(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 113434695 A
(43)申请公布日 2021.09.24

(21)申请号 202110716036.X

(22)申请日 2021.06.25

(71)申请人 平安科技(深圳)有限公司
地址 518000 广东省深圳市福田区福田街道福安社区益田路5033号平安金融中心23楼

(72)发明人 王思瀚

(74)专利代理机构 广州三环专利商标代理有限公司 44202
代理人 熊永强

(51)Int.Cl.
G06F 16/36(2019.01)
G06F 40/289(2020.01)
G06K 9/62(2006.01)

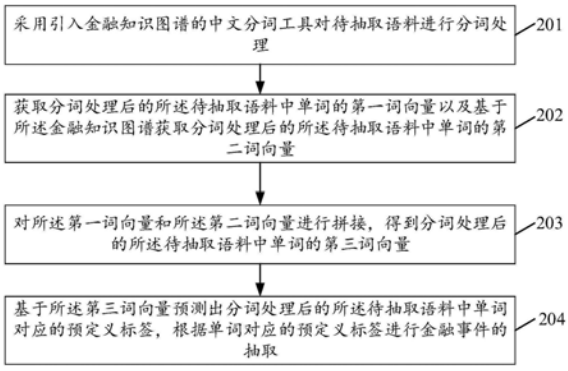
权利要求书2页 说明书13页 附图5页

(54)发明名称

金融事件抽取方法、装置、电子设备及存储介质

(57)摘要

本申请实施例提供了一种金融事件抽取方法、装置、电子设备及存储介质，其中，该金融事件抽取方法包括：采用引入金融知识图谱的中文分词工具对待抽取语料进行分词处理，获取分词处理后的待抽取语料中单词的第一词向量以及基于金融知识图谱获取分词处理后的待抽取语料中单词的第二词向量，对第一词向量和第二词向量进行拼接，得到分词处理后的待抽取语料中单词的第三词向量，基于第三词向量预测出分词处理后的待抽取语料中单词对应的预定义标签，根据单词对应的预定义标签进行金融事件的抽取。本申请实施例有利于提升金融事件抽取的精度。



CN 113434695 A

1.一种金融事件抽取方法,其特征在于,所述方法包括:

采用引入金融知识图谱的中文分词工具对待抽取语料进行分词处理;

获取分词处理后的所述待抽取语料中单词的第一词向量以及基于所述金融知识图谱获取分词处理后的所述待抽取语料中单词的第二词向量;

对所述第一词向量和所述第二词向量进行拼接,得到分词处理后的所述待抽取语料中单词的第三词向量;

基于所述第三词向量预测出分词处理后的所述待抽取语料中单词对应的预定义标签,根据单词对应的预定义标签进行金融事件的抽取。

2.根据权利要求1所述的方法,其特征在于,所述基于所述金融知识图谱获取分词处理后的所述待抽取语料中单词的第二词向量,包括:

以所述金融知识图谱中的目标节点、与所述目标节点直接相连的节点和连接该两个节点的边构建三元组,所述目标节点为所述金融知识图谱中代表分词处理后的所述待抽取语料中的实体的节点;

对所述三元组进行向量化处理,得到分词处理后的所述待抽取语料中的实体的第四词向量;

将分词处理后的所述待抽取语料中不属于所述金融知识图谱的节点的单词映射为零向量;

将所述第四词向量和所述零向量确定为所述第二词向量。

3.根据权利要求1或2所述的方法,其特征在于,在采用引入金融知识图谱的中文分词工具对待抽取语料进行分词处理之前,所述方法还包括:

将金融知识图谱的节点的名称添加到金融名词表中;

将所述金融名词表加入预设中文分词工具的词表中,得到引入金融知识图谱的中文分词工具。

4.根据权利要求1-3任一项所述的方法,其特征在于,所述基于所述第三词向量预测出分词处理后的所述待抽取语料中单词对应的预定义标签,包括:

利用预设的注意力机制计算所述第三词向量中每个词向量的权重;

根据每个词向量的权重计算得到分词处理后的所述待抽取语料中单词的注意力向量;

根据所述注意力向量预测出分词处理后的所述待抽取语料中单词对应的预定义标签。

5.根据权利要求4所述的方法,其特征在于,所述根据所述注意力向量预测出分词处理后的所述待抽取语料中单词对应的预定义标签,包括:

对所述注意力向量进行编码,得到待分类向量;

将所述待分类向量输入训练好的序列标注模型进行分类预测,得到分词处理后的所述待抽取语料中单词对应的预定义标签。

6.根据权利要求1-3任一项所述的方法,其特征在于,在将所述金融名词表加入预设中文分词工具的词表中,得到引入金融知识图谱的中文分词工具之后,所述方法还包括:

获取样本语料,采用引入金融知识图谱的中文分词工具对所述样本语料进行分词处理;

获取分词处理后的所述样本语料中单词的第五词向量以及基于所述金融知识图谱获取分词处理后的所述样本语料中单词的第六词向量;

对所述第五词向量和所述第六词向量进行拼接，得到分词处理后的所述样本语料中单词的第七词向量；

将所述七词向量输入序列标注模型进行训练，得到分词处理后的所述样本语料中单词对应的预定义标签；

根据分词处理后的所述样本语料中单词对应的标注数据和单词对应的预定义标签确定目标损失；

对所述序列标注模型的参数进行调整，以最小化所述目标损失，获得训练好的序列标注模型。

7.一种金融事件抽取装置，其特征在于，所述装置包括：

分词模块，用于采用引入金融知识图谱的中文分词工具对待抽取语料进行分词处理；

编码模块，用于获取分词处理后的所述待抽取语料中单词的第一词向量以及基于所述金融知识图谱获取分词处理后的所述待抽取语料中单词的第二词向量；

拼接模块，用于对所述第一词向量和所述第二词向量进行拼接，得到分词处理后的所述待抽取语料中单词的第三词向量；

处理模块，用于基于所述第三词向量预测出分词处理后的所述待抽取语料中单词对应的预定义标签，根据单词对应的预定义标签进行金融事件的抽取。

8.根据权利要求7所述的装置，其特征在于，在基于所述金融知识图谱获取分词处理后的所述待抽取语料中单词的第二词向量方面，所述编码模块具体用于：

以所述金融知识图谱中的目标节点、与所述目标节点直接相连的节点和连接该两个节点的边构建三元组，所述目标节点为所述金融知识图谱中代表分词处理后的所述待抽取语料中的实体的节点；

对所述三元组进行向量化处理，得到分词处理后的所述待抽取语料中的实体的第四词向量；

将分词处理后的所述待抽取语料中不属于所述金融知识图谱的节点的单词映射为零向量；

将所述第四词向量和所述零向量确定为所述第二词向量。

9.一种电子设备，包括输入设备和输出设备，其特征在于，还包括：

处理器，适于实现一条或多条指令；以及，

计算机存储介质，所述计算机存储介质存储有一条或多条指令，所述一条或多条指令适于由所述处理器加载并执行如权利要求1-6任一项所述的方法。

10.一种计算机存储介质，其特征在于，所述计算机存储介质存储有一条或多条指令，所述一条或多条指令适于由处理器加载并执行如权利要求1-6任一项所述的方法。

金融事件抽取方法、装置、电子设备及存储介质

技术领域

[0001] 本申请涉及事件抽取技术领域，尤其涉及一种金融事件抽取方法、装置、电子设备及存储介质。

背景技术

[0002] 事件抽取是信息抽取领域的一个重要研究方向，所谓事件抽取是指将非结构化文本中的事件信息以结构化的形式进行呈现。作为事件抽取下的一个分支，金融事件抽取指的是从金融相关的非结构化材料，例如研究报告、新闻等中，抽取出对于已发生的金融事件的简洁描述，抽取出的事件集合可服务于更上层的应用，比如投管分析。目前的金融事件抽取多是基于序列标注模型，通过定义标签体系为(B,beginning，表示事件主体描述的开始；I,intermediate，表示事件描述的主要内容；E,end，表示事件描述的结束；O,others，表示不相关的字符)，预测语料中每个位置的字符对应的标签完成事件抽取。现有方案虽然适用于绝大多数事件抽取场景，但在某些具体领域，比如金融，其抽取精度还有待提升。

