

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2022 年 11 月 17 日 (17.11.2022)



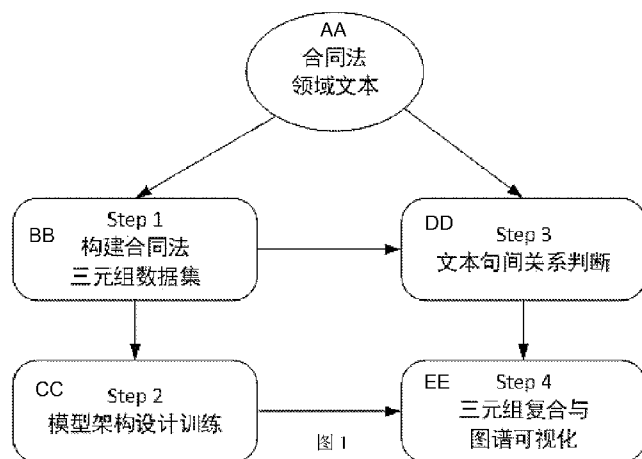
(10) 国际公布号
WO 2022/237013 A1

- (51) 国际专利分类号:
G06F 16/36 (2019.01) *G06K 9/62* (2022.01)
G06F 40/211 (2020.01) *G06N 3/04* (2006.01)
- (21) 国际申请号: PCT/CN2021/116053
- (22) 国际申请日: 2021 年 9 月 1 日 (01.09.2021)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
202110513432.2 2021年5月11日 (11.05.2021) CN
- (71) 申请人: 西安交通大学 (XI'AN JIAOTONG UNIVERSITY) [CN/CN]; 中国陕西省西安市咸宁西路28号, Shaanxi 710049 (CN)。
- (72) 发明人: 郑庆华 (ZHENG, Qinghua); 中国陕西省西安市咸宁西路28号, Shaanxi 710049 (CN)。 马昆

明 (MA, Kunming); 中国陕西省西安市咸宁西路28号, Shaanxi 710049 (CN)。 刘均 (LIU, Jun); 中国陕西省西安市咸宁西路28号, Shaanxi 710049 (CN)。 李星熠 (LI, Xingyi); 中国陕西省西安市咸宁西路28号, Shaanxi 710049 (CN)。 马黛露丝 (MA, Dailusi); 中国陕西省西安市咸宁西路28号, Shaanxi 710049 (CN)。 王佳欣 (WANG, Jiaxin); 中国陕西省西安市咸宁西路28号, Shaanxi 710049 (CN)。 朱海萍 (ZHU, Haiping); 中国陕西省西安市咸宁西路28号, Shaanxi 710049 (CN)。 麻珂欣 (MA, Kexin); 中国陕西省西安市咸宁西路28号, Shaanxi 710049 (CN)。 李鸿轩 (LI, Hongxuan); 中国陕西省西安市咸宁西路28号, Shaanxi 710049 (CN)。 魏笔凡 (WEI, Bifan); 中国陕西省西安市咸宁西路28号, Shaanxi 710049 (CN)。 张玲玲 (ZHANG, Lingling); 中国陕西省西安市咸宁西路28号, Shaanxi 710049 (CN)。

(54) **Title:** ENTITY RELATIONSHIP JOINT EXTRACTION-BASED LEGAL KNOWLEDGE GRAPH CONSTRUCTION METHOD AND DEVICE

(54) 发明名称: 基于实体关系联合抽取的法律知识图谱构建方法及设备



AA Text in the field of contract law
BB Step 1, construction of a triple data set of the contract law
CC Step 2, model architecture design and training
DD Step 3, determination of relationships between text sentences
EE Step 4, triple compounding and graph visualization

(57) **Abstract:** An entity relationship joint extraction-based legal knowledge graph construction method and device, the construction method comprising: construction of a triple data set; design of model architecture and training of a model, the model architecture comprising a model encoding layer, a head entity extraction layer and a relationship-tail entity extraction layer; determination of relationships between text sentences; and triple compounding and graph visualization. The design of the model architecture of the present invention uses a Chinese bert pre-training model as an encoder, which has a good effect on Chinese text encoding. An entity extraction part uses two BiLSTM binary classifiers to distinguish between start and end positions of entities, which may effectively extract the entities in the form of phrases in the text. In the present invention, a head entity is extracted first, and a tail entity corresponding to an entity relationship is then extracted from the extracted head entity. When extracting the entity relationship and the tail entity, not only is encoding information of a sentence used, but encoding information of the head entity is also incorporated. In the present invention, a highly accurate legal knowledge graph may be obtained.

(74) 代理人: 西安通大专利代理有限公司(XI'AN TONG DA PATENT AGENCY CO., LTD.);
中国陕西省西安市高新区锦业路 32
号, Shaanxi 710075 (CN)。

(81) 指定国(除另有指明, 要求每一种可提供的国家
保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG,
BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU,
CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB,
GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT,
JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK,
LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX,
MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL,
PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL,
ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US,
UZ, VC, VN, WS, ZA, ZM, ZW。

(84) 指定国(除另有指明, 要求每一种可提供的地区
保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ,
NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM,
AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG,
CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU,
IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT,
RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI,
CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告(条约第21条(3))。

(57) 摘要: 一种基于实体关系联合抽取的法律知识图谱构建方法及设备, 构建方法包括: 三元组数据集的构建; 模型架构的设计和模型的训练; 模型架构包括模型编码层、头实体抽取层以及关系-尾实体抽取层; 文本句间关系判断; 以及三元组复合与图谱可视化; 本发明模型架构的设计采用了中文 bert 预训练模型作为编码器, 对中文的文本编码效果好。实体抽取部分采用两个 BiLSTM 二分类器来判别实体的起始位置和结束位置, 可以有效地抽取出文本中短语形式的实体。本发明先抽取头实体, 再由抽取到的头实体抽取对应实体关系的尾实体, 抽取实体关系和尾实体时不仅用到了句子的编码信息, 还融入了头实体的编码信息。本发明能够得到准确率较高的法律知识图谱。

