

事件检测

技术领域

[01]本说明书涉及自然语言处理领域，特别涉及一种事件检测方法和系统。

背景技术

[02]事件检测或事件抽取是人工智能技术的一项重要应用，其可以从海量数据中高效的获取人们关心的事件。例如，从大量的金融领域的新闻、报道中及时抽取目标风险事件，帮助投资人有效规避投资风险。然而，随着新事件的出现，已有的事件检测或抽取算法可能也要针对新事件进行更新，进而增加技术更新或升级的成本。

[03]因此，希望提供一种事件检测方法和系统，能够有效从海量数据中识别已知事件外，当出现新事件时，也能通过花费较少的成本实现对新事件的兼容。

发明内容

[04]本说明书一个方面提供一种事件检测方法，所述方法包括：获取待处理文本；基于抽取模型从待处理文本中抽取一组或多组实例数据；其中，每组实例数据包括第一实体实例、所述第一实体实例对应的第一实体类型、第二实体实例、所述第二实体实例对应的第二实体类型以及两个实体类型之间的关系描述；基于一组或多组实例数据，确定一个或多个抽取三元组，进而得到抽取图谱；其中，抽取三元组包括实例数据中的第一实体类型、第二实体类型以及两个实体类型之间的关系描述；获取一个或多个候选事件的图谱本体定义数据，并基于此得到对应每个候选事件的本体定义图谱；其中，事件的图谱本体定义数据包括用于定义实体的实体类型以及用于定义实体类型间关系的关系描述；确定抽取图谱分别与一个或多个候选事件的本体定义图谱的相似度；基于各相似度，从一个或多个候选事件中确定待处理文本对应的事件。

[05]本说明书另一个方面提供一种通过事件检测系统，所述系统包括：文本获取模块，用于获取待处理文本；抽取模块，用于基于抽取模型从待处理文本中抽取一组或多组实例数据；其中，每组实例数据包括第一实体实例、所述第一实体实例对应的第一实体类型、第二实体实例、所述第二实体实例对应的第二实体类型以及两个实体类型之间的关系描述；抽取图谱获取模块，用于基于一组或多组实例数据，确定一个或多个抽取三元组，进而得到抽取图谱；其中，抽取三元组包括实例数据中的第一实体类型、第二实体类型以及两个实体类型之间的关系描述；本体定义图谱获取模块，用于获取一个或多个

候选事件的图谱本体定义数据并基于此得到对应每个候选事件的图谱本体定义图谱；其中，事件的图谱本体定义数据包括用于定义实体的实体类型以及用于定义实体类型间关系的关系描述；相似度确定模块，用于确定抽取图谱分别与一个或多个候选事件的图谱本体定义图谱的相似度；事件确定模块，用于基于各相似度，从一个或多个候选事件中确定待处理文本对应的事件。

[06]本说明书另一个方面提供一种计算机可读存储介质，其特征在于，所述存储介质存储计算机指令，当所述计算机指令被处理器执行时实现事件检测方法。

[07]本说明书另一个方面提供一种事件检测装置，其特征在于，所述装置包括至少一个处理器以及至少一个存储器；所述至少一个存储器用于存储计算机指令；所述至少一个处理器用于执行所述计算机指令中的至少部分指令以实现事件检测方法。

附图说明

[08]本说明书将以示例性实施例的方式进一步说明，这些示例性实施例将通过附图进行详细描述。这些实施例并非限制性的，在这些实施例中，相同的编号表示相同的结构。

[09]图 1 是根据本说明书一些实施例所示的事件检测方法的示例性流程图；

[10]图 2 是根据本说明书一些实施例所示的事件的图谱本体定义数据的示意图；

[11]图 3 是根据本说明书一些实施例所示的抽取模型的结构示意图；

[12]图 4 是根据本说明书一些实施例所示的确定抽取图谱分别与一个或多个候选事件的图谱本体定义图谱的相似度的示意图。

具体实施方式

[13]为了更清楚地说明本说明书实施例的技术方案，下面将对实施例描述中所需要使用的附图作简单的介绍。显而易见地，下面描述中的附图仅仅是本说明书的一些示例或实施例，对于本领域的普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图将本说明书应用于其它类似情景。除非从语言环境中显而易见或另做说明，图中相同标号代表相同结构或操作。

[14]应当理解，本说明书中所使用的“系统”、“装置”、“单元”和/或“模组”是用于区分不同级别的不同组件、元件、部件、部分或装配的一种方法。然而，如果其他词语可实现相同的目的，则可通过其他表达来替换所述词语。

[15]如本说明书和权利要求书所示，除非上下文明确提示例外情形，“一”、“一个”、“一种”和/或“该”等词并非特指单数，也可包括复数。一般说来，术语“包括”与“包含”仅提示包括已明确标识的步骤和元素，而这些步骤和元素不构成一个排它性的罗列，方法或者设备也可能包含其它的步骤或元素。

[16]本说明书中使用了流程图用来说明根据本说明书的实施例的系统所执行的操作。应当理解的是，前面或后面操作不一定按照顺序来精确地执行。相反，可以按照倒序或同时处理各个步骤。同时，也可以将其他操作添加到这些过程中，或从这些过程移除某一步或数步操作。

[17]在信息迸发的时代，每天都会有大量信息出现，加之信息的表达形式灵活多变，因此，如何使已有的事件检测方案能较好的兼容新事件检测是一个值得研究的问题。在一些实施例中，可以通过事件检测模型对文本数据进行处理，以从文本中检测和/或抽取事件，从而帮助用户快速筛选感兴趣信息。例如，事件检测模型可以从新闻报道“...A公司起诉B公司，一审判决A公司败诉...”中检测出事件“A公司败诉于B公司”。然而，事件检测模型受限于训练语料。当出现新的事件时，事件检测模型可能因为未曾“见过”新事件的关键词或触发词，进而难以从新闻报道等数据中发现新的事件。为此，一种解决方案是不断搜集新的训练语料，对事件检测模型进行继续训练，提高其检测能力。然而搜集新的训练语料，或者对模型进行再训练，都会耗费较多的人力和时间成本。