发明内容

[0003] 针对上述问题，本申请提供了一种金融事件抽取方法、装置、电子设备及存储介质，有利于提升金融事件抽取的精度。

[0004] 为实现上述目的，本申请实施例第一方面提供了一种金融事件抽取方法，该方法包括：

[0005] 采用引入金融知识图谱的中文分词工具对待抽取语料进行分词处理；

[0006] 获取分词处理后的所述待抽取语料中单词的第一词向量以及基于所述金融知识图谱获取分词处理后的所述待抽取语料中单词的第二词向量；

[0007] 对所述第一词向量和所述第二词向量进行拼接，得到分词处理后的所述待抽取语料中单词的第三词向量；

[0008] 基于所述第三词向量预测出分词处理后的所述待抽取语料中单词对应的预定义标签，根据单词对应的预定义标签进行金融事件的抽取。

[0009] 结合第一方面，在一种可能的实施方式中，所述基于所述金融知识图谱获取分词处理后的所述待抽取语料中单词的第二词向量，包括：

[0010] 以所述金融知识图谱中的目标节点、与所述目标节点直接相连的节点和连接该两个节点的边构建三元组，所述目标节点为所述金融知识图谱中代表分词处理后的所述待抽取语料中的实体的节点；

[0011] 对所述三元组进行向量化处理，得到分词处理后的所述待抽取语料中的实体的第四词向量；

[0012] 将分词处理后的所述待抽取语料中不属于所述金融知识图谱的节点的单词映射为零向量；

[0013] 将所述第四词向量和所述零向量确定为所述第二词向量。

[0014] 结合第一方面，在一种可能的实施方式中，在采用引入金融知识图谱的中文分词工具对待抽取语料进行分词处理之前，所述方法还包括：

[0015] 将金融知识图谱的节点的名称添加到金融名词表中；

[0016] 将所述金融名词表加入预设中文分词工具的词表中，得到引入金融知识图谱的中文分词工具。

[0017] 结合第一方面，在一种可能的实施方式中，所述基于所述第三词向量预测出分词处理后的所述待抽取语料中单词对应的预定义标签，包括：

[0018] 利用预设的注意力机制计算所述第三词向量中每个词向量的权重；

[0019] 根据每个词向量的权重计算得到分词处理后的所述待抽取语料中单词的注意力向量；

[0020] 根据所述注意力向量预测出分词处理后的所述待抽取语料中单词对应的预定义标签。

[0021] 结合第一方面，在一种可能的实施方式中，所述根据所述注意力向量预测出分词处理后的所述待抽取语料中单词对应的预定义标签，包括：

[0022] 对所述注意力向量进行编码，得到待分类向量；

[0023] 将所述待分类向量输入训练好的序列标注模型进行分类预测，得到分词处理后的所述待抽取语料中单词对应的预定义标签。

[0024] 结合第一方面，在一种可能的实施方式中，在将所述金融名词表加入预设中文分词工具的词表中，得到引入金融知识图谱的中文分词工具之后，所述方法还包括：

[0025] 获取样本语料，采用引入金融知识图谱的中文分词工具对所述样本语料进行分词处理；

[0026] 获取分词处理后的所述样本语料中单词的第五词向量以及基于所述金融知识图谱获取分词处理后的所述样本语料中单词的第六词向量；

[0027] 对所述第五词向量和所述第六词向量进行拼接，得到分词处理后的所述样本语料中单词的第七词向量；

[0028] 将所述第七词向量输入序列标注模型进行训练，得到分词处理后的所述样本语料中单词对应的预定义标签；

[0029] 根据分词处理后的所述样本语料中单词对应的标注数据和单词对应的预定义标签确定目标损失；

[0030] 对所述序列标注模型的参数进行调整，以最小化所述目标损失，获得训练好的序列标注模型。

[0031] 本申请实施例第二方面提供了一种金融事件抽取装置，该装置包括：

[0032] 分词模块，用于采用引入金融知识图谱的中文分词工具对待抽取语料进行分词处理；

[0033] 编码模块，用于获取分词处理后的所述待抽取语料中单词的第一词向量以及基于所述金融知识图谱获取分词处理后的所述待抽取语料中单词的第二词向量；

[0034] 拼接模块，用于对所述第一词向量和所述第二词向量进行拼接，得到分词处理后的所述待抽取语料中单词的第三词向量；

[0035] 处理模块，用于基于所述第三词向量预测出分词处理后的所述待抽取语料中单词

对应的预定义标签，根据单词对应的预定义标签进行金融事件的抽取。

[0036] 本申请实施例第三方面提供了一种电子设备，该电子设备包括输入设备和输出设备，还包括处理器，适于实现一条或多条指令；以及，计算机存储介质，所述计算机存储介质存储有一条或多条指令，所述一条或多条指令适于由所述处理器加载并执行如下步骤：

[0037] 采用引入金融知识图谱的中文分词工具对待抽取语料进行分词处理；

[0038] 获取分词处理后的所述待抽取语料中单词的第一词向量以及基于所述金融知识图谱获取分词处理后的所述待抽取语料中单词的第二词向量；

[0039] 对所述第一词向量和所述第二词向量进行拼接，得到分词处理后的所述待抽取语料中单词的第三词向量；

[0040] 基于所述第三词向量预测出分词处理后的所述待抽取语料中单词对应的预定义标签，根据单词对应的预定义标签进行金融事件的抽取。

[0041] 本申请实施例第四方面提供了一种计算机存储介质，所述计算机存储介质存储有一条或多条指令，所述一条或多条指令适于由处理器加载并执行如下步骤：

[0042] 采用引入金融知识图谱的中文分词工具对待抽取语料进行分词处理；

[0043] 获取分词处理后的所述待抽取语料中单词的第一词向量以及基于所述金融知识图谱获取分词处理后的所述待抽取语料中单词的第二词向量；

[0044] 对所述第一词向量和所述第二词向量进行拼接，得到分词处理后的所述待抽取语料中单词的第三词向量；

[0045] 基于所述第三词向量预测出分词处理后的所述待抽取语料中单词对应的预定义标签，根据单词对应的预定义标签进行金融事件的抽取。

[0046] 本申请的上述方案至少包括以下有益效果：与现有技术相比，本申请实施例通过采用引入金融知识图谱的中文分词工具对待抽取语料进行分词处理，获取分词处理后的待抽取语料中单词的第一词向量以及基于金融知识图谱获取分词处理后的待抽取语料中单词的第二词向量，对第一词向量和第二词向量进行拼接，得到分词处理后的待抽取语料中单词的第三词向量，基于第三词向量预测出分词处理后的待抽取语料中单词对应的预定义标签，根据单词对应的预定义标签进行金融事件的抽取。由于本申请在金融事件抽取时引入了金融知识图谱的信息，引入金融知识图谱的中文分词工具在对待抽取语料进行分词时有利于确保重要专业词汇的完整性，并确保提取出的第一词向量和第二词向量的质量更高，那么，采用更高质量的第一词向量和第二词向量拼接成的第三词向量进行预定义标签的预测，有利于提高标签预测的精度，从而提升金融事件抽取的精度。另外，预定义的标签对原有标签体系进行了细化，其只关注事件主体和事件谓语，在保留事件语义的前提下，去除了大量修饰性的无用信息，极大地提升了金融事件描述的精简性，有利于降低上层应用对抽取出的金融事件做进一步处理的成本。