基于实体关系联合抽取的法律知识图谱构建方法及设备

技术领域

本发明属于电子信息领域，具体涉及一种基于实体关系联合抽取的法律知识图谱构建方法及设备。

背景技术

知识图谱 (Knowledge Graph)，在图书情报界称为知识域可视化或知识领域映射地图，是显示知识发展进程与结构关系的一系列各种不同的图形，用可视化技术描述知识资源及其载体，挖掘、分析、构建、绘制和显示知识及它们之间的相互联系。知识图谱是通过将应用数学、图形学、信息可视化技术、信息科学等学科的理论与方法与计量学引文分析、共现分析等方法结合，并利用可视化的图谱形象地展示学科的核心结构、发展历史、前沿领域以及整体知识架构达到多学科融合目的的现代理论，能为学科研究提供切实的、有价值的参考。

知识图谱分为通用知识图谱与领域知识图谱两类。这两种知识图谱主要存在覆盖范围和使用方式上的差异。通用知识图谱面向通用领域，主要包含了大量现实世界中的常识性知识，覆盖面广。领域知识图谱又称为行业知识图谱或垂直知识图谱，是面向某一特定领域的，是由该领域的专业数据构成的行业知识库，因其基于行业数据构建，有着严格而丰富的数据模式，所以对该领域知识的深度、知识准确性有着更高的要求。领域知识图谱是面向特定领域的知识图谱，如电商、金融、医疗等。相比较而言，领域知识图谱的知识来源更多、规模化扩展要求更迅速、知识结构更加复杂、知识质量要求更高、知识的应用形式也更加广泛。

面向法律领域的知识图谱构建与研究现今仍然较为匮乏，在法律发展较为快速全面的如今，法律领域知识图谱的需求也渐渐浮现。领域知识图谱的构建需要大量该领域的信息，如何从海量的无结构或半结构中抽取出有价值的信息，引起了众多学者的关注，信息抽取技术

应运而生。其中构建知识图谱主要用到信息抽取中的实体抽取和关系抽取子任务。实体抽取技术，又称命名实体识别技术，是指识别抽取文本中具有特定意义的实体，主要包括人名、地名、机构名、专有名词等，以及时间、数量、货币、比例数值等文字。关系抽取的主要任务是抽取文本中实体之间的关系。通常将实体和实体之间的关系形式化地描述为三元组的关系 $\langle h, r, t \rangle$ ，其中 h ， t 表示头实体和尾实体， r 表示实体之间的关系。例如，在“《霸王别姬》的导演是陈凯歌”这句话中，“《霸王别姬》”和“陈凯歌”都是实体，两个实体之间的关系是“导演”关系，用三元组就可以表示为 $\langle \text{《霸王别姬》}, \text{导演}, \text{陈凯歌} \rangle$ ，信息抽取的主要目的就是在大量无结构或半结构文本中抽取三元组形式的数据，最终由大量的三元组构成知识图谱。然而，传统的信息抽取技术采用的是一种 pipeline 的方式，对于非结构化的文本，先进行实体抽取，然后在实体抽取结果的基础上进行关系抽取，这样做有一个很大的弊端，一旦实体抽取的结果出错，将会很大程度上影响关系抽取的准确率，这样就会导致错误的传递。

发明内容

本发明目的在于针对现有技术中法律领域知识图谱构建准确率不能保证的问题，提供一种基于实体关系联合抽取的法律知识图谱构建方法及设备，得到准确率较高的知识图谱。

为了实现上述目的，本发明有以下的技术方案：

一种基于实体关系联合抽取的法律知识图谱构建方法，包括以下步骤：

-三元组数据集的构建；

将法律文本语句拆分成短句的形式；

将短句中缺省的主语补全；

从短句中抽取三元组，构建三元组数据集；

-模型架构的设计和模型的训练；

模型架构包括模型编码层、头实体抽取层以及关系-尾实体抽取层；

具体的，模型编码层使用 bert 预训练模型；

头实体抽取层使用两个 BiLSTM 作为二分类器，将文本的编码作为分类器的输入，输出信息中，第一个 BiLSTM 二分类器对应的实体起始位置输出为 1，其余位置输出都为 0，第二个 BiLSTM 二分类器对应的实体结束位置输出为 1，其余位置输出都为 0；

关系-尾实体抽取层，将头实体的编码信息与句子的编码相结合作为输入，对于每个头实体，找到每个关系下可能存在的尾实体，最终得到完整的三元组；

-文本句间关系判断；

对于未进行短句拆分的法律文本语句，判断各个短句之间的关系；

-三元组复合与图谱可视化；

根据模型抽取出的三元组结合文本句间关系，得到法律文本对应的复合三元组；

法律知识图谱的可视化。

作为本发明基于实体关系联合抽取的法律知识图谱构建方法的一种优选方案，构建三元组数据集时将主语和宾语分别作为头实体和尾实体，将谓语作为关系。

作为本发明基于实体关系联合抽取的法律知识图谱构建方法的一种优选方案，根据标注的关系确定关系集合，合并语义相同或相似的关系。

作为本发明基于实体关系联合抽取的法律知识图谱构建方法的一种优选方案，头实体抽取层以 bert 编码层输出的特征向量 x_i 作为输入，输出抽取到实体的起始和结束标志；

$$p_i^{h.start} = BiLSTM(W_s x_i + b_s)$$

$$p_i^{h.end} = BiLSTM(W_e x_i + b_e)$$

其中 x_i 为每个词的特征向量， W_s ， W_e 为两个二分类器能够训练的权重矩阵， b_s ， b_e 为各自的偏置向量； p_i^{start} 为实体的起始位置的标志，当其值趋近于 1 时，表示该位置是实体的起始位置； p_i^{end} 为实体的结束位置的标志，当其值趋近于 1 时，表示该位置是实体的结束位置。

作为本发明基于实体关系联合抽取的法律知识图谱构建方法的一种优选方案，关系-尾实体抽取层采用两个结构相同的 BiLSTM 组成，该层模型的输入包括句子的特征向量 h_s ，并融入了上一层抽取的头实体编码 h_{head}^k ，其中 k 表示第 k 个头实体；

将 $h_s + h_{head}^k$ 作为该层的输入向量 x_i ，具体的计算公式如下：

$$p_i^{t.start} = BiLSTM(W_s^r(x_i + h_{head}^k) + b_s^r)$$

$$p_i^{t.end} = BiLSTM(W_e^r(x_i + h_{head}^k) + b_e^r)$$

其中，向量 h_s 和 h_{head}^k 是直接向量相加的关系，维度相同；对于第 k 个头实体，取其开始位置到结束位置词向量的平均值作为向量 h_{head}^k 的表示； W_s^r 和 W_e^r 表示起始位置和结束位置能够训练的参数矩阵；对于每个头实体，遍历关系集合中所有的关系，重复上面的计算公式，找到每个关系下可能存在的尾实体，最终得到完整的三元组。