[18]本说明书的一些实施例提出了一种事件检测方案，其基于事件图谱本体定义数据检测事件或事件要素，事件图谱本体定义数据中主要包括实体类型和基于实体类型之间关系的概括描述。因此，在一定程度上可以做到至少对于相同领域的新出现的事件，只需要为新事件定义事件本体定义数据，便可实现对新事件的有效检测，大大降低了技术升级或更新换代的成本。

[19]图1是根据本说明书一些实施例所示的事件检测方法的示例性流程图。在一些实施例中，流程100可以由处理设备执行，或者由设置于处理设备上的事件检测系统实现。所述事件检测系统可以包括文本获取模块、抽取模块、抽取图谱获取模块、本体定义图谱获取模块、相似度确定模块以及事件确定模块。

[20]如图1所示，事件检测方法100可以包括以下步骤。

[21]步骤110，获取待处理文本。

[22]具体地，该步骤110可以由文本获取模块执行。

[23]待处理文本是需要检测事件的文本。在一些实施例中，待处理文本可以是篇章级文本。例如，新闻、论文、研报、时事评论等。在一些实施例中，待识别文本可以是句子级文本。例如，前述任意篇章级文本中包括的语句。示例性地，待处理文本可以是新闻文本“A公司的股权被冻结，120亿新能源投资何以为继？...”。

[24]在一些实施例中，文本获取模块可以直接从文字形式的信息中获取待处理文本。例如，文本获取模块可以从文本数据库中获取待处理文本。又例如，文本获取模块还可以从网页文本中爬取待处理文本。

[25]在一些实施例中，文本获取模块还可以基于文字识别技术从图片消息获取的待处理文本。在一些实施例中，待处理文本还可以基于自动语音识别（Automatic Speech Recognition, ASR）技术从语音信息中获取。

[26]在一些实施例中，待处理文本可以包括多个字或词，如，中文字符、日文字符或英文等西文单词等。

[27]步骤120，基于抽取模型从待处理文本中抽取一组或多组实例数据。

[28]具体地，该步骤120可以由抽取模块执行。

[29]实例数据由待处理文本中的实体类型、实体实例（即实体对应的数据实例）以及实体类型之间的关系描述构成。

[30]其中，实体类型是对客观个体和行为的广泛抽象，其可以指物理世界中的有形物体，如人、执法机关、公司主体等，也可以指无形的对象，如话语、歌曲、电影、资产、金额等，还可以指行为动作，如管理动词、惩罚动词、保全动词等。实体实例可以是实体类型的抽象概念下实际存在的例子，如公司主体可以具体是公司A、公司B、公司C等，资产可以具体是不动产、200亿元股权、10万元现金等，保全动词可以具体是查封、冻结、扣押等。示例性地，待处理文本“A公司的股权被冻结，120亿新能源投资何以为继”中的实体实例包括：A公司、股权、冻结...，对应的实体类型包括：公司主体、资产、保全动词...。

[31]实体实例之间可以具有关系，实体实例之间的关系可以由其实体类型之间的关系描述来定义。例如，资产这一实体类型与公司主体这一实体类型之间的关系是属于，惩罚动词这一实体类型与资产这一实体类型之间的关系可以是作用对象。进而，实体实例之间也可以具有相应的关系描述。前述待处理文本中，“股权”属于“A公司”、“冻结”的作用对象是“股权”等。

[32]在一些实施例中，可以定义更加抽象和适用范围广的关系描述。例如，关系描述可以包括动宾关系、主谓关系、定中关系和修饰关系。其中，当第一实体类型是动词类实体，第二实体类型是名词类实体时，两者的关系描述可以为动宾关系(Verb-Object, VOB)。例如，第一实体类型“保全动词”和第二实体类型“资产”之间的关系描述为动宾关系。当第一实体类型是名词类型，第二实体类型是动词类型时，关系描述可以是主谓关系(Subject-Verb, SBV)。例如，第一实体类型“执法机关”和第二实体类型“惩罚动词”之间的关系描述为主谓关系。当第一实体类型是领属、范围、质料、形式、性质、数量、用途、时间、处所等修饰词等，第二实体类型是中心语时，关系描述可以是定中关系(Attribute, ATT)。例如，第一实体类型“公司主体”和第二实体类型“资产”之间的关系描述为定中关系。当第一实体类型是被修饰词，第二实体类型是修饰词时，关系描述可以是修饰关系(Modify, MOD)。例如，第一实体类型“资产”和第二实体类型“涉及金额”之间的关系描述为修饰关系。

[33]在一些实施例中，每组实例数据可以包括第一实体实例、第一实体类型、第二实体实例、第二实体类型以及两个实体类型之间的关系描述。

[34]继续上述示例，第1组实例数据中，第一实体实例可以是股权，对应的第二实体实例可以是冻结。第一实体类型和第二实体类型分别是第一实体实例和第二实体实例对应的广泛抽象。例如，第1组实例数据中，第一实体实例“股权”对应的第一实体类型是资产，第二实体实例“冻结”对应的第二实体类型是保全动词。第一实体实例“冻结”对应的第一实体类型“保全动词”和第二实体实例“股权”对应的第二实体类型“资产”之间的关系描述为动宾关系。第1组实例数据可以表示为【保全动词：冻结，VOB，资产：股权】。

[35]在一些实施例中，抽取模块可以从待处理文本中抽取多组实例数据。

[36]可以理解，第一实体实例和第二实体实例之间的关系是相对的。在一些实施例中，第一实体实例和第二实体实例可以交换，作为新的一组实例数据。继续上述示例，第2组实例数据中，第一实体实例可以是第1组实例数据中的第二实体实例“股权”，第二实体实例是第1组实例数据中的第一实体实例“冻结”；对应的第一实体类型和第二实体类型分别为“资产”和“保全动词”。相应地，两个实体类型之间的关系描述相应发生改变。例如，第2组实例数据中，第一实体类型“资产”和第二实体类型“保全动词”之间的关系描述为主谓关系。第2组实例数据可以表示为【资产：股权，SBV，保全动词：冻结】。