附图说明

[0047] 为了更清楚地说明本申请实施例或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本申请的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

[0048] 图1为本申请实施例提供的一种应用环境的示意图；

[0049] 图2为本申请实施例提供的一种金融事件抽取方法的流程示意图；

[0050] 图3为本申请实施例提供的一种金融事件抽取的模型结构示意图；

[0051] 图4为本申请实施例提供的另一种金融事件抽取方法的流程示意图；

[0052] 图5为本申请实施例提供的一种金融事件抽取装置的结构示意图；

[0053] 图6为本申请实施例提供的另一种金融事件抽取装置的结构示意图；

[0054] 图7为本申请实施例提供的一种电子设备的结构示意图。

具体实施方式

[0055] 为了使本技术领域的人员更好地理解本申请方案，下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本申请一部分的实施例，而不是全部的实施例。基于本申请中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都应当属于本申请保护的范围。

[0056] 本申请说明书、权利要求书和附图中出现的术语“包括”和“具有”以及它们任何变形，意图在于覆盖不排他的包含。例如包含了一系列步骤或单元的过程、方法、系统、产品或设备没有限定于已列出的步骤或单元，而是可选地还包括没有列出的步骤或单元，或可选地还包括对于这些过程、方法、产品或设备固有的其它步骤或单元。此外，术语“第一”、“第二”和“第三”等是用于区别不同的对象，而并非用于描述特定的顺序。

[0057] 在本申请中提及“实施例”意味着，结合实施例描述的特定特征、结构或特性可以包含在本申请的至少一个实施例中。在说明书中的各个位置出现该短语并不一定均是指相同的实施例，也不是与其它实施例互斥的独立的或备选的实施例。本领域技术人员显式地和隐式地理解的是，本申请所描述的实施例可以与其它实施例相结合。

[0058] 本申请实施例提供一种金融事件抽取方法，可基于图1所示的应用环境实施，请参见图1，该应用环境中包括终端设备和服务器，终端设备和服务器通信连接，其连接的方式可以是串口连接、无线网络连接、蓝牙连接、网络直连。其中，终端设备可以包括各种具有输入能力和通信能力的设备，其可以是平板电脑、掌上电脑、笔记本电脑等，服务器可以包括各种具有程序代码运行能力和通信能力的设备，其可以是独立的物理服务器，也可以是服务器集群或者分布式系统，还可以是提供云服务、云数据库、云计算、云函数、云存储、网络服务、云通信、中间件服务、域名服务、安全服务、以及大数据和人工智能平台等基础云计算服务的云服务器。

[0059] 具体实施中，终端设备通过通信部件接收用户输入的程序指令和待抽取语料，并由通信接口将该程序指令和待抽取语料发送给服务器，由服务器执行金融事件的抽取操作。服务器在接收到该程序指令和待抽取语料的情况下，基于提供的程序运行环境和基础运行该程序指令，以对待抽取语料进行分词、对分词后的待抽取语料进行向量化处理，采用序列标注模型对最终得到的向量进行预测，得到分词后的待抽取语料中单词对应的预定义标签，由此抽取出对应的金融事件。由于整个抽取过程中引入了金融知识图谱的相关信息，有利于辅助金融事件的抽取，并提升金融事件抽取的精度。

[0060] 基于图1所示的应用环境，以下结合其他附图对本申请实施例提供的金融事件抽

取方法进行详细阐述。

[0061] 请参见图2,图2为本申请实施例提供的一种金融事件抽取方法的流程示意图,该方法应用于电子设备,如图2所示,包括步骤201-204:

[0062] 201:采用引入金融知识图谱的中文分词工具对待抽取语料进行分词处理。

[0063] 本申请具体实施例中,在采用引入金融知识图谱的中文分词工具对待抽取语料进行分词处理之前,所述方法进一步包括:

[0064] 将金融知识图谱的节点的名称添加到金融名词表中;

[0065] 将所述金融名词表加入预设中文分词工具的词表中,得到引入金融知识图谱的中文分词工具。

[0066] 示例性的,该预设中文分词工具可以是Jieba分词工具,金融知识图谱是知识图谱在金融领域的垂直应用,其以图谱节点来表征金融领域中的重要实体,比如产业、标地、国家、机构、重要从业人员等等,图谱中的边代表各类实体间的关系属性。简而言之,金融知识图谱是一个大而全的金融信息集合,在现有中文分词工具中加入金融名词表,有利于准确识别到待抽取语料中的金融领域重要实体,在分词时有利于确保金融领域专有名词的完整性,以提升后续词向量的质量。

[0067] 202:获取分词处理后的所述待抽取语料中单词的第一词向量以及基于所述金融知识图谱获取分词处理后的所述待抽取语料中单词的第二词向量。

[0068] 本申请具体实施例中,第一词向量是指经过分词处理后的待抽取语料中每个单词的向量化表示,具体可采用词嵌入的方法得到,比如词嵌入模型,该词嵌入模型可以是BERT(Bidirectional Encoder Representations from Transformers,基于转换器的双向编码表征)模型、Word2Vec(Word to Vector,一个将单词转换成向量形式的工具)模型,等等。

[0069] 示例性的,上述基于所述金融知识图谱获取分词处理后的所述待抽取语料中单词的第二词向量,包括:

[0070] 以所述金融知识图谱中的目标节点、与所述目标节点直接相连的节点和连接该两个节点的边构建三元组;

[0071] 对所述三元组进行向量化处理,得到分词处理后的所述待抽取语料中的实体的第四词向量;

[0072] 将分词处理后的所述待抽取语料中不属于所述金融知识图谱的节点的单词映射为零向量;

[0073] 将所述第四词向量和所述零向量确定为所述第二词向量。

[0074] 举例说明,假设待抽取语料为“张XX在国会证词中释放强烈降息信号”,“张XX”为分词处理后的待抽取语料中的实体,若金融知识图谱中还存在与之直接相连的节点“经理”,连接这两个节点之间的边表示两个实体之间的关系,比如“A公司”,则由“张XX、A公司、经理”可构建一个三元组,其中,“张XX”表示头实体,“A公司”表示关系,“经理”表示尾实体,如图3所示,采用TransE算法可将该三元组进行向量化处理,得到实体和关系的向量化表示,由此得到分词处理后的待抽取语料中的实体的第四词向量,比如“张XX”的词向量,而对于分词处理后的待抽取语料中未在金融知识图谱中出现的单词,将其词向量置为零向量,所有的第四词向量和零向量都为第二词向量。该实施方式中,借助金融知识图谱中的知识,在待抽取语料中缺少金融概念或实体别名时,比如“张XX”是A公司经理这一隐藏信息,也能

抽取出相应的事件，从而能有效提升事件抽取模型的召回率，进而提升事件抽取的精度。

[0075] 203：对所述第一词向量和所述第二词向量进行拼接，得到分词处理后的所述待抽取语料中单词的第三词向量。

[0076] 本申请具体实施例中，请继续参见图3，记第一词向量为v_embed，记第二词向量为v_graph，则分词处理后的待抽取语料中每个单词的第三词向量表示为v_t=[v_embed,v_graph]，将该第三词向量v_t作为序列标注模型的输入，或者将基于该第三词向量v_t得到的新的向量作为序列标注模型的输入。

[0077] 204：基于所述第三词向量预测出分词处理后的所述待抽取语料中单词对应的预定义标签，根据单词对应的预定义标签进行金融事件的抽取。

[0078] 本申请具体实施例中，由于金融事件的内容可大致分为两类：一类为对政府或专业机构政策性的描述，另一类是对具体产业或大类资产价值走势变化的描述。用户真正感兴趣信息往往只涉及明确的事件主体和事件谓语，因此，想要精确地进行事件抽取，就需要抽取出这些信息，并排除掉文本中修饰类的内容。