本发明还提供一种基于实体关系联合抽取的法律知识图谱构建系统，包括：

三元组数据集构建模块，用于将法律文本语句拆分成短句的形式，并能够将短句中缺省的主语补全，最终从短句中抽取三元组，构建三元组数据集；

模型建立和训练模块，用于对模型架构中的模型编码层、头实体抽取层以及关系-尾实体抽取层分别构建，经过训练得到能够抽取出三元组的模型；

句间关系判断模块，用于对未进行短句拆分的法律文本语句，判断各个短句之间的关系；

知识图谱可视化模块，用于根据模型抽取出的三元组结合文本句间关系，得到法律文本对应的复合三元组，并实现法律知识图谱的可视化。

本发明还提供一种终端设备，包括存储器、处理器以及存储在所述存储器中并可在所述处理器上运行的计算机程序，所述的处理器执行所述的计算机程序时实现所述基于实体关系联合抽取的法律知识图谱构建方法的步骤。

本发明还提供一种计算机可读存储介质，所述计算机可读存储介质存储有计算机程序，

所述的计算机程序被处理器执行时实现所述基于实体关系联合抽取的法律知识图谱构建方法的步骤。

相较于现有技术，本发明有如下的有益效果：

现有的知识图谱构建方法往往采用的是 pipeline 的思想，先进行实体抽取，并以实体抽取的结果进行关系抽取，这样会导致错误的累计。本发明利用一种基于实体关系联合抽取的模型，对合同法领域的非结构化文本进行三元组抽取，最终构建出法律领域的知识图谱，本发明能够避免 pipeline 方法导致的错误传递，准确率高。本发明模型架构的设计采用了中文 bert 预训练模型作为编码器，对中文的文本编码效果好。由于法律领域的实体是短语形式，因此，实体抽取部分采用两个 BiLSTM 二分类器来判别实体的起始位置和结束位置，可以有效地抽取出文本中短语形式的实体。本发明给定一段文本，先抽取头实体，再由抽取到的头实体抽取对应实体关系的尾实体。抽取实体关系和尾实体时不仅用到了句子的编码信息，还融入了头实体的编码信息。本发明能够得到准确率较高的法律知识图谱，构建完成的知识图谱可以结合深度学习技术实现合同法领域的问答推理及相关推荐等功能。

附图说明

图 1 是本发明实施例合同法知识图谱构建方法流程图；

图 2 是本发明实施例模型训练构建的数据集样例图；

图 3 是本发明实施例从文本抽取三元组的模型架构图；

图 4 是本发明实施例抽取三元组模型的训练和使用示意图；

图 5 是本发明实施例构建的知识图谱可视化示意图。

具体实施方式

下面结合附图及实施例对本发明做进一步的详细说明。

本发明提出一种基于实体关系联合抽取的法律知识图谱构建方法，实施例以合同法为例

进行说明，本发明可以利用给定的合同法文本，同时进行实体抽取和关系抽取，最终得到完整的三元组信息。由抽取到的三元组首尾连接即可构成合同法知识图谱。构建完成的知识图谱可以结合深度学习技术实现合同法领域的问答推理及相关推荐等功能。

实体抽取：对于任何一个完整的合同法文本语句，都可以将其分解为(h,r,t)的形式，h 表示头实体，r 表示实体关系，t 表示尾实体。实体抽取表示抽取文本中的头实体和尾实体。

关系抽取：这里的关系指的是实体之间的关系或实体的属性，这一步通常在头实体和尾实体抽取完成后抽取对应的实体关系。

联合抽取：不同于以往的实体抽取和关系抽取各自独立分别进行，联合抽取所抽取到的实体和关系互相影响，采用联合抽取可以减少由实体抽取导致的错误传递问题。

参见图 1，本发明实施例基于实体关系联合抽取的合同法图谱构建方法，包括以下步骤：

步骤一、合同法三元组数据集的构建，包括：

1.1) 将复杂的合同法文本语句拆分成简单短句的形式。

根据合同法文本语句本身特点，绝大部分合同法文本都是由两个及以上短句构成，且短句之间都存在一定的逻辑关系。如合同法第九条“当事人订立合同，应当具有相应的民事权利能力和民事行为能力”。为了准确抽取文本中的三元组，需要将其拆分成多个短句的形式。本例可以得到两个短句“当事人订立合同”和“应当具有相应的民事权利能力和民事行为能力”。

1.2) 利用零指代消解方面的技术解决因短句拆分导致的主语缺失问题。

经过第 1 步得到的拆分后的文本短句会存在部分短句中主语缺失的现象，这会影响后续三元组的抽取工作，因此需要将缺省的主语补全。本方法采用开源工具 pyltp 结合依存句法分析的方法对缺省部分进行句法分析，并补全缺省的主语。第 1 步中的例子经过主语补全后的结果为短句“当事人订立合同”和“(当事人)应当具有相应的民事权利能力和民事行为能力”。

1.3) 合同法三元组数据集的构建；

对于进行主语补全的短句，可以从中抽取出需要的三元组。为了保证三元组抽取模型的性能，需要人工标注三元组数据以训练模型。一个完整的句子通常由主语、谓语和宾语三部分构成，因此在标注数据时，将句子的主语和宾语分别作为头实体和尾实体，将谓语作为关系初步标注。标注部分三元组后，根据标注的关系确定关系集合，合并语义相同或相似的关系，如关系“概念”和关系“定义”语义相似，将关系“定义”统一化为关系“概念”。图 2 是部分标注的三元组示例。最终得到人工标定的三元组 798 个，实体关系集合中包含 25 个关系。