[37]在一些实施例中，多组实例数据的实体实例和关系描述可以部分相同。继续上述示例，第3组实例数据可以是【资产：股权，MOD，涉及金额：120亿】，其中，第一实

体实例和第2组实例数据中的第一实体实例相同。

[38]具体地，抽取模块可以利用抽取模型处理待处理文本，得到待处理文本的标注序列以及关系矩阵。

[39]标注序列用于标记待处理文本中属于实体实例的字或词，以及所述字或词所属的实体类型。关系矩阵用于标记待处理文本中任意两个字或词之间的关系描述。

[40]在一些实施例中，抽取模型包括以下模型中的一种或多：BERT、Transformer、StanfordNLP 或 LTP。

[41]关于抽取模型的详细描述可以参见图3及其相关描述，在此不再赘述。

[42]抽取模块可以基于标注序列确定待处理文本中的实体实例及其实体类型。

[43]如图3所示，抽取模型处理“A公司的股权被冻结，120亿新能源投资何以为继？”，获得标注序列：“B-co”、“I-co”...“B-pro”、“I-pro”...“B-pre”、“I-pre”...“O”、“O”、“O”、“O”。抽取模块可以基于其中的实体标注“B-co”、“I-co”、“B-pro”、“I-pro”、“B-pre”、“I-pre”获取对应的实体实例：A公司，股权，冻结，及其实体类型公司主体、资产、保全动词。

[44]抽取模块可以基于关系矩阵确定待处理文本中任意两个实体实例之间的关系描述，并将其作为对应两个实体类型之间的关系描述。如图3所示，待处理文本中每个字或词可以对应一个关系向量 r ，该关系向量的维度可以与待处理文本的总字数或总词数相同，关系向量中的元素反映向量对应的字或词与待处理文本中其他各字或各词的关系描述，关系向量还可以包括反映其对应的字或词与自身关系描述的元素，如元素值可以默认为null。多个字或词的关系向量组成了所述关系矩阵。可以理解，关系矩阵中的元素可以包括前述的VOB、MOD等关系描述，也可以包括null，其中null表示无效或空。在一些实施例中，抽取模块可以基于两个实体实例各自的首字或首词之间的关系描述，确定这两个实体实例或对应的两个实体类型之间的关系描述。

[45]如图3所示，抽取模块可以基于关系矩阵中的实体实例“A公司”的首字“A”对应的关系向量 $[r_1]$ 中和实体实例“股权”的首字“股”对应的元素，确定“A公司”和“股权”之间的关系描述为定中关系。进一步地，将定中关系作为“A公司”对应的实体类型“公司主体”和“股权”对应的实体类型“资产”之间的关系描述。又例如，抽取模块可以基于关系矩阵中的实体实例“冻结”的首字“冻”对应的关系向量 $[r_{v1}]$ 中和实体实例“股权”的首字“股”对应的元素，确定“冻结”和“股权”之间的关系描述为动宾关系。进一步地，将动宾关系作为“冻结”对应的实体类型“保全动词”和“股权”对应的实体类型“资产”之间的关系描述。

又例如，关系矩阵中“冻”对应的关系向量【 r_{vi} 】中和“被”对应的元素为 null，表示两者无明确关系。

[46]步骤 130，基于一组或多组实例数据，确定一个或多个抽取三元组，进而得到抽取图谱。

[47]具体地，该步骤 130 可以由抽取图谱获取模块执行。

[48]抽取三元组是从实例数据抽取的三个元素的集合。在一些实施例中，抽取三元组包括实例数据中的第一实体类型、第二实体类型以及两个实体类型之间的关系描述。

[49]例如，抽取图谱获取模块可以从第 1 组实例数据【保全动词：冻结，VOB，资产：股权】抽取第 1 个三元组【保全动词，VOB，资产】，从第 2 个实例数据【资产：股权，SBV，保全动词：冻结】中抽取第 2 组三元组【资产，SBV，保全动词】，从第 3 组实例数据【资产：股权，MOD，涉及金额：120 亿】中抽取第 3 个三元组【资产，MOD，涉及金额】...

[50]进一步地，抽取图谱获取模块可以基于一个或多个抽取三元组，构建抽取图谱。

[51]抽取图谱是由一个或多个三元组中实体实例以及实体实例类型之间关系描述构成的网络图。在一些实施例中，抽取图谱中一个或多个三元组的实体类型可以用节点表示，一个或多个三元组的实体类型之间的关系描述可以用连接对应节点的边表示。

[52]例如，抽取图谱中可以以第 1 个三元组【保全动词，VOB，资产】中的实体类型“保全动词”和“资产”分别为两个节点，以“保全动词”和“资产”之间的关系描述“VOB”为连接所述两个节点的边。

[53]在一些实施例中，多个三元组中相同的实体类型可以在抽取图谱中用同一个节点表示。例如，第 3 个三元组【资产，MOD，涉及金额】中实体类型“资产”和第 1 个三元组【保全动词，VOB，资产】中的实体类型“资产”相同，则在抽取图谱中第 3 个三元组中实体类型“资产”和第 1 个三元组中实体类型“资产”可以用同一个节点表示。

[54]在一些实施例中，抽取图谱中的边具有方向性，由第一实体类型指向第二实体类型。例如，第 1 个三元组中的关系描述“VOB”对应的边是“保全动词”指向“资产”。第 2 个三元组【资产，SBV，保全动词】中的关系描述“SBV”对应的边是“资产”指向“保全动词”。

[55]步骤 140，获取一个或多个候选事件的图谱本体定义数据，并基于此得到对应每个候选事件的图谱本体定义图谱。

[56]具体地，该步骤 140 可以由本体定义图谱获取模块执行。

[57]本体定义图谱是指由一系列实体实例类型以及实体实例类型之间关系描述构成的图。在一些实施例中，本体定义图谱中的实体实例类型可以用节点表示，实体实例类型之间的关系描述可以用连接节点的边表示。

[58]多个候选事件可以分别对应多种事件类型。例如，第一候选事件、第二候选事件、...第 N 候选事件可以分别对应收购事件类型、败诉事件类型、...保全事件类型。

[59]在一些实施例中，每个候选事件可以对应一个本体定义图谱。事件的本体定义图谱可以基于事件的图谱本体定义数据，如 schema，生成。而事件的图谱本体定义数据可以根据事件的一般要素，人工制定或编写。