[0079] 本申请在现有常规序列标注的标签体系下，作了进一步的细化，预定义标签体系为：

[0080] SB:subject beginning，表示事件主体描述的开始。

[0081] SI:subject intermediate，表示事件主体描述的主要内容。

[0082] SE:subject end，表示事件主体描述的结束。

[0083] PB:predicate beginning，表示事件谓语描述的开始。

[0084] PI:predicate intermediate，表示事件谓语的主要内容。

[0085] PE:predicate end，表示事件谓语描述的结束。

[0086] O:others，表示不相关的字符。

[0087] 在预定义了标签体系的基础上，采用序列标注模型对第三词向量进行分类预测，序列标注模型可以是LSTM(Long Short-Term Memory，长短期记忆网络)模型，也可以是Bi-LSTM(Bidirectional LSTM Networks，双向长短期记忆网络)模型，或者还可以是BI-LSTM-CRF(Bidirectional LSTM Networks-Conditional random field，双向长短期记忆网络-条件随机场)模型，等等。相较于序列标注模型的常规输入，本申请将第二词向量v_graph拼接在第一词向量v_embed后，输入向量的维度有所增加，但序列标注模型的处理方式与原来保持一致，其最后一层输出的标签则为预定义的标签，如图3中的：Y1,Y2,…Yn-1,Yn，其中，n表示单词的数量，具体可如下所示：

[0088] 张XX在国会证词中释放强烈降息信号

[0089] SB SI SE O O O O O O PB PI PI PI PI PI PI PE

[0090] 由于本申请仅抽取事件主体和事件谓语，则上述例子中，根据预测出的预定义标签，最终抽取出的金融事件为“张XX释放强烈降息信号”。

[0091] 示例性的，在将所述金融名词表加入预设中文分词工具的词表中，得到引入金融知识图谱的中文分词工具之后，所述方法还包括：

[0092] 获取样本语料，采用引入金融知识图谱的中文分词工具对所述样本语料进行分词处理；

[0093] 获取分词处理后的所述样本语料中单词的第五词向量以及基于所述金融知识图

谱获取分词处理后的所述样本语料中单词的第六词向量；

[0094] 对所述第五词向量和所述第六词向量进行拼接，得到分词处理后的所述样本语料中单词的第七词向量；

[0095] 将所述第七词向量输入序列标注模型进行训练，得到分词处理后的所述样本语料中单词对应的预定义标签；

[0096] 根据分词处理后的所述样本语料中单词对应的标注数据和单词对应的预定义标签确定目标损失；

[0097] 对所述序列标注模型的参数进行调整，以最小化所述目标损失，获得训练好的序列标注模型。

[0098] 本申请具体实施例中，第五词向量是指对分词处理后的样本语料中的单词进行向量化处理得到的词向量，第六词向量是指基于金融知识图谱采用 TransE算法对三元组进行向量化处理得到的词向量，训练阶段的处理可参照上述步骤201-204中的描述，此处不再赘述。目标损失可以是最大似然损失，其公式表示如下：

[0099] $$\mathrm{L}(\theta)=\sum_{k=1}^{N}\sum_{j=1}^{T}\log P_{\theta}\left(y_j|x_{1:j}\right);$$

[0100] 其中，L(θ)表示目标损失的值，N表示样本语料的数量，k表示N条样本语料中的第k条，T表示样本语料的长度，P_{θ}表示序列标注模型输出的预定义标签的概率分布，y_j表示位置j上的单词对应的预定义标签，$x_{1:j}$表示位置j以及位置j之前所有位置上的单词的拼接词向量，θ表示序列标注模型的参数。

[0101] 该实施方式中，金融知识图谱的引入，使得序列标注模型的训练不再需要大量的标注数据，打破了完全由大数据驱动模型训练的瓶颈，实现了数据和知识双驱动的训练模式，有利于降低训练开销。

[0102] 示例性的，上述基于所述第三词向量预测出分词处理后的所述待抽取语料中单词对应的预定义标签，包括：

[0103] 利用预设的注意力机制计算所述第三词向量中每个词向量的权重；

[0104] 根据每个词向量的权重计算得到分词处理后的所述待抽取语料中单词的注意力向量；

[0105] 根据所述注意力向量预测出分词处理后的所述待抽取语料中单词对应的预定义标签。

[0106] 本申请具体实施例中，在得到第三词向量后，采用预设的注意力机制捕捉待抽取语料中的主题信息，以通过学习单词的重要性，为待抽取语料中的单词分配不同的权重，相较于传统的注意力机制依赖于seq2seq (Sequence-to-sequence，一个自然语言处理中的编码-解码模型)解码器中的隐向量，预设的注意力机制为经过改进的注意力机制，其改进点在于通过主题标记器s_t对第三词向量进行标记，得到每个词向量的权重$\langle\alpha_t^i\rangle_{i=1}^{n}$，其中，t表示第t时刻。

[0107] 其中，$\alpha_t^i=softmax\left(o_t^i\right)$；

[0108] $$o_t^i=w_a^T\tanh\left(W^{(act)}(s_t\oplus v_t)+b^{(act)}\right);$$

[0109] 其中，w_a^T、$W^{(act)}$、$b^{(act)}$都是在以损失函数最小化的训练阶段学习到的参数，s_t表

示主题标记器s_t标记出的隐向量，⊕是用于连接两个向量的运算符号，也就是说预设的注意力机制依赖的是主题标记器s_t标记出的隐向量。将α_t^i与v_t^i相乘得到第i个第三词向量对应的注意力向量，利用该注意力向量进行金融事件的抽取有利于使序列标注模型更关注于待抽取语料的主题信息，进一步抽取到有用信息。

[0110] 示例性的，上述根据所述注意力向量预测出分词处理后的所述待抽取语料中单词对应的预定义标签，包括：

[0111] 对所述注意力向量进行编码，得到待分类向量；

[0112] 将所述待分类向量输入训练好的序列标注模型进行分类预测，得到分词处理后的所述待抽取语料中单词对应的预定义标签。

[0113] 本申请具体实施例中，对于预设的注意力机制输出的注意力向量，再次采用主题标记器s_t对其进行编码，得到待分类向量s'_t，将待分类向量s'_t输入训练好的序列标注模型进行分类预测，最终输出预定义标签，以使待抽取语料的主题信息更容易被序列标注模型捕捉到。

[0114] 可以看出，本申请实施例通过采用引入金融知识图谱的中文分词工具对待抽取语料进行分词处理，获取分词处理后的待抽取语料中单词的第一词向量以及基于金融知识图谱获取分词处理后的待抽取语料中单词的第二词向量，对第一词向量和第二词向量进行拼接，得到分词处理后的待抽取语料中单词的第三词向量，基于第三词向量预测出分词处理后的待抽取语料中单词对应的预定义标签，根据单词对应的预定义标签进行金融事件的抽取。由于本申请在金融事件抽取时引入了金融知识图谱的信息，引入金融知识图谱的中文分词工具在对待抽取语料进行分词时有利于确保重要专业词汇的完整性，并确保提取出的第一词向量和第二词向量的质量更高，那么，采用更高质量的第一词向量和第二词向量拼接成的第三词向量进行预定义标签的预测，有利于提高标签预测的精度，从而提升金融事件抽取的精度。另外，预定义的标签对原有标签体系进行了细化，其只关注事件主体和事件谓语，在保留事件语义的前提下，去除了大量修饰性的无用信息，极大地提升了金融事件描述的精简性，有利于降低上层应用对抽取出的金融事件做进一步处理的成本。

