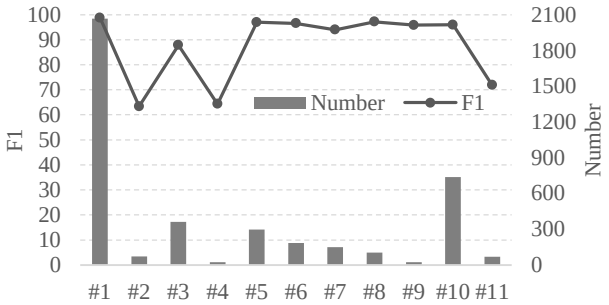

图 5　各实体类别的数量及对应的 F1 值（Number≥20）

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较差。分析原因，首先这三类实体数量比较少，数据不足导致模型学习不充分；死亡类实体数量虽然很少，但由于其表达如“死亡”、“致死”等比较固定，因而也能获得较好的效果；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多种多样，例如财产损失可能会涉及各种物品损失的表达，导致模型学习较为困难。

3.3　关系抽取

表 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基本一致。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，平移嵌入任务的间隔参数取默认值 1.0。

参照 SemEval-2010 Task 8[35]多关系分类任务的官方评测标准，我们取宏平均（Macro-averaged）的准确率、召回率及 F1 值进行效果评估，唯一的区别是学术研究标准任务评测未考虑“其他”关系类，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”类关系抽取的表现往往要低于平均值。

表 5　关系抽取超参数设置

参数	取值	参数	取值
max length	400	entity emb size	128
learning rate	2e-5	rel emb size	768
weight decay	0.01	λ	1e-5
batch size	16	γ	1.0

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	89.98	89.27
BERT-Multitask	**91.53**	**91.75**	**91.64**

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-Base，准确率、召回率及 F1 值都获得了全面提升，综合指标 F1 提升高达 2.37，表明融入平移嵌入的多任务联合的语义关系抽取模型能够明显改善关系抽取的效果，验证了融入语义信息约束的有效性。在模型 BERT-Multitask 训练完成后，利用该模型同时可以得到一件非常有价值的副产物，即一种结合了上下文及三元组语义关系的实体和关系的向量嵌入表示。

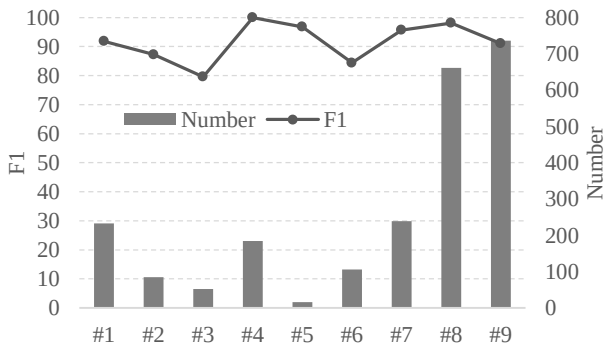

图 6　各关系类别的数量及对应的 F1 值

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 值在 95 以上的有 4 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。而对于“搭乘”和“发生事故”这两类关系，它们的抽取效果则要表现较差一些，经分析发现，涉及这两类关系的数据量相对较少，且这两类关系的两实体的相隔距离往往较远。

3.4 案情知识图谱自动构建

在“机动车交通事故责任纠纷”案由下，案情知识图谱构建的文本类型包括类结构化文本与非结构化文本。类结构化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”，非结构化文本只包含“法院认定事实”段落类型。司法案件的案情知识图谱自动构建流程图如图 7 所示。

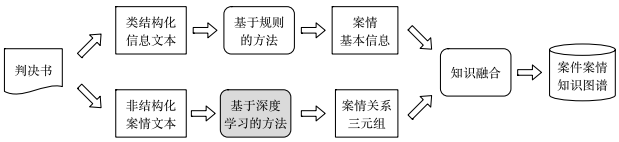

图 7 案情知识图谱自动构建流程图

对于类结构化文本，我们采用基于规则的方法进行抽取，以获取案情的基本信息；对于非结构化文本，我们采用基于深度学习的方法进行抽取，以获取案情关系知识三元组；对案件案情的基本信息和知识三元组进行知识融合，最终获得该案件的案情知识图谱。

面向司法案件的案情知识图谱构建的详细步骤如下：

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$Text_1$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。基于训练的实体识别模型 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）的集合；

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，融合主要实现 *Info* 和 *Triples* 的实体对齐。主要采用一些基于规则的方法：例如利用两实体之间的固定表达制定规则，如“如下简称”、“简写为”等类似表达；利用实体自身的特点，如“川 A×××××号小轿车”与“川 A×××××号”的实体类别和关系约束，可根据车牌号进行对齐。实体对齐处理后得到案情知识 *Knowledge*；