步骤二、模型架构的设计和模型的训练，包括：

2.1) 模型架构的设计；

三元组训练数据标注完成后，下面进行实验模型的设计，本模型的设计主要考虑以下几个方面：第一，bert 预训练模型使用了双向 Transformer，同时在预训练过程中使用 Masked Language Model(MLM) 捕获词语级别的表示，这使得词向量从先前只包含前文信息变成了可以学习上下文的信息，在预训练过程中使用 Next Sentence Prediction(NSP)捕获句子级别的表示。因此在编码层使用 bert 预训练模型可以更好地表征句子的深层含义。第二，合同法法条文本中的实体与通用领域的实体有所不同，不仅包含词语实体，还包含短语实体，这样再用传统的 NER 方法无法准确抽取短语实体。因此考虑使用两个 BiLSTM 作为二分类器，将文本的编码作为分类器的输入，输出信息中，第一个 BiLSTM 二分类器对应的实体起始位置输出为 1 其余位置输出都为 0，第二个 BiLSTM 二分类器对应的实体结束位置输出为 1，其余位置输出都为 0。分别提取实体的起始位置和结束位置的位置编码，这样可以根据需要很好地抽取短语实体。第三，为了使得实体关系和尾实体的抽取充分利用头实体的编码信息，对于模型的实体关系和尾实体抽取部分的输入不只是整个句子的编码信息，而是将头实体的编码信息与句子的编码相结合，这对于实体关系和尾实体的抽取有很好的效果提升。

通过以上的分析，设计了合同法领域实体关系联合抽取的算法模型，模型框架图如图所

示。该模型共分为三个部分，分别是模型编码层，头实体抽取层，关系-尾实体抽取层，模型架构图参考图 3，各个部分的具体内容如下：

(a) 模型编码层

本发明的模型编码层采用的是哈工大讯飞联合实验室在更大规模语料上基于全词 Mask 的中文预训练模型 BERT-wwm-ext，该模型在多项基准测试上获得了进一步的性能提升。利用该模型可以将输入的文本转化为特征向量的形式。

(b) 头实体抽取层

该层主要由两个结构相同的 BiLSTM 组成，以 bert 编码层输出的特征向量 x_i 作为输入，输出抽取到实体的起始和结束标志。

$$p_i^{h.start} = BiLSTM(W_s x_i + b_s)$$

$$p_i^{h.end} = BiLSTM(W_e x_i + b_e)$$

其中 x_i 为每个词的特征向量， W_s ， W_e 为两个二分类器可以训练的权重矩阵， b_s ， b_e 为各自的偏置向量。 p_i^{start} 为实体的起始位置的标志，当其值趋近于 1 时，表示该位置是实体的起始位置； p_i^{end} 为实体的结束位置的标志，当其值趋近于 1 时，表示该位置是实体的结束位置；图中示例，对于文本“当事人应当具有相应的民事权利能力和民事行为能力”，经过头实体抽取层可以抽取到的实体分别为“当事人”和“具有相应的民事权利能力和民事行为能力”，图 3 中黑色标记分别为“当事人”实体的起始位置和结束位置，浅灰色标记则为另一个实体的起始位置和结束位置。

(c) 关系-尾实体抽取层

该层的与头实体抽取层相似，也是由两个结构相同的 BiLSTM 组成，该层模型的输入不再只是句子的特征向量 h_s ，它还融入了上一层抽取的头实体编码 h_{head}^k ，其中 k 表示第 k 个头实体。将 $h_s + h_{head}^k$ 作为该层的输入向量。具体公式为：

$$p_i^{t.start} = BiLSTM(W_s^r(x_i + h_{head}^k) + b_s^r)$$

$$p_i^{t.end} = BiLSTM(W_e^r(x_i + h_{head}^k) + b_e^r)$$

其中，向量 h_s 和 h_{head}^k 是直接向量相加的关系，因此必须要保证维度相同，所以对于第 k 个头实体，取其开始位置到结束位置词向量的平均值作为向量 h_{head}^k 的表示； W_s^r 和 W_e^r 表示起始位置和结束位置可训练的参数矩阵，与头实体抽取层不同，这里表示对于每个头实体，遍历关系集合中所有的关系，重复上面的计算公式，从而找到每个关系下可能存在的尾实体，最终可以得到完整的三元组。对于图 3 中的示例，头实体“当事人”在关系为“应当”时对应的尾实体为“具有相应的民事权利能力和民事行为能力”，因此得到三元组（当事人，应当，具有相应的民事权利能力和民事行为能力）。

2.2) 模型的使用；

参见图 4，中间部分为模型的训练过程，模型的输入为合同法法条文本，分别经过 bert 编码层，头实体抽取层，关系-尾实体抽取层得到模型输出的三元组，然后利用给定的损失函数不断地进行迭代优化，当损失函数值趋于稳定时停止迭代，完成模型的训练，保存训练完成的模型。

对于未包含在测试集中的合同法文本三元组的抽取，利用训练好的模型，将其作为模型的输入，模型的输出即为文本对应的三元组。

步骤三、文本句间关系的判断。

对于未进行短句拆分的合同法文本，利用开源工具 `pyltp` 结合规则匹配的方法判断各个短句之间的关系。句间关系包含“条件”、“转折”、“并列”、“因果”四种关系，其中共得到因果关系 85 例，条件关系 194 例，转折关系 34 例，并列关系 8 例。例如，对于合同法文本“当事人订立合同，应当具有相应的民事权利能力和民事行为能力”，两个短句中抽取到的句间关系的“条件关系”。

步骤四、三元组复合与图谱可视化，包括：

4.1) 三元组的整合；

将模型抽取出的三元组，结合过程 3 得到的句间关系，可以得到合同法文本对应的复合三元组。对于例子“当事人订立合同，应当具有相应的民事权利能力和民事行为能力”，最终可以得到的三元组形式为 ((当事人, 订立, 合同), 条件, (当事人, 应当, 具有相应的民事权利能力和民事行为能力))。由此将抽取到的所有三元组进行整合拼接，可以得到完整的合同法知识图谱。

4.2) 合同法知识图谱的可视化；

参见图 5，是合同法知识图谱可视化后的部分示意图。构建完成的合同法知识图谱可以结合深度学习技术实现合同法领域的问答推理及相关推荐等功能。

一种基于实体关系联合抽取的法律知识图谱构建系统，包括：

三元组数据集构建模块，用于将法律文本语句拆分成短句的形式，并能够将短句中缺省的主语补全，最终从短句中抽取三元组，构建三元组数据集；

模型建立和训练模块，用于对模型架构中的模型编码层、头实体抽取层以及关系-尾实体抽取层分别构建，经过训练得到能够抽取出三元组的模型；

句间关系判断模块，用于对未进行短句拆分的法律文本语句，判断各个短句之间的关系；

知识图谱可视化模块，用于根据模型抽取出的三元组结合文本句间关系，得到法律文本对应的复合三元组，并实现法律知识图谱的可视化。

一种终端设备，包括存储器、处理器以及存储在所述存储器中并可在所述处理器上运行的计算机程序，所述的处理器执行所述的计算机程序时实现所述基于实体关系联合抽取的法律知识图谱构建方法的步骤。