[60]图 2 是根据本说明书一些实施例所示的事件的图谱本体定义数据的示意图。

[61]图谱本体定义数据是指对本体定义图谱包括的实体类型、实体类型之间关系描述进行定义的数据。

[62]相应地，每组图谱本体定义数据和一个候选事件类型相对应。如图 2 所示，收购类图谱本体定义数据、败诉类图谱本体定义数据...保全类图谱本体定义数据，分别对应收购事件类型、败诉事件类型、...保全事件类型。

[63]在一些实施例中，事件的图谱本体定义数据包括用于定义实体的实体类型以及用于定义实体类型间关系的关系描述。

[64]在一些实施例中，图谱本体定义数据中的实体类型可以由词表或抽取规则定义属于该实体类型的实体实例。可以理解，满足抽取规则的实体实例即属于对应的实体类型。具体的，对应于枚举类的实体类型，则可以由词表进行定义。如实体类型“保全动词”对应有词表：扣押、冻结、查封等。对于无法枚举的实体类型，则可以通过正则表达式，关键词匹配、限制数据格式等抽取规则定义。如实体类型“执法机关”，可以定义抽取规则为基于关键词“法院”匹配进行文本抽取，又如实体类型“涉及金额”，可以基于数据格式进行抽取。

[65]图谱本体定义数据可以包括关系描述，并通过定义数据将关系描述与图谱本体定义数据中的实体类型结合，进而定义不同实体类型之间的关系描述。以图 2 中保全类图谱定义数据为例，保全类图谱定义数据中包含 6 条定义数据，第 1 条定义数据指定第一实体类型“保全动词”，第二实体类型“公司主体”的关系描述定义为“VOB”；第 2 条定义数据指定第一实体类型“公司主体”，第二实体类型“保全动词”之间关系描述定义为“ATT”；

第3条定义数据指定第一实体类型“保全动词”，第二实体类型“资产”之间的关系描述定义为“VOB”；....第6条定义数据指定第一实体类型“资产”，第二实体类型“保全动词”之间的关系描述定义为“SBV”。

[66]图谱本体定义数据可以基于预设的本体定义数据集获取。具体地，事件的图谱本体定义数据中的实体类型以及关系描述来自预设的本体定义数据集。

[67]预设的本体定义数据集可以是针对某一特定领域，如金融领域或教育领域，制定的实体类型以及关系描述的集合。可以认为，本体定义数据集可以包括了对应领域较为全面的实体类型及关系描述，使得该领域中的不同事件的实体类型和关系描述都可以在该数据集中找到，或者使得该数据集中的实体类型和关系描述可以通用。在为不同事件制定图谱本体定义数据时，可以从该数据集中选取实体类型以及关系描述，通过词表或抽取规则对实体类型做进一步定义，以及通过定义数据指定实体类型间的关系描述则可以获得各事件自己的图谱本体定义数据。

[68]在一些实施例中，还可以基于所述预设的本体定义数据集训练抽取模型，进而使得步骤120抽取的实体实例及其关系描述直接映射为候选事件的图谱本体定义数据中的实体类型及关系描述，进一步提高后续的图匹配准确度。关于训练抽取模型的详细描述可以参见图3及其相关描述，在此不再赘述。

[69]在一些实施例中，本体定义图谱获取模块还可以基于一个或多个候选事件的图谱本体定义获取一个或多个候选事件的本体定义图谱。

[70]具体地，对于每一个候选事件类型的图谱本体定义数据，可以将其中的第一实体类型和第二实体类型作为对应候选事件类型的本体定义图谱中的节点，将基于定义数据为对应节点建立边。

[71]如图2所示，本体定义图谱获取模块可以将保全类候选事件的图谱本体定义数据中第1条定义数据中的第一实体类型“保全动词”和第二实体类型“公司主体”作为保全类候选事件的本体定义图谱中的两个节点，并将在节点“保全动词”与节点“公司主体”建立边“VOB”。

[72]与抽取图谱类似地，候选事件的本体定义图谱中的边具有方向性，由第一实体类型指向第二实体类型。继续以图2保全类候选事件对应的本体定义图谱为例，第1条定义数据中的关系描述定义“VOB”对应的边是“保全动词”指向“资产”。第2条定义数据中的关系描述定义“SBV”对应的边是“资产”指向“保全动词”。

[73]与抽取图谱类似地，多条定义数据中相同的实体定义可以在候选事件的本体定义图谱中用同一个节点表示，详细描述可以参见步骤 130，在此不再赘述。

[74]类似地，本体定义图谱获取模块可以获取收购类候选事件的本体定义图谱、败诉类本体定义图谱、....。

[75]步骤 150，确定抽取图谱分别与一个或多个候选事件的本体定义图谱的相似度。

[76]具体地，该步骤 150 可以由相似度确定模块执行。

[77]在一些实施例中，相似度确定模块可以利用图匹配模型处理所述抽取图谱以及所述候选事件的本体定义图谱，得到两者之间的相似度。

[78]在一些实施例中，图匹配模型可以包括但不限于图匹配网络（Graph Matching Network，GMN）模型、图神经网络（Graph Neural Network，GNN）模型、图卷积神经网络（Graph Convolutional Network，GCN）和图嵌入模型(Graph Embedding Models，GEM)等。

[79]以 GMN 模型为例，GMN 模型可以先分别获取抽取图谱和候选事件的本体定义图谱的每个节点以及每条边的（初始）表示向量，基于 Attention 机制获取抽取图谱的每个表示向量和候选事件的本体定义图谱的每个表示向量之间的注意力大小，然后聚合抽取图谱和候选事件的本体定义图谱的每个节点以及每条边的表示向量和所述每个表示向量之间的注意力大小，获取包含了抽取图谱和候选事件的本体定义图谱的节点、边以及相互关系的跨图信息（cross-graph），再基于跨图信息获取所述相似度。