[0115] 请参见图4，图4本申请实施例提供的另一种金融事件抽取方法的流程示意图，同样可基于图1所示的应用环境实施，如图4所示，包括步骤401-408：

[0116] 401：采用引入金融知识图谱的中文分词工具对待抽取语料进行分词处理；

[0117] 402：获取分词处理后的所述待抽取语料中单词的第一词向量；

[0118] 403：以所述金融知识图谱中的目标节点、与所述目标节点直接相连的节点和连接该两个节点的边构建三元组，所述目标节点为所述金融知识图谱中代表分词处理后的所述待抽取语料中的实体的节点；

[0119] 404：对所述三元组进行向量化处理，得到分词处理后的所述待抽取语料中的实体的第四词向量；

[0120] 405：将分词处理后的所述待抽取语料中不属于所述金融知识图谱的节点的单词映射为零向量；

[0121] 406：将所述第四词向量和所述零向量确定为第二词向量；

[0122] 407：对所述第一词向量和所述第二词向量进行拼接，得到分词处理后的所述待抽

取语料中单词的第三词向量；

[0123] 408：基于所述第三词向量预测出分词处理后的所述待抽取语料中单词对应的预定义标签，根据单词对应的预定义标签进行金融事件的抽取。

[0124] 其中，步骤401-408的具体实施方式在图2所示的实施例中已有相关说明，且能达到相同或相似的有益效果，为避免重复，此处不再赘述。

[0125] 基于上述金融事件抽取方法实施例的描述，请参见图5，图5为本申请实施例提供的一种金融事件抽取装置的结构示意图，如图5所示，该装置包括：

[0126] 分词模块501，用于采用引入金融知识图谱的中文分词工具对待抽取语料进行分词处理；

[0127] 编码模块502，用于获取分词处理后的所述待抽取语料中单词的第一词向量以及基于所述金融知识图谱获取分词处理后的所述待抽取语料中单词的第二词向量；

[0128] 拼接模块503，用于对所述第一词向量和所述第二词向量进行拼接，得到分词处理后的所述待抽取语料中单词的第三词向量；

[0129] 处理模块504，用于基于所述第三词向量预测出分词处理后的所述待抽取语料中单词对应的预定义标签，根据单词对应的预定义标签进行金融事件的抽取。

[0130] 在一种可能的实施方式中，在基于所述金融知识图谱获取分词处理后的所述待抽取语料中单词的第二词向量方面，编码模块502具体用于：

[0131] 以所述金融知识图谱中的目标节点、与所述目标节点直接相连的节点和连接该两个节点的边构建三元组，所述目标节点为所述金融知识图谱中代表分词处理后的所述待抽取语料中的实体的节点；

[0132] 对所述三元组进行向量化处理，得到分词处理后的所述待抽取语料中的实体的第四词向量；

[0133] 将分词处理后的所述待抽取语料中不属于所述金融知识图谱的节点的单词映射为零向量；

[0134] 将所述第四词向量和所述零向量确定为所述第二词向量。

[0135] 在一种可能的实施方式中，分词模块501还用于：

[0136] 将金融知识图谱的节点的名称添加到金融名词表中；

[0137] 将所述金融名词表加入预设中文分词工具的词表中，得到引入金融知识图谱的中文分词工具。

[0138] 在一种可能的实施方式中，在基于所述第三词向量预测出分词处理后的所述待抽取语料中单词对应的预定义标签方面，处理模块504具体用于：

[0139] 利用预设的注意力机制计算所述第三词向量中每个词向量的权重；

[0140] 根据每个词向量的权重计算得到分词处理后的所述待抽取语料中单词的注意力向量；

[0141] 根据所述注意力向量预测出分词处理后的所述待抽取语料中单词对应的预定义标签。

[0142] 在一种可能的实施方式中，在根据所述注意力向量预测出分词处理后的所述待抽取语料中单词对应的预定义标签方面，处理模块504具体用于：

[0143] 对所述注意力向量进行编码，得到待分类向量；

[0144] 将所述待分类向量输入训练好的序列标注模型进行分类预测，得到分词处理后的所述待抽取语料中单词对应的预定义标签。

[0145] 在一种可能的实施方式中，如图6所示，该装置还包括训练模块505，该训练模块505用于：

[0146] 获取样本语料，采用引入金融知识图谱的中文分词工具对所述样本语料进行分词处理；

[0147] 获取分词处理后的所述样本语料中单词的第五词向量以及基于所述金融知识图谱获取分词处理后的所述样本语料中单词的第六词向量；

[0148] 对所述第五词向量和所述第六词向量进行拼接，得到分词处理后的所述样本语料中单词的第七词向量；

[0149] 将所述第七词向量输入序列标注模型进行训练，得到分词处理后的所述样本语料中单词对应的预定义标签；

[0150] 根据分词处理后的所述样本语料中单词对应的标注数据和单词对应的预定义标签确定目标损失；

[0151] 对所述序列标注模型的参数进行调整，以最小化所述目标损失，获得训练好的序列标注模型。

[0152] 根据本申请的一个实施例，图5或图6所示的金融事件抽取装置的各个单元可以分别或全部合并为一个或若干个另外的单元来构成，或者其中的某个 (些)单元还可以再拆分为功能上更小的多个单元来构成，这可以实现同样的操作，而不影响本申请的实施例的技术效果的实现。上述单元是基于逻辑功能划分的，在实际应用中，一个单元的功能也可以由多个单元来实现，或者多个单元的功能由一个单元实现。在本申请的其它实施例中，基于金融事件抽取装置也可以包括其它单元，在实际应用中，这些功能也可以由其它单元协助实现，并且可以由多个单元协作实现。

[0153] 根据本申请的另一个实施例，可以通过在包括中央处理单元(CPU)、随机存取存储介质(RAM)、只读存储介质(ROM)等处理元件和存储元件的例如计算机的通用计算设备上运行能够执行如图2或图4中所示的相应方法所涉及的各步骤的计算机程序(包括程序代码)，来构造如图5或图6中所示的金融事件抽取装置设备，以及来实现本申请实施例的金融事件抽取方法。所述计算机程序可以记载于例如计算机可读记录介质上，并通过计算机可读记录介质装载于上述计算设备中，并在其中运行。

[0154] 基于上述方法实施例和装置实施例的描述，本申请实施例还提供一种电子设备。请参见图7，该电子设备至少包括处理器701、输入设备702、输出设备703以及计算机存储介质704。其中，电子设备内的处理器701、输入设备702、输出设备703以及计算机存储介质704可通过总线或其他方式连接。

[0155] 计算机存储介质704可以存储在电子设备的存储器中，所述计算机存储介质704用于存储计算机程序，所述计算机程序包括程序指令，所述处理器701 用于执行所述计算机存储介质704存储的程序指令。处理器701(或称CPU (Central Processing Unit，中央处理器))是电子设备的计算核心以及控制核心，其适于实现一条或多条指令，具体适于加载并执行一条或多条指令从而实现相应方法流程或相应功能。

[0156] 在一个实施例中，本申请实施例提供的电子设备的处理器701可以用于进行一系

列金融事件的抽取处理：

[0157] 采用引入金融知识图谱的中文分词工具对待抽取语料进行分词处理；

[0158] 获取分词处理后的所述待抽取语料中单词的第一词向量以及基于所述金融知识图谱获取分词处理后的所述待抽取语料中单词的第二词向量；

[0159] 对所述第一词向量和所述第二词向量进行拼接，得到分词处理后的所述待抽取语料中单词的第三词向量；

[0160] 基于所述第三词向量预测出分词处理后的所述待抽取语料中单词对应的预定义标签，根据单词对应的预定义标签进行金融事件的抽取。

[0161] 再一个实施例中，处理器701执行所述基于所述金融知识图谱获取分词处理后的所述待抽取语料中单词的第二词向量，包括：

[0162] 以所述金融知识图谱中的目标节点、与所述目标节点直接相连的节点和连接该两个节点的边构建三元组，所述目标节点为所述金融知识图谱中代表分词处理后的所述待抽取语料中的实体的节点；