一种计算机可读存储介质，所述计算机可读存储介质存储有计算机程序，所述的计算机

程序被处理器执行时实现所述基于实体关系联合抽取的法律知识图谱构建方法的步骤。

所述的计算机程序可以被分割成一个或多个模块/单元，所述一个或者多个模块/单元被存储在所述存储器中，并由所述处理器执行，以完成本发明的知识图谱构建方法。

处理器可以是中央处理单元(CentralProcessingUnit, CPU)，还可以是其他通用处理器、数字信号处理器(DigitalSignalProcessor, DSP)、专用集成电路(ApplicationSpecificIntegratedCircuit, ASIC)、现成可编程门阵列(Field-ProgrammableGateArray, FPGA)或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等。存储器可用于存储计算机程序和/或模块，所述处理器通过运行或执行存储在所述存储器内的计算机程序和/或模块，以及调用存储在存储器内的数据，实现本发明知识图谱构建系统的各种功能。

以上所述的仅仅是本发明的较佳实施例，并不用以对本发明的技术方案进行任何限制，本领域技术人员应当理解的是，在不脱离本发明精神和原则的前提下，该技术方案还可以进行若干简单的修改和替换，这些修改和替换也均属于权利要求书所涵盖的保护范围之内。

权利要求书

1.一种基于实体关系联合抽取的法律知识图谱构建方法，其特征在于，包括以下步骤：

-三元组数据集的构建；

将法律文本语句拆分成短句的形式；

将短句中缺省的主语补全；

从短句中抽取三元组，构建三元组数据集；

-模型架构的设计和模型的训练；

模型架构包括模型编码层、头实体抽取层以及关系-尾实体抽取层；

具体的，模型编码层使用 bert 预训练模型；

头实体抽取层使用两个 BiLSTM 作为二分类器，将文本的编码作为分类器的输入，输出信息中，第一个 BiLSTM 二分类器对应的实体起始位置输出为 1，其余位置输出都为 0，第二个 BiLSTM 二分类器对应的实体结束位置输出为 1，其余位置输出都为 0；

关系-尾实体抽取层，将头实体的编码信息与句子的编码相结合作为输入，对于每个头实体，找到每个关系下可能存在的尾实体，最终得到完整的三元组；

-文本句间关系判断；

对于未进行短句拆分的法律文本语句，判断各个短句之间的关系；

-三元组复合与图谱可视化；

根据模型抽取出的三元组结合文本句间关系，得到法律文本对应的复合三元组；

法律知识图谱的可视化。

2.根据权利要求 1 所述基于实体关系联合抽取的法律知识图谱构建方法，其特征在于：构建三元组数据集时将主语和宾语分别作为头实体和尾实体，将谓语作为关系。

3.根据权利要求 2 所述基于实体关系联合抽取的法律知识图谱构建方法，其特征在于：根据标注的关系确定关系集合，合并语义相同或相似的关系。

4.根据权利要求 1 所述基于实体关系联合抽取的法律知识图谱构建方法,其特征在于:
头实体抽取层以 bert 编码层输出的特征向量 x_i 作为输入,输出抽取到实体的起始和结束标志;

$$p_i^{h.start} = BiLSTM(W_s x_i + b_s)$$

$$p_i^{h.end} = BiLSTM(W_e x_i + b_e)$$

其中 x_i 为每个词的特征向量, W_s , W_e 为两个二分类器能够训练的权重矩阵, b_s , b_e 为各自的偏置向量; p_i^{start} 为实体的起始位置的标志,当其值趋近于 1 时,表示该位置是实体的起始位置; p_i^{end} 为实体的结束位置的标志,当其值趋近于 1 时,表示该位置是实体的结束位置。

5.根据权利要求 1 所述基于实体关系联合抽取的法律知识图谱构建方法,其特征在于:
关系-尾实体抽取层采用两个结构相同的 BiLSTM 组成,该层模型的输入包括句子的特征向量 h_s ,并融入了上一层抽取的头实体编码 h_{head}^k ,其中 k 表示第 k 个头实体;

将 $h_s + h_{head}^k$ 作为该层的输入向量 x_i ,具体的计算公式如下:

$$p_i^{t.start} = BiLSTM(W_s^r(x_i + h_{head}^k) + b_s^r)$$

$$p_i^{t.end} = BiLSTM(W_e^r(x_i + h_{head}^k) + b_e^r)$$

其中,向量 h_s 和 h_{head}^k 是直接向量相加的关系,维度相同;对于第 k 个头实体,取其开始位置到结束位置词向量的平均值作为向量 h_{head}^k 的表示; W_s^r 和 W_e^r 表示起始位置和结束位置能够训练的参数矩阵;对于每个头实体,遍历关系集合中所有的关系,重复上面的计算公式,找到每个关系下可能存在的尾实体,最终得到完整的三元组。

6.一种基于实体关系联合抽取的法律知识图谱构建系统,其特征在于,包括:

三元组数据集构建模块,用于将法律文本语句拆分成短句的形式,并能够将短句中缺省的主语补全,最终从短句中抽取三元组,构建三元组数据集;

模型建立和训练模块,用于对模型架构中的模型编码层、头实体抽取层以及关系-尾实体抽取层分别构建,经过训练得到能够抽取出三元组的模型;

句间关系判断模块,用于对未进行短句拆分的法律文本语句,判断各个短句之间的关系;

知识图谱可视化模块,用于根据模型抽取出的三元组结合文本句间关系,得到法律文本对应的复合三元组,并实现法律知识图谱的可视化。

7.一种终端设备,包括存储器、处理器以及存储在所述存储器中并可在所述处理器上运行的计算机程序,其特征在于:所述的处理器执行所述的计算机程序时实现如权利要求 1 至 5 中任意一项所述基于实体关系联合抽取的法律知识图谱构建方法的步骤。

8.一种计算机可读存储介质,所述计算机可读存储介质存储有计算机程序,其特征在于:所述的计算机程序被处理器执行时实现如权利要求 1 至 5 中任意一项所述基于实体关系联合抽取的法律知识图谱构建方法的步骤。