Step 7：知识存储与可视化。将 *Knowledge* 写入 Neo4j 图数据库进行存储与可视化展示。

对新输入的一份司法判决书，通过上述流程自动生成的案件案情知识图谱如图 8 所示，结果验证了该构建流程的可行性和有效性。根据该流程，我们在“机动车交通事故责任纠纷”案由下，选取了 20 万余份一审判决书进行了案情知识图谱自动构建，获得了一个大规模司法案件的案情知识图谱。

图 8 司法案件的案情知识图谱自动构建示例

4 结语

本文致力于面向司法案件的案情知识图谱自动构建的研究与实现。对于知识图谱构建涉及的两个重要的 NLP 任务进行了重点研究：针对实体识别任务，对比研究了两种基于 BERT 的实体识别模型，采用 CRF 进行解码输出可使结果得到

进一步提高；针对关系抽取任务，我们提出了一种融合平移嵌入的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的案情知识图谱自动构建流程，通过实验对该流程的可行性和有效性进行了验证，并构建了一个大规模案件的案情知识图谱。

本文研究的案情知识图谱自动构建的一个重要的前期工作是段落类型标记任务，目前采用基于规则的方法，非结构化案情事实文本的提取效果相对较差，下一步将结合一些监督学习的方法提升该任务的效果；为进一步提升与合理评价案情图谱构建质量，下一步将继续加大数据标注规模，算法与构建流程中也充分考虑法律知识并结合文书写作自身特点，并建立合理的案情知识图谱构建质量评价体系；围绕已构建的大规模案件的案情知识图谱，进行类案精准推送与检索等司法应用的研究。

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] Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledge base[J]. Communications of the ACM. 2014, 57(10): 78-85.

[3] Bizer C, Lehmann J, Kobilarov G, et al. DBpedia-A crystallization point for the Web of Data[J]. Journal of Web Semantics. 2009, 7(3): 154-165.

[4] Suchanek F M, Kasneci G, Weikum G. YAGO: A core of semantic knowledge unifying WordNet and Wikipedia[C]//Proceedings of the 16th International Conference on World Wide Web. ACM, 2007: 697-706.

[5] Niu X, Sun X, Wang H, et al. Zhishi.me-weaving chinese linking open data[C]//Proceedings of the 10th international conference on the semantic web. 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] Tang J, Zhang J, Yao L, et al. Arnetminer: extraction and mining of academic social networks[C]//Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2008: 990-998.

[8] Wang R, Yan Y, Wang J, et al. AceKG: A Large-scale Knowledge Graph for Academic Data Mining[C]//Proceedings of the 27th ACM International Conference on Information and Knowledge Management. ACM, 2018: 1487-1490.

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

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

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

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

[13] 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.

[14] 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.

[15] 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). Association for Computational Linguistics, 2016, 2: 207-212.

[16] Zeng D, Liu K, Lai S, et al. Relation classification via convolutional deep neural network[J]. Dublin City University and Association for Computational Linguistics, 2014: 2335–2344.

[17] 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. The COLING 2016 Organizing Committee, 2016: 2526-2536.

[18] Wang L, Cao Z, De Melo G, et al. Relation classification via multi-level attention cnns[J]. Association for Computational Linguistics, 2016: 1298–1307.

[19] 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.

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

[21] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.
[22] Peters M E, Neumann M, Iyyer M, et al. Deep contextualized word representations[J]. Association for Computational Linguistics, 2018: 2227–2237.
[23] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in neural information processing systems. Curran Associates, Inc, 2017: 5998-6008.
[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] 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.
[26] Alt C, Hübner M, Hennig L. Improving Relation Extraction by Pre-trained Language Representations[C]//AKBC 2019: Conference on Automated Knowledge Base Construction , 2019.
[27] 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
[28] Lei Ba J, Kiros J R, Hinton G E. Layer normalization[J]. arXiv preprint arXiv:1607.06450, 2016.
[29] Lafferty J, McCallum A, Pereira F C N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[C]//Proceedings of the Eighteenth International Conference on Machine Learning, 2001: 282-289.
[30] Forney G D. The viterbi algorithm[J]. Proceedings of the IEEE, 1973, 61(3): 268-278.
[31] Bordes A, Usunier N, Garcia-Duran A, et al. Translating embeddings for modeling multi-relational data[C]//Advances in neural information processing systems. Curran Associates, Inc, 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

洪文兴(1980—)，博士，副教授，主要研究领域为数据挖掘、大数据分析、推荐系统、系统工程。
E-mail：hwx@xmu.edu.cn

翁洋(1979—)，通信作者，博士，副教授，主要研究领域为统计机器学习、数据挖掘、分布式估计理论。
E-mail：wengyang@scu.edu.cn

胡志强(1993—)，硕士研究生，主要研究领域为自然语言处理、知识图谱。
E-mail：huzhiqiang@stu.xmu.edu.cn