[0163] 对所述三元组进行向量化处理，得到分词处理后的所述待抽取语料中的实体的第四词向量；

[0164] 将分词处理后的所述待抽取语料中不属于所述金融知识图谱的节点的单词映射为零向量；

[0165] 将所述第四词向量和所述零向量确定为所述第二词向量。

[0166] 再一个实施例中，在采用引入金融知识图谱的中文分词工具对待抽取语料进行分词处理之前，处理器701还用于：

[0167] 将金融知识图谱的节点的名称添加到金融名词表中；

[0168] 将所述金融名词表加入预设中文分词工具的词表中，得到引入金融知识图谱的中文分词工具。

[0169] 再一个实施例中，处理器701执行所述基于所述第三词向量预测出分词处理后的所述待抽取语料中单词对应的预定义标签，包括：

[0170] 利用预设的注意力机制计算所述第三词向量中每个词向量的权重；

[0171] 根据每个词向量的权重计算得到分词处理后的所述待抽取语料中单词的注意力向量；

[0172] 根据所述注意力向量预测出分词处理后的所述待抽取语料中单词对应的预定义标签。

[0173] 再一个实施例中，处理器701执行所述根据所述注意力向量预测出分词处理后的所述待抽取语料中单词对应的预定义标签，包括：

[0174] 对所述注意力向量进行编码，得到待分类向量；

[0175] 将所述待分类向量输入训练好的序列标注模型进行分类预测，得到分词处理后的所述待抽取语料中单词对应的预定义标签。

[0176] 再一个实施例中，在将所述金融名词表加入预设中文分词工具的词表中，得到引入金融知识图谱的中文分词工具之后，处理器701还用于：

[0177] 获取样本语料，采用引入金融知识图谱的中文分词工具对所述样本语料进行分词处理；

[0178] 获取分词处理后的所述样本语料中单词的第五词向量以及基于所述金融知识图谱获取分词处理后的所述样本语料中单词的第六词向量；

[0179] 对所述第五词向量和所述第六词向量进行拼接，得到分词处理后的所述样本语料中单词的第七词向量；

[0180] 将所述第七词向量输入序列标注模型进行训练，得到分词处理后的所述样本语料中单词对应的预定义标签；

[0181] 根据分词处理后的所述样本语料中单词对应的标注数据和单词对应的预定义标签确定目标损失；

[0182] 对所述序列标注模型的参数进行调整，以最小化所述目标损失，获得训练好的序列标注模型。

[0183] 示例性的，电子设备包括但不仅限于处理器701、输入设备702、输出设备703以及计算机存储介质704。还可以包括内存、电源、应用客户端模块等。输入设备702可以是键盘、触摸屏、射频接收器等，输出设备703可以是扬声器、显示器、射频发送器等。本领域技术人员可以理解，所述示意图仅仅是电子设备的示例，并不构成对电子设备的限定，可以包括比图示更多或更少的部件，或者组合某些部件，或者不同的部件。

[0184] 需要说明的是，由于电子设备的处理器701执行计算机程序时实现上述的金融事件抽取方法中的步骤，因此上述金融事件抽取方法的实施例均适用于该电子设备，且均能达到相同或相似的有益效果。

[0185] 本申请实施例还提供了一种计算机存储介质(Memory)，所述计算机存储介质是电子设备中的记忆设备，用于存放程序和数据。可以理解的是，此处的计算机存储介质既可以包括终端中的内置存储介质，当然也可以包括终端所支持的扩展存储介质。计算机存储介质提供存储空间，该存储空间存储了终端的操作系统。并且，在该存储空间中还存放了适于被处理器701加载并执行的一条或多条的指令，这些指令可以是一个或一个以上的计算机程序(包括程序代码)。需要说明的是，此处的计算机存储介质可以是高速RAM存储器，也可以是非不稳定的存储器(non-volatile memory)，例如至少一个磁盘存储器；可选的，还可以是至少一个位于远离前述处理器701的计算机存储介质。在一个实施例中，可由处理器701加载并执行计算机存储介质中存放的一条或多条指令，以实现上述有关金融事件抽取方法的相应步骤。

[0186] 示例性的，计算机存储介质的计算机程序包括计算机程序代码，所述计算机程序代码可以为源代码形式、对象代码形式、可执行文件或某些中间形式等。所述计算机可读介质可以包括：能够携带所述计算机程序代码的任何实体或装置、记录介质、U盘、移动硬盘、磁碟、光盘、计算机存储器、只读存储器(ROM，Read-Only Memory)、随机存取存储器(RAM，Random Access Memory)、电载波信号、电信信号以及软件分发介质等。

[0187] 需要说明的是，由于计算机存储介质的计算机程序被处理器执行时实现上述的金融事件抽取方法中的步骤，因此上述金融事件抽取方法的所有实施例均适用于该计算机存储介质，且均能达到相同或相似的有益效果。

[0188] 以上对本申请实施例进行了详细介绍，本文中应用了具体个例对本申请的原理及实施方式进行了阐述，以上实施例的说明只是用于帮助理解本申请的方法及其核心思想；同时，对于本领域的一般技术人员，依据本申请的思想，在具体实施方式及应用范围上均会