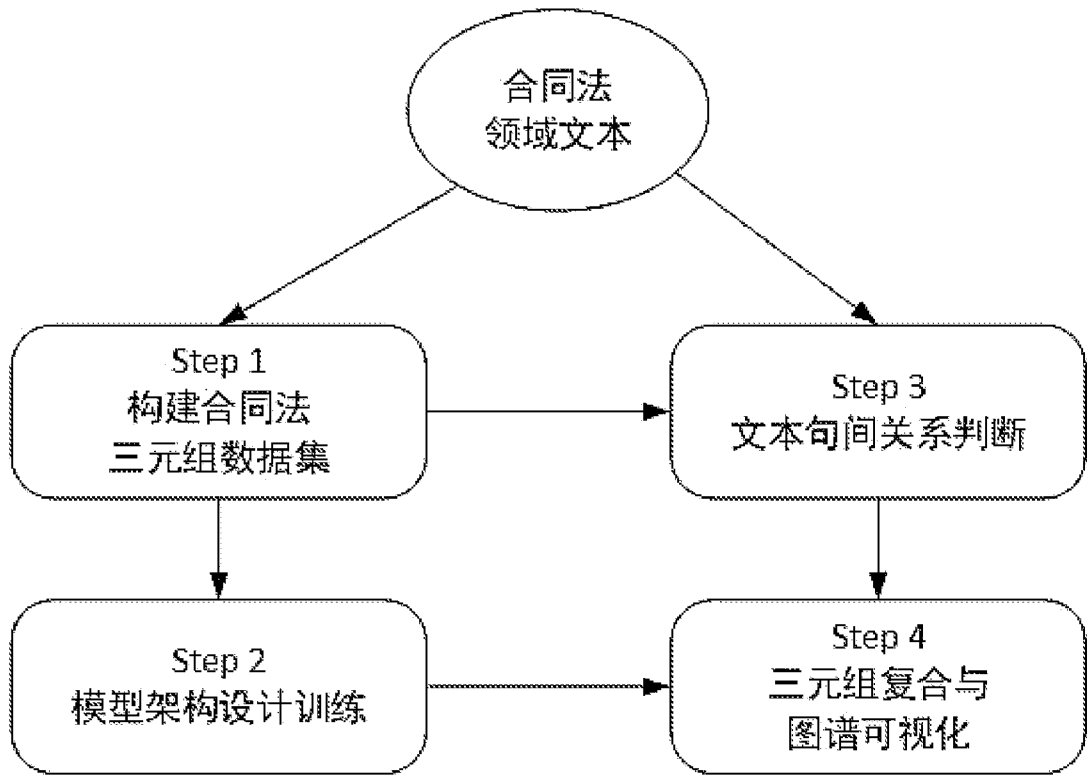


图 1

sentence	head1	tail1	relation1
承租人应当妥善保管租赁物	承租人	妥善保管租赁物	应当
承租人应当承担损害赔偿责任。	承租人	承担损害赔偿责任	应当
承租人在租赁物需要维修时可以要求出租人在合理期限内维修	承租人	要求出租人在合理期限内维修	可以
承租人可以自行维修	承租人	自行维修	可以
因承租人保管不善造成租赁物毁损、灭失的	承租人	租赁物毁损、灭失的	致使
维修费用由出租人负担。	出租人	维修费用	负担
承运人应当向旅客及时告知有关不能正常运输的重要事由和	承运人	向旅客及时告知有关不能正常运输的重要事由和安全	应当
旅客、托运人或者收货人应当支付票款或者运输费用。	旅客、托运人或	支付票款或者运输费用	应当
承运人未按照约定路线或者通常路线运输增加票款或者运输	承运人	按照约定路线或者通常路线运输增加票款或者运输费	应当
旅客、托运人或者收货人可以拒绝支付增加部分的票款或者	旅客、托运人或	拒绝支付增加部分的票款或者运输费用	可以