[80]图 4 是根据本说明书一些实施例所示的确定抽取图谱分别与一个或多个候选事件的本体定义图谱的相似度的示意图。

[81]以图 4 为例，相似度确定模块可以先分别获取抽取图谱 420 中的节点“公司主体”、“保全动词”、“资产”和边“E（公司主体，VOB，保全动词）”、边“E（保全动词，SBV，保全动词）”...等对应的表示向量 N1、N2、N3、E1、E2、...，以及保全类候选事件的本体定义图谱 410 中的节点“公司主体”、“保全动词”、“资产”、“涉事金额”和边“e（公司主体，VOB，保全动词）”、边“e（保全动词，SBV，保全动词）”...等对应的表示向量 n1、n2、n3、n4、e1、e2...，然后分别获取 N1 和 n1、n2、n3、n4、e1、e2...之间的注意力大小，获取注意力向量 a1；类似地，可以分别获取 N2 和 n1、n2、n3、n4、e1、e2...之间的注意力大小，获取注意力向量 a2；.....进一步地，GMN 模型（即匹配模型 430）可以聚合抽取图谱的每个节点以及每条边的表示向量 N1、N2、N3、E1、E2、...，候选

事件的本体定义图谱的每个节点以及每条边的表示向量 $n1$ 、 $n2$ 、 $n3$ 、 $n4$ 、 $e1$ 、 $e2$...，和注意力向量 $a1$ 、 $a2$ 、 $a3$ 、...，获取跨图信息（cross-graph），然后基于跨图信息获取抽取图谱 420 和保全类候选事件的本体定义图谱 410 之间的相似度 440 为 0.8。

[82]类似地，相似度确定模块可以获取抽取图谱 420 和收购类候选事件的本体定义图谱、败诉类候选事件的本体定义图谱...的相似度。

[83]在一些实施例中，还可以基于 GEM 或 GCN 处理抽取图谱和候选事件的本体定义图谱，分别得到两个图谱的向量表示，再通过计算两个向量表示的距离确定两者的相似度。

[84]步骤 160，基于各相似度，从所述一个或多个候选事件中确定所述待处理文本对应的事件。

[85]具体地，该步骤 160 可以由事件确定模块执行。

[86]在一些实施例中，事件确定模块可以确定多个相似度中最大值对应的候选事件，将该候选事件作为待处理文本对应的事件。

[87]例如，抽取图谱和保全类候选事件的本体定义图谱、收购类候选事件的本体定义图谱、败诉类候选事件的本体定义图谱...的相似度分别为 0.8、0.5、0.4...，则确定相似度最大值 0.8 对应的保全类事件作为待处理文本对应的事件。

[88]在一些实施例中，事件确定模块还可以进一步用于确定待处理文本中的事件要素。

[89]具体地，事件检测系统可以基于所述一组或多组实例数据，确定一个或多个实例三元组。所述实例三元组包括实例数据中的第一实体实例、第二实体实例及其分别对应的两个实体类型之间的关系描述。例如，第 1 组实例三元组可以包括待抽取文本“A 公司的股权被冻结，120 亿新能源投资何以为继？...”对应的第 1 组实例数据【保全动词：冻结，VOB，资产：股权】中的第一实体实例“冻结”、第二实体实例“股权”及其分别对应的两个实体类型之间的关系描述“VOB”，即第 1 组实例三元组为【冻结，VOB，资产】；类似地，第 2 组实例三元组为【股权，SBV，冻结】、第 3 组实例三元组为【股权，MOD，120 亿】，...【A 公司，ATT，股权】。

[90]进一步地，事件检测系统可以基于所述一组或多组实例三元组，确定待处理文本对应事件的事件要素。

[91]事件要素包括组成事件的元素和元素之间的关系。在一些实施例中，可以将每组实例三元组的实例实体作为事件中的元素，将事件中的元素基于元素之间的关系进行结构

化表达。

[92]例如，先分别基于第1组实例三元组【冻结，VOB，股权】、第2组实例三元组【股权，SBV，冻结】、第3组实例三元组【股权，MOD，120 亿】，...【A 公司，ATT，股权】获取元素“冻结”、“股权”、“120 亿”、“A 公司”...，再基于元素之间的关系“VOB”、“SBV”等，将元素结构化表达为“A 公司股权冻结，120 亿...”，进而得到该保全类事件的事件要素。

[93]图3是根据本说明书一些实施例所示的抽取模型的结构示意图。

[94]在一些实施例中，抽取模块可以利用抽取模型处理待处理文本，得到待处理文本的标注序列以及关系矩阵。

[95]如图3所示，抽取模型300可以包括特征提取层310、标注序列层320和关系识别层330。

[96]具体地，特征提取层310可以提取待处理文本的特征向量。

[97]在一些实施例中，特征提取层310可以对待处理文本进行编码，得到融合了待处理文本信息的特征向量。

[98]在一些实施例中，在特征提取层310对待处理文本进行编码之前，可以对待处理文本做如下处理：在待处理文本之前添加[CLS]；在待处理文本中每句话之间通过分隔符[SEP]分割，以进行区分。例如，待处理文本“A 公司的股权被冻结，120 亿新能源投资何以为继”处理后为“[CLS]A 公司的股权被冻结[SEP]120 亿新能源投资何以为继”。

[99]在一些实施例中，特征提取层310可以基于待处理文本分别得到对应的字符向量和位置向量。

[100]字符向量(token embedding)是表征待处理文本的字符特征信息的向量。如图3所示，待处理文本“A 公司的股权被冻结，120 亿新能源投资何以为继”包括的22个字符【w₁】【w₂】...【w_{a1}】【w_{a2}】...【w_{as}】的特征信息可以分别用22个字符向量【t₁】【t₂】...【t_{a1}】【t_{a2}】...【t_{as}】表征。示例性地，字符【A】的特征信息可以用字符向量[2,3,3]表征。在实际应用场景中，向量表示的维度可以更高。在一些实施例中，字符向量可以通过查询词向量表或词嵌入模型获取。在一些实施例中，词嵌入模型可以包括但不限于：Word2vec 模型、词频-逆向文件频率模型(Term Frequency-Inverse Document Frequency, TF-IDF)或 SSWE-C(skip-gram based combined-sentiment word embedding)模型等。

[101]位置向量 (position embedding) 是反映该字符在待处理文本中位置的向量, 如指示该字符是待处理文本中的第 1 个字符, 或第 2 个字符等。在一些实施例中, 待处理文本的位置向量可以通过余弦正弦编码获取。在一些实施例中, 还可以包括分段向量 (segment embedding), 反映字符所在的分段。如字符【A】位于待处理文本的第 1 句 (分段) 中。