有改变之处,综上所述,本说明书内容不应理解为对本申请的限制。


通信
终端设备
服务器

图1

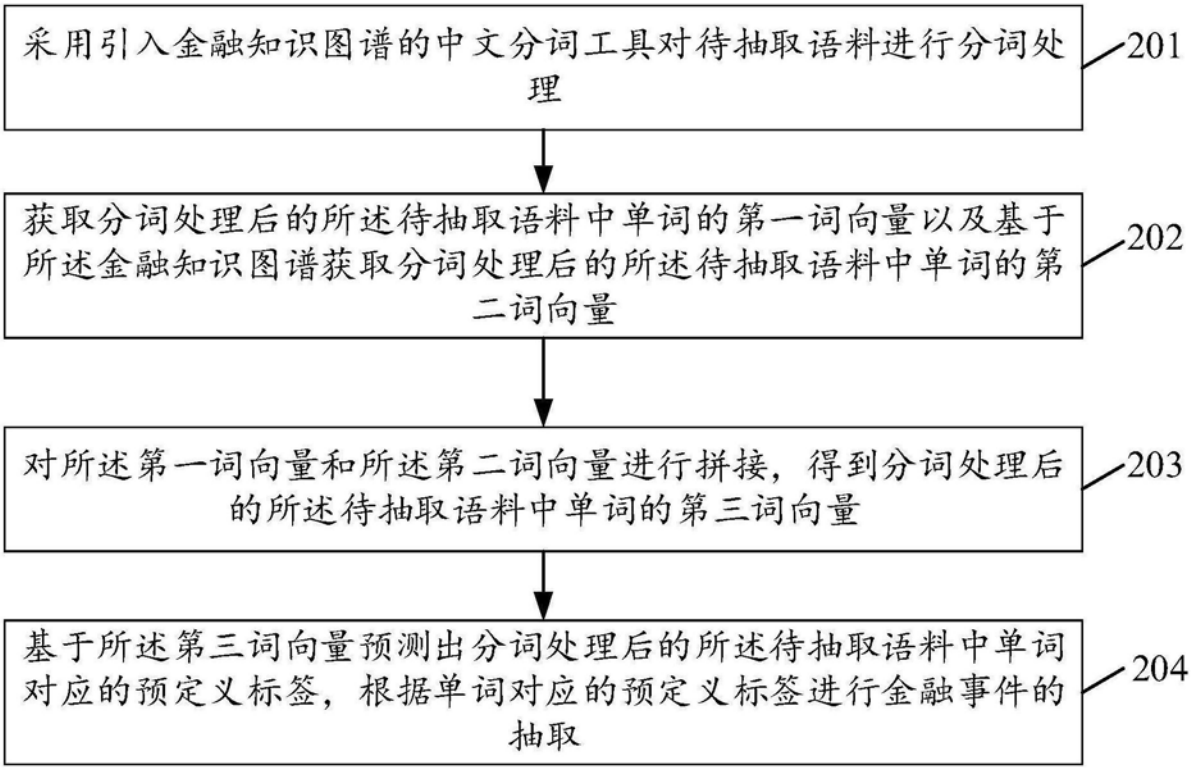

采用引入金融知识图谱的中文分词工具对待抽取语料进行分词处理
201
获取分词处理后的所述待抽取语料中单词的第一词向量以及基于所述金融知识图谱获取分词处理后的所述待抽取语料中单词的第二词向量
202
对所述第一词向量和所述第二词向量进行拼接，得到分词处理后的所述待抽取语料中单词的第三词向量
203
基于所述第三词向量预测出分词处理后的所述待抽取语料中单词对应的预定义标签，根据单词对应的预定义标签进行金融事件的抽取
204

图2

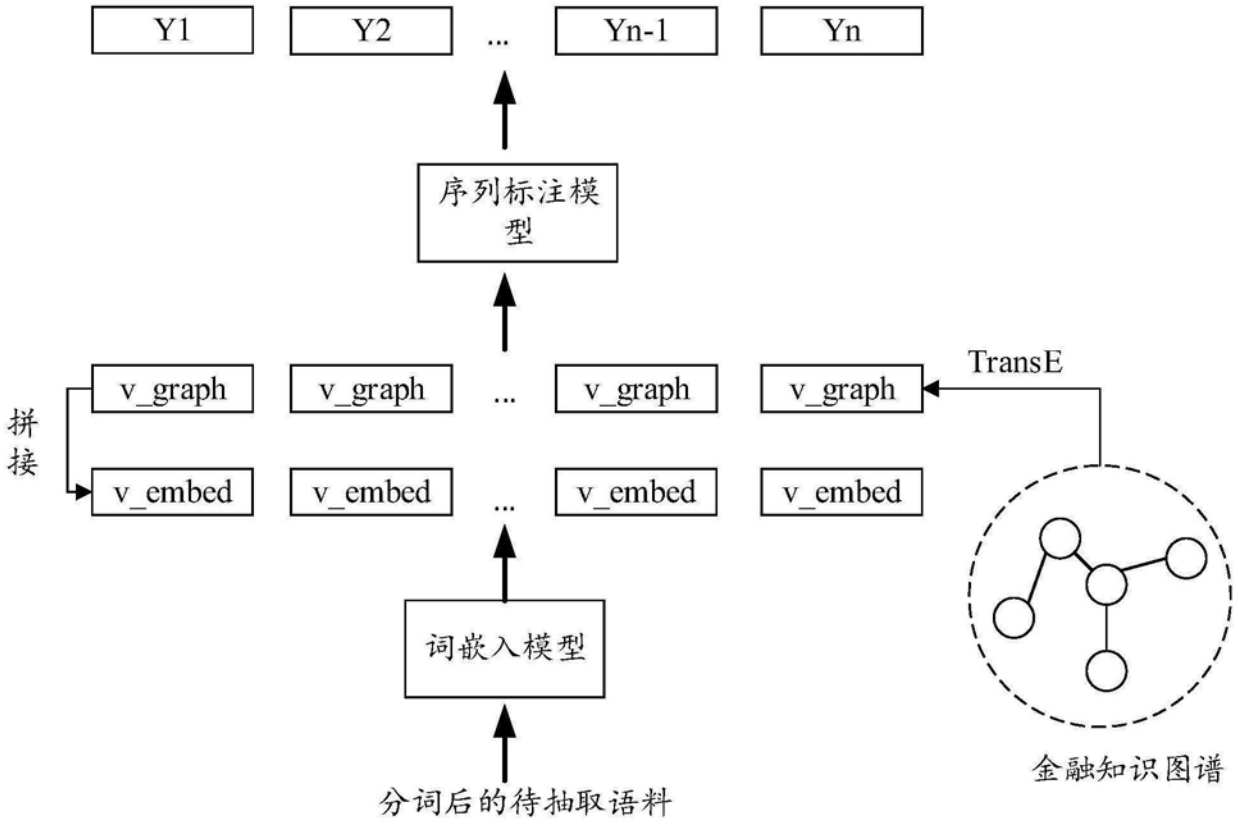

Y1
Y2
...
Yn-1
Yn
序列标注模型
v_graph
v_graph
...
v_graph
v_graph
TransE
拼接
v_embed
v_embed
...
v_embed
v_embed
词嵌入模型
分词后的待抽取语料
金融知识图谱