图 2

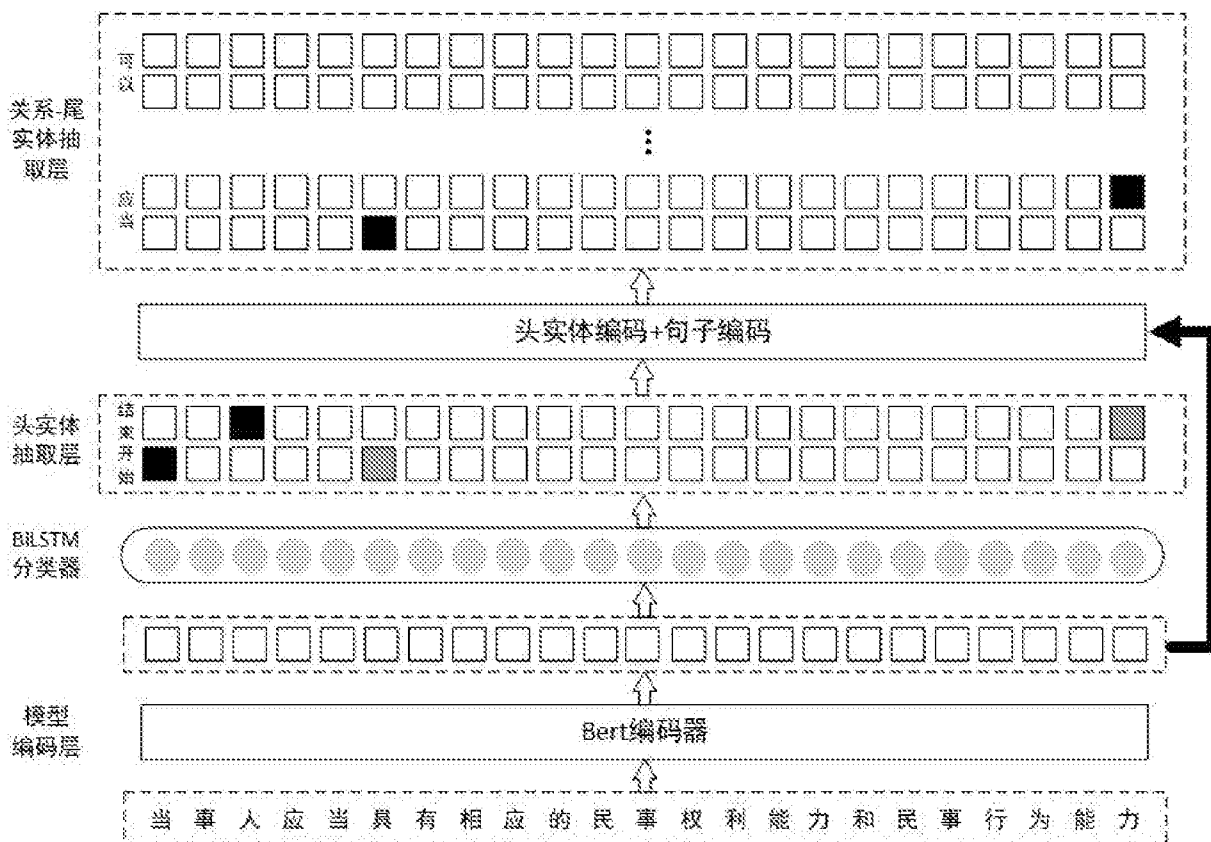


图 3

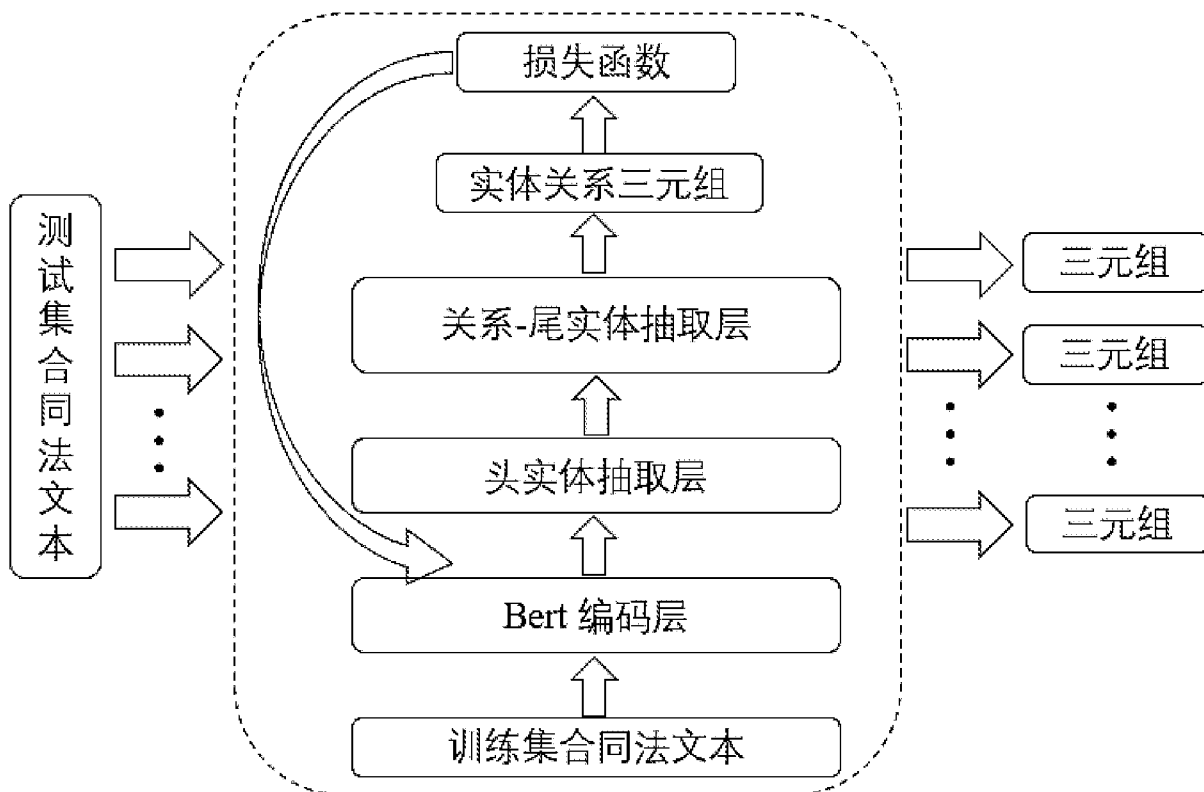


图 4

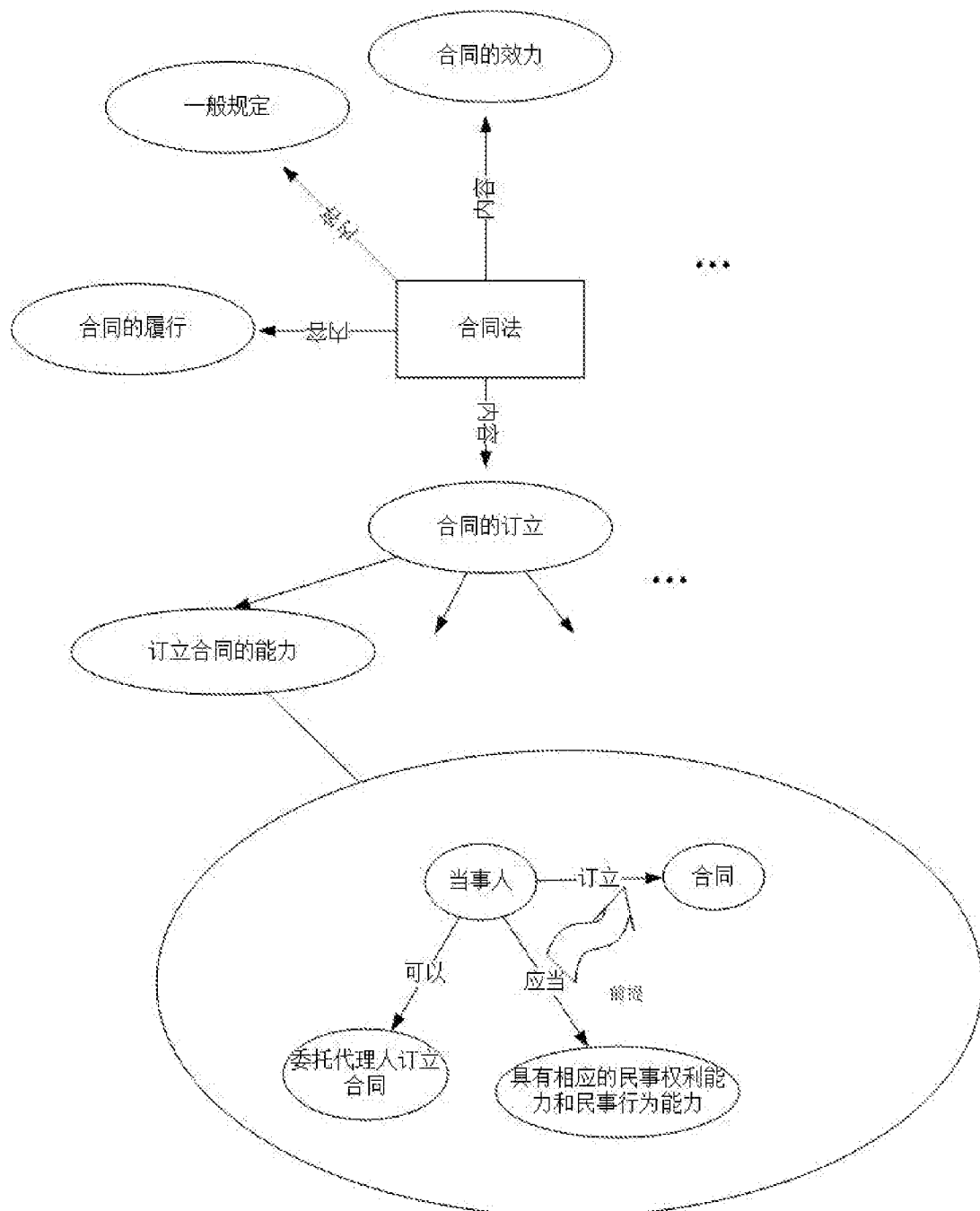


图 5

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2021/116053

A. CLASSIFICATION OF SUBJECT MATTER

G06F 16/36(2019.01)i; G06F 40/211(2020.01)i; G06K 9/62(2022.01)i; G06N 3/04(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F; G06K; G06N

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNPAT; WPI; EPODOC; CNKI; IEEE: 实体, 关系, 联合, 抽取, 提取, 知识图谱, 三元组, entity, relationship, joint, extraction, knowledge, graph, KG, triple

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 113204649 A (XI'AN JIAOTONG UNIVERSITY) 03 August 2021 (2021-08-03) claims 1-8	1-8
Y	CN 111858940 A (XIANGTAN UNIVERSITY) 30 October 2020 (2020-10-30) description, paragraphs [0053]-[0111], and figure 2	1-8
Y	WEI, Zhepei et al. "A Novel Cascade Binary Tagging Framework for Relational Triple Extraction." <i>arXiv:1909.03227v4 [cs.CL]</i> , 22 June 2020 (2020-06-22), pages 1-8, figure 2	1-8
Y	CN 109977237 A (NANJING AEGIS INFORMATION TECHNOLOGY CO., LTD.) 05 July 2019 (2019-07-05) description, paragraphs [0022]-[0026]	1-8
A	CN 110781254 A (SICHUAN UNIVERSITY et al.) 11 February 2020 (2020-02-11) entire document	1-8
A	CN 108874878 A (ZHONGAN INFORMATION TECHNOLOGY SERVICE CO., LTD.) 23 November 2018 (2018-11-23) entire document	1-8



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

19 January 2022

Date of mailing of the international search report

11 February 2022

Name and mailing address of the ISA/CN

China National Intellectual Property Administration (ISA/
CN)
No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing
100088, China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2021/116053

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2020073933 A1 (NATIONAL UNIVERSITY OF DEFENSE TECHNOLOGY) 05 March 2020 (2020-03-05) entire document	1-8

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2021/116053