[102]进一步地, 特征提取层 310 可以先将待处理文本的各类向量进行融合, 如拼接或叠加, 再对融合后的向量进行编码, 得到特征向量。

[103]如图 3 所示, 特征提取层 310 可以基于字符向量【t₁】【t₂】...【t_{a1}】【t_{a2}】...【t_{as}】和位置向量 (未示出), 获取特征向量【T₁】【T₂】...【T_{a1}】【T_{a2}】...【T_{as}】。

[104]示例性的特征提取层可以由 BERT 模型或 Transformer 实现。

[105]进一步地, 标注序列层 320 可以基于特征向量获取标注序列。

[106]标注序列是与待处理文本中多个字或多个词分别对应的多个标注按照顺序排列的结果。在一些实施例中, 标注可以包括实体标注, 用于指示对应的字或词是否属于实体实例, 进一步的, 实体标注可以进一步分为公司主体实体标注、资产实体标注, 以便进一步指示对应的字或词所属的实体类型。由此, 标注序列可以用于标记待处理文本中属于实体实例的字或词, 以及所述字或词所属的实体类型。

[107]在一些实施例中, 所述实体标注可以是汉字、数字、字母和符号等中的至少一种。例如, 可以用 B 表示实体实例的首字或首词, I 表示实体实例的非首字或非首词。又例如, 实体标注 B-co 或 I-co 可以标记待处理文本中实体类型为“公司主体”的字或词。又例如, 实体标注 B-pro 或 I-pro 可以标记待处理文本中实体类型为“资产”的字或词。

[108]如图 3 所示, 标注序列层 320 可以基于特征向量【T₁】【T₂】...【T_{a1}】【T_{a2}】...【T_{as}】, 分别将待处理文本“A 公司的股权被冻结, 120 亿新能源投资何以为继”中的实体实例“A 公司、股权、冻结...”标记为“B-co、I-co, B-pro、I-pro, B-pre、I-pre, ...”, 分别表示“公司主体的第一个字、公司主体的非第一个字, 资产的第一个字、资产的非第一个字, 保全动词的第一个字、保全动词的非第一个字”。

[109]在一些实施例中, 标注还可以包括非实体标注。非实体标注也可以是汉字、数字、字母和符号等中的至少一种。待处理文本中不属于实体实例的字或词可以用相同的非实体标注进行标注。如图 3 所示, 标注序列层 320 用 4 个“O”标注待处理文本中不属于实体实例的词“何以为继”。在一些实施例中, 待处理文本中不属于实体实例的字或词也可

以不进行任何标注。

[110]具体地,标注序列层 320 可以基于特征向量,获取待处理文本中每个字或词分别属于不同实体类型的概率和不属于任何实体的概率,然后将概率最大值对应的实体类型的实体标注或者不属于实体的非实体标注作为所述字或词的标注。

[111]以图 3 为例,标注序列层 320 可以基于特征向量 $\mathbf{T}_1$, 获取“A”属于公司主体第一个字的概率为 0.8、属于公司主体非第一个字的概率为 0.5、属于人的第一个字的概率为 0.3、属于人的非第一个字的概率为 0.3、属于冻结动词的第一个字的概率为....., 不属于实体的概率为 0.2, 然后将概率最大值 0.8 对应的实体类型“公司主体”第一个字的实体标注“B-co”作为“A”字的实体标注。

[112]与此类似地,标注序列层 320 可以获取待处理文本“A 公司的股权被冻结, 120 亿新能源投资何以为继”中每个字或词的标注(实体标注或非实体标注), 并按照所述字或词在待处理文本中的顺序排列, 从而获取标注序列: “B-co”、“I-co”...“B-pro”、“I-pro”...“B-pre”、“I-pre”...“O”、“O”、“O”、“O”。

[113]在一些实施例中,标注序列层 320 可以包括但不限于 N 元(N-Gram)模型、条件随机场(Conditional Random Fields, CRF)模型和隐马尔可夫模型(Hidden Markov Model, HMM)中的一种或多种。

[114]更进一步地,关系识别层 330 可以基于特征向量和标注序列,获取关系矩阵。

[115]关系矩阵可以用于标记待处理文本中任意两个字或词之间的关系描述。在一些实施例中,关系矩阵中的每一个元素可以标记两个字或词之间的关系描述。所述关系矩阵的维度基于待处理文本的字符个数确定。例如,待处理文本包含 N 个字或词,则关系矩阵的维度为 $N \times N$ 。其中,关系矩阵中的每一列 $1 \times N$ 的向量,可以标记待处理文本中其中一个字或词和其余所有字或词之间的关系描述,N 列 $1 \times N$ 的向量可以分别标记待处理文本中 N 个字或词和其余所有字或词之间的关系描述。

[116]在一些实施例中,关系识别层 330 可以先通过词嵌入网络将标注序列中每一个标注嵌入对应的特征向量中,获取标记向量。

[117]如图 3 所示,关系识别层 330 可以先将标注序列中的标注:“B-co”、“I-co”...“B-pro”、“I-pro”...“B-pre”、“I-pre”...“O”、“O”、“O”、“O”分别嵌入特征向量 $\mathbf{T}_1$ $\mathbf{T}_2$... $\mathbf{T}_{a1}$ $\mathbf{T}_{a2}$ $\mathbf{T}_{as}$, 获取标记向量 $\mathbf{e}_1$ $\mathbf{e}_2$... $\mathbf{e}_{a1}$ $\mathbf{e}_{a2}$ $\mathbf{e}_{as}$ 。

[118]进一步地,关系识别层 330 可以基于标记向量,通过多头非线性激活层

(Multi-sigmoid layer)，获取关系矩阵。关于关系矩阵的更多内容还可以参见步骤 120 的相关描述。

[119]上述实施例给出了抽取模型的一种实现结构，在又一些实施例中，抽取模型可以基于端到端的模型实现，如基于 BERT 的多头选择模型、斯坦福中文语法分析工具 StanfordNLP 或者哈工大中文语言分析工具 LTP 实现。