图3

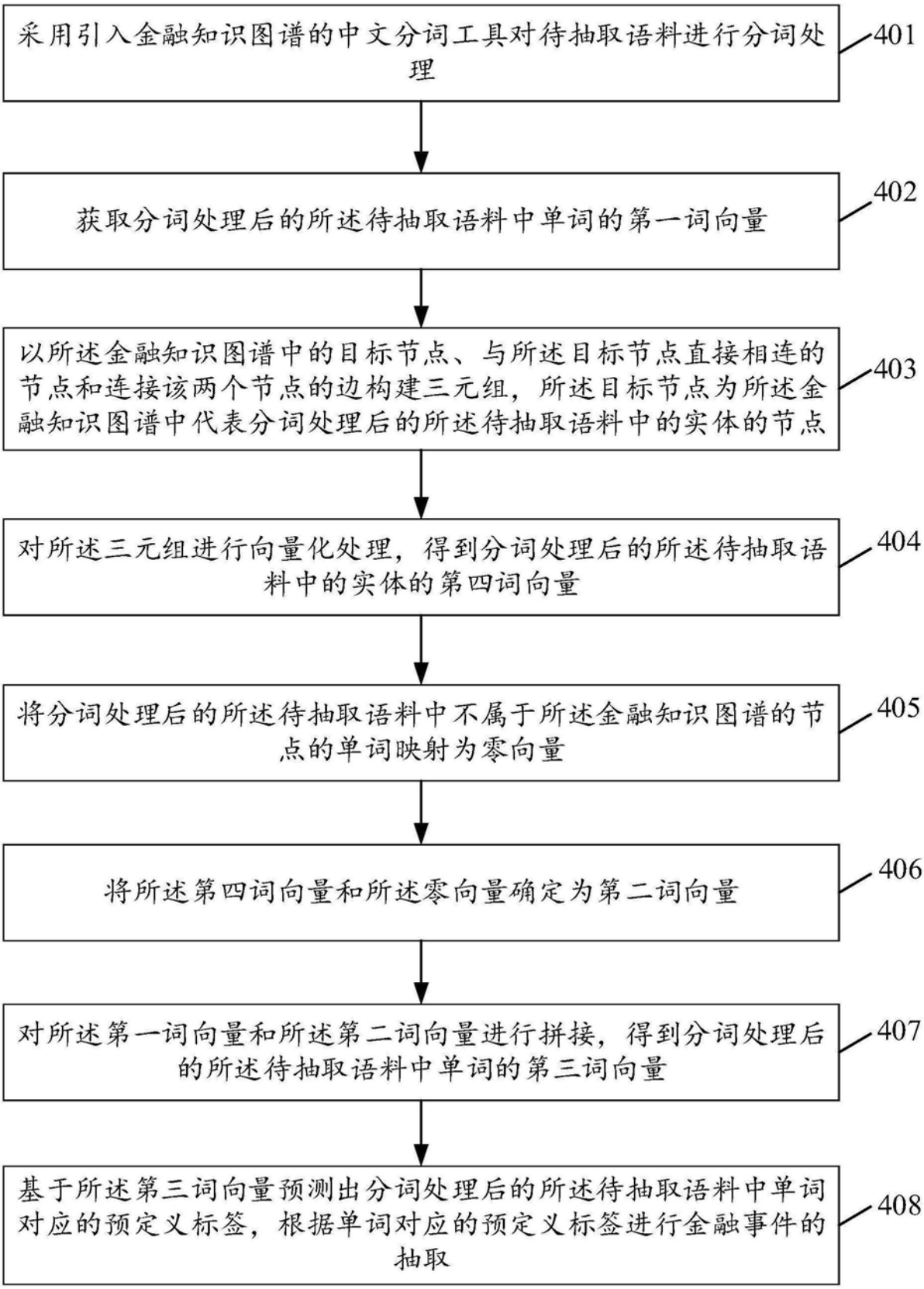

采用引入金融知识图谱的中文分词工具对待抽取语料进行分词处理
401
获取分词处理后的所述待抽取语料中单词的第一词向量
402
以所述金融知识图谱中的目标节点、与所述目标节点直接相连的节点和连接该两个节点的边构建三元组，所述目标节点为所述金融知识图谱中代表分词处理后的所述待抽取语料中的实体的节点
403
对所述三元组进行向量化处理，得到分词处理后的所述待抽取语料中的实体的第四词向量
404
将分词处理后的所述待抽取语料中不属于所述金融知识图谱的节点的单词映射为零向量
405
将所述第四词向量和所述零向量确定为第二词向量
406
对所述第一词向量和所述第二词向量进行拼接，得到分词处理后的所述待抽取语料中单词的第三词向量
407
基于所述第三词向量预测出分词处理后的所述待抽取语料中单词对应的预定义标签，根据单词对应的预定义标签进行金融事件的抽取
408

图4

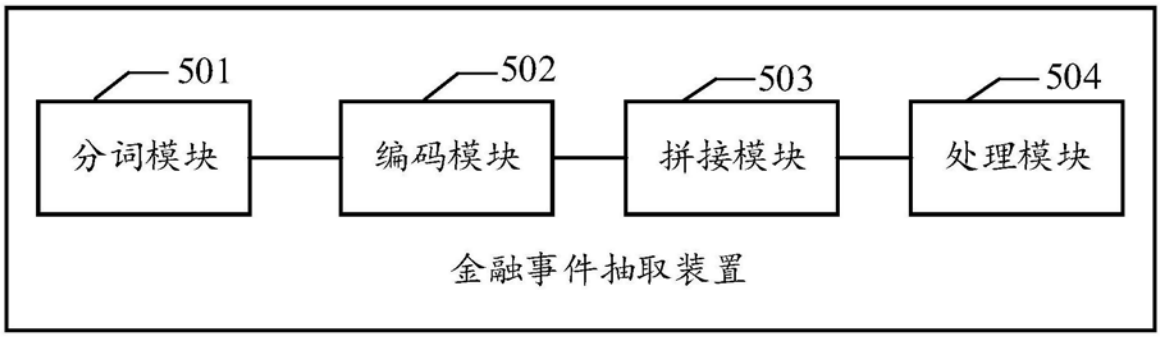

501
分词模块
502
编码模块
503
拼接模块
504
处理模块
金融事件抽取装置

图5

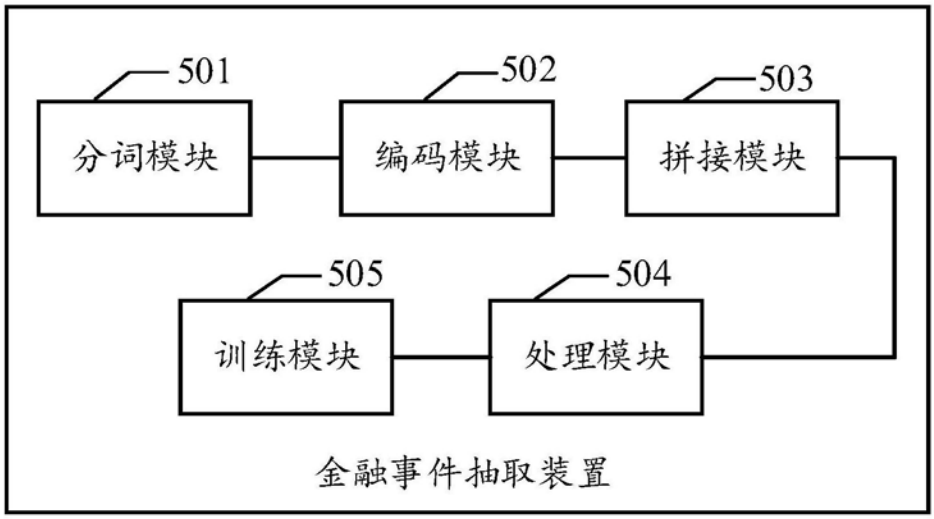

501
分词模块
502
编码模块
503
拼接模块
505
训练模块
504
处理模块
金融事件抽取装置

图6

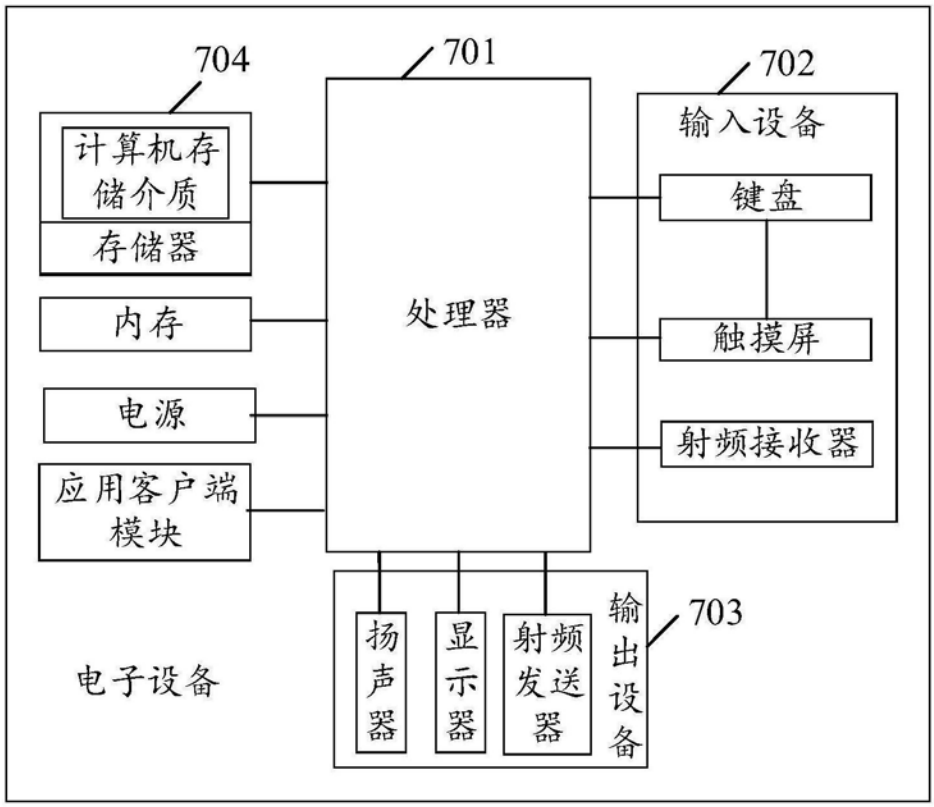

704
701
702
计算机存储介质
存储器
内存
电源
应用客户端模块
处理器
输入设备
键盘
触摸屏
射频接收器
扬声器
显示器
射频发送器
输出设备
703
电子设备

图7