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)		Publication date (day/month/year)
CN	113204649	A	03 August 2021	None		
CN	111858940	A	30 October 2020	None		
CN	109977237	A	05 July 2019	None		
CN	110781254	A	11 February 2020	None		
CN	108874878	A	23 November 2018	None		
US	2020073933	A1	05 March 2020	CN	109165385	A 08 January 2019

国际检索报告

国际申请号

PCT/CN2021/116053

A. 主题的分类

G06F 16/36(2019.01)i; G06F 40/211(2020.01)i; G06K 9/62(2022.01)i; G06N 3/04(2006.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G06F; G06K; G06N

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称, 和使用的检索词(如使用))

CNPAT;WPI;EPDOC;CNKI;IEEE: 实体, 关系, 联合, 抽取, 提取, 知识图谱, 三元组, entity, relationship, joint, extraction, knowledge, graph, KG, triple

C. 相关文件

类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求
PX	CN 113204649 A (西安交通大学) 2021年8月3日 (2021 - 08 - 03) 权利要求1-8	1-8
Y	CN 111858940 A (湘潭大学) 2020年10月30日 (2020 - 10 - 30) 说明书第[0053]-[0111]段, 图2	1-8
Y	WEI, Zhepei 等. "A Novel Cascade Binary Tagging Framework for Relational Triple Extraction." arXiv:1909.03227v4 [cs.CL], 2020年6月22日 (2020 - 06 - 22), 第1-8页, 图2	1-8
Y	CN 109977237 A (南京擎盾信息科技有限公司) 2019年7月5日 (2019 - 07 - 05) 说明书第[0022]-[0026]段	1-8
A	CN 110781254 A (四川大学 等) 2020年2月11日 (2020 - 02 - 11) 全文	1-8
A	CN 108874878 A (众安信息技术服务有限公司) 2018年11月23日 (2018 - 11 - 23) 全文	1-8
A	US 2020073933 A1 (NATIONAL UNIVERSITY OF DEFENSE TECHNOLOGY) 2020年3月5日 (2020 - 03 - 05) 全文	1-8

☐ 其余文件在C栏的续页中列出。☒ 见同族专利附件。

* 引用文件的具体类型:

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件(如具体说明的)

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期

2022年1月19日

国际检索报告邮寄日期

2022年2月11日

ISA/CN的名称和邮寄地址

中国国家知识产权局(ISA/CN)

中国北京市海淀区蓟门桥西土城路6号 100088

传真号 (86-10)62019451

受权官员

金霞

电话号码 86-(10)-53961524

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2021/116053

检索报告引用的专利文件			公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN	113204649	A	2021年8月3日	无	
CN	111858940	A	2020年10月30日	无	
CN	109977237	A	2019年7月5日	无	
CN	110781254	A	2020年2月11日	无	
CN	108874878	A	2018年11月23日	无	
US	2020073933	A1	2020年3月5日	CN 109165385 A	2019年1月8日