[120]在一些实施例中，所述抽取模型可以使用训练样本训练得到。具体的，将带有标签的训练样本输入抽取模型，通过训练更新抽取模型的参数。

[121]在一些实施例中，训练样本的标签可以基于预设的本体定义数据集中的实体类型以及关系描述确定。继续上述示例，训练样本“X 公司股权被冻结”的标签可以包括该训练样本已知的标注序列及关系矩阵，其中，已知的标注序列指示“X 公司”、“股权”和“冻结”对应的实体类型分别为预设的本体定义数据集中的实体类型“公司主体”、“资产”和“保全动词”，已知的关系矩阵指示“公司主体”和“资产”之间的关系描述“ATT”，“资产”和“保全动词”之间的关系描述“SBV”等。

[122]进一步，可以利用抽取模型处理所述训练样本，得到模型预测的标注序列以及关系矩阵，参照已知的标注序列和关系矩阵调整抽取模型的参数，以减小预测结果与已知标签之间的差异。

[123]关于预设的本体定义数据集中的详细描述可以参见步骤 140，在此不再赘述。

[124]本说明书实施例还提供一种计算机可读存储介质。所述存储介质存储计算机指令，当计算机读取存储介质中的计算机指令后，计算机实现前述的事件检测方法。

[125]本说明书实施例还提供一种事件检测装置，其特征在于，所述装置包括至少一个处理器以及至少一个存储器；所述至少一个存储器用于存储计算机指令；所述至少一个处理器用于执行所述计算机指令中的至少部分指令以实现前述的事件检测方法。

[126]本说明书实施例可能带来的有益效果包括但不限于：（1）基于事件图谱本体定义数据检测事件或事件要素，对于新出现的事件，只需要花费较低的成本即可实现兼容检测。（2）基于预设的本体定义数据集训练抽取模型，从而可以使得抽取模型将待抽取文本中的实例数据映射为事件图谱本体定义数据中实体类型和关系描述，从而提高后续图匹配的准确性（3）实体类型之间的关系描述进一步抽象和概括，提高了新事件检测的兼容性。

[127]需要说明的是，不同实施例可能产生的有益效果不同，在不同的实施例里，可能产

生的有益效果可以是以上任意一种或几种的组合，也可以是其他任何可能获得的有益效果。

[128]上文已对基本概念做了描述，显然，对于本领域技术人员来说，上述详细披露仅仅作为示例，而并不构成对本说明书的限定。虽然此处并没有明确说明，本领域技术人员可能会对本说明书进行各种修改、改进和修正。该类修改、改进和修正在本说明书中被建议，所以该类修改、改进、修正仍属于本说明书示范实施例的精神和范围。

[129]同时，本说明书使用了特定词语来描述本说明书的实施例。如“一个实施例”、“一实施例”、和/或“一些实施例”意指与本说明书至少一个实施例相关的某一特征、结构或特点。因此，应强调并注意的是，本说明书中在不同位置两次或多次提及的“一实施例”或“一个实施例”或“一个替代性实施例”并不一定是指同一实施例。此外，本说明书的一个或多个实施例中的某些特征、结构或特点可以进行适当的组合。

[130]此外，本领域技术人员可以理解，本说明书的各方面可以通过若干具有可专利性的种类或情况进行说明和描述，包括任何新的和有用的工序、机器、产品或物质的组合，或对他们的任何新的和有用的改进。相应地，本说明书的各个方面可以完全由硬件执行、可以完全由软件（包括固件、常驻软件、微码等）执行、也可以由硬件和软件组合执行。以上硬件或软件均可被称为“数据块”、“模块”、“引擎”、“单元”、“组件”或“系统”。此外，本说明书的各方面可能表现为位于一个或多个计算机可读介质中的计算机产品，该产品包括计算机可读程序编码。

[131]计算机存储介质可能包含一个内含有计算机程序编码的传播数据信号，例如在基带上或作为载波的一部分。该传播信号可能有多种表现形式，包括电磁形式、光形式等，或合适的组合形式。计算机存储介质可以是除计算机可读存储介质之外的任何计算机可读介质，该介质可以通过连接至一个指令执行系统、装置或设备以实现通讯、传播或传输供使用的程序。位于计算机存储介质上的程序编码可以通过任何合适的介质进行传播，包括无线电、电缆、光纤电缆、RF、或类似介质，或任何上述介质的组合。

[132]本说明书各部分操作所需的计算机程序编码可以用任意一种或多种程序语言编写，包括面向对象编程语言如 Java、Scala、Smalltalk、Eiffel、JADE、Emerald、C++、C#、VB.NET、Python 等，常规程序化编程语言如 C 语言、Visual Basic、Fortran2003、Perl、COBOL2002、PHP、ABAP，动态编程语言如 Python、Ruby 和 Groovy，或其他编程语言等。该程序编码可以完全在用户计算机上运行、或作为独立的软件包在用户计算机上运行、或部分在用户计算机上运行部分在远程计算机运行、或完全在远程计算机或处理

设备上运行。在后种情况下，远程计算机可以通过任何网络形式与用户计算机连接，比如局域网（LAN）或广域网（WAN），或连接至外部计算机（例如通过因特网），或在云计算环境中，或作为服务使用如软件即服务（SaaS）。

[133]此外，除非权利要求中明确说明，本说明书所述处理元素和序列的顺序、数字字母的使用、或其他名称的使用，并非用于限定本说明书流程和方法的顺序。尽管上述披露中通过各种示例讨论了一些目前认为有用的发明实施例，但应当理解的是，该类细节仅起到说明的目的，附加的权利要求并不仅限于披露的实施例，相反，权利要求旨在覆盖所有符合本说明书实施例实质和范围的修正和等价组合。例如，虽然以上所描述的系统组件可以通过硬件设备实现，但是也可以只通过软件的解决方案得以实现，如在现有的处理设备或移动设备上安装所描述的系统。

[134]应当注意的是，为了简化本说明书披露的表述，从而帮助对一个或多个发明实施例的理解，前文对本说明书实施例的描述中，有时会将多种特征归并至一个实施例、附图或对其的描述中。但是，这种披露方法并不意味着本说明书对象所需要的特征比权利要求中提及的特征多。实际上，实施例的特征要少于上述披露的单个实施例的全部特征。

[135]一些实施例中使用了描述成分、属性数量的数字，应当理解的是，此类用于实施例描述的数字，在一些示例中使用了修饰词“大约”、“近似”或“大体上”来修饰。除非另外说明，“大约”、“近似”或“大体上”表明所述数字允许有 $\pm 20\%$ 的变化。相应地，在一些实施例中，说明书和权利要求中使用的数值参数均为近似值，该近似值根据个别实施例所需特点可以发生改变。在一些实施例中，数值参数应考虑规定的有效数位并采用一般位数保留的方法。尽管本说明书一些实施例中用于确认其范围广度的数值域和参数为近似值，在具体实施例中，此类数值的设定在可行范围内尽可能精确。

[136]针对本说明书引用的每个专利、专利申请、专利申请公开物和其他材料，如文章、书籍、说明书、出版物、文档等，特此将其全部内容并入本说明书作为参考。与本说明书内容不一致或产生冲突的申请历史文件除外，对本说明书权利要求最广范围有限制的文件（当前或之后附加于本说明书中的）也除外。需要说明的是，如果本说明书附属材料中的描述、定义、和/或术语的使用与本说明书所述内容有不一致或冲突的地方，以本说明书的描述、定义和/或术语的使用为准。

[137]最后，应当理解的是，本说明书中所述实施例仅用以说明本说明书实施例的原则。其他的变形也可能属于本说明书的范围。因此，作为示例而非限制，本说明书实施例的替代配置可视为与本说明书的教导一致。相应地，本说明书的实施例不仅限于本说明书

明确介绍和描述的实施例。

权利要求书

1、一种事件检测方法，所述方法包括：

获取待处理文本；

基于抽取模型从待处理文本中抽取一组或多组实例数据；其中，每组实例数据包括第一实体实例、所述第一实体实例对应的第一实体类型、第二实体实例、所述第二实体实例对应的第二实体类型以及两个实体类型之间的关系描述；

基于所述一组或多组实例数据，确定一个或多个抽取三元组，进而得到抽取图谱；其中，抽取三元组包括实例数据中的第一实体类型、第二实体类型以及两个实体类型之间的关系描述；

获取一个或多个候选事件的图谱本体定义数据，并基于此得到对应每个候选事件的本体定义图谱；其中，事件的图谱本体定义数据包括用于定义实体的实体类型以及用于定义实体类型间关系的关系描述；

确定所述抽取图谱分别与所述一个或多个候选事件的本体定义图谱的相似度；

基于各相似度，从所述一个或多个候选事件中确定所述待处理文本对应的事件。

2、如权利要求 1 所述的方法，还包括：为事件定义其图谱本体定义数据；所述事件的图谱本体定义数据中的实体类型以及关系描述来自预设的本体定义数据集。

3、如权利要求 1 或 2 所述的方法，所述抽取模型使用训练样本训练得到，所述训练样本的标签基于预设的本体定义数据集中的实体类型以及关系描述确定。

4、如权利要求 1 所述的方法，所述关系描述包括以下关系中的一种或多种：动宾关系、主谓关系、定中关系和修饰关系。

5、如权利要求 1 所述的方法，所述基于抽取模型从待处理文本中抽取一组或多组实例数据，包括：

利用抽取模型处理待处理文本，得到待处理文本的标注序列以及关系矩阵；

基于标注序列确定待处理文本中的实体实例及其实体类型；

基于关系矩阵确定待处理文本中任意两个实体实例之间的关系描述，并将其作为对应两个实体类型之间的关系描述。

6、如权利要求 5 所述的方法，标注序列用于标记待处理文本中属于实体实例的字或词，以及所述字或词所属的实体类型；所述关系矩阵用于标记待处理文本中任意两个字或词之间的关系描述。

7、如权利要求 1 或 5 所述的方法，所述抽取模型包括以下模型中的一种或多：BERT、Transformer、StanfordNLP 或 LTP。

8、如权利要求 1 所述的方法，所述确定所述抽取图谱分别与所述一个或多个候选事件的本体定义图谱的相似度，包括对于任一候选事件的本体定义图谱：

利用图匹配模型处理所述抽取图谱以及所述候选事件的本体定义图谱，得到两者之间的相似度。

9、如权利要求 1 所述的方法，还包括：

基于所述一组或多组实例数据，确定一个或多个实例三元组；其中，实例三元组包括实例数据中的第一实体实例、第二实体实例及其分别对应的两个实体类型之间的关系描述；

基于所述一组或多组实例三元组，确定待处理文本对应事件的事件要素。

10、一种事件检测系统，所述系统包括：

文本获取模块，用于获取待处理文本；

抽取模块，用于基于抽取模型从待处理文本中抽取一组或多组实例数据；其中，每组实例数据包括第一实体实例、所述第一实体实例对应的第一实体类型、第二实体实例、所述第二实体实例对应的第二实体类型以及两个实体类型之间的关系描述；

抽取图谱获取模块，用于基于所述一组或多组实例数据，确定一个或多个抽取三元组，进而得到抽取图谱；其中，抽取三元组包括实例数据中的第一实体类型、第二实体类型以及两个实体类型之间的关系描述；

本体定义图谱获取模块，用于获取一个或多个候选事件的本体定义数据并基于此得到对应每个候选事件的本体定义图谱；其中，事件的本体定义数据包括用于定义实体的实体类型以及用于定义实体类型间关系的关系描述；

相似度确定模块，用于确定所述抽取图谱分别与所述一个或多个候选事件的本体定义图谱的相似度；

事件确定模块，用于基于各相似度，从所述一个或多个候选事件中确定所述待处理文本对应的事件。

11、一种计算机可读存储介质，其特征在于，所述存储介质存储计算机指令，当所述计算机指令被处理器执行时实现如权利要求 1~9 中任一项所述的方法。

12、一种事件检测装置，其特征在于，所述装置包括至少一个处理器以及至少一个存储器；所述至少一个存储器用于存储计算机指令；所述至少一个处理器用于执行所述计算机指令中的至少部分指令以实现如权利要求 1~9 中任一项所述的方法。