(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 111723564 B
(45)授权公告日 2022.12.09

(21)申请号 202010464119.X

(22)申请日 2020.05.27

(65)同一申请的已公布的文献号
申请公布号 CN 111723564 A

(43)申请公布日 2020.09.29

(73)专利权人 西安交通大学
地址 710049 陕西省西安市咸宁西路28号

(72)发明人 赵银亮 屈垠岑 刘硕 酒冲冲 李椿茂

(74)专利代理机构 西安通大专利代理有限责任公司 61200
专利代理师 姚咏华

(51)Int.Cl.
G06F 40/242(2020.01)
G06F 40/289(2020.01)
G06F 40/211(2020.01)
G06F 40/253(2020.01)
G06F 16/35(2019.01)
G06K 9/62(2022.01)
G06Q 50/18(2012.01)

审查员 常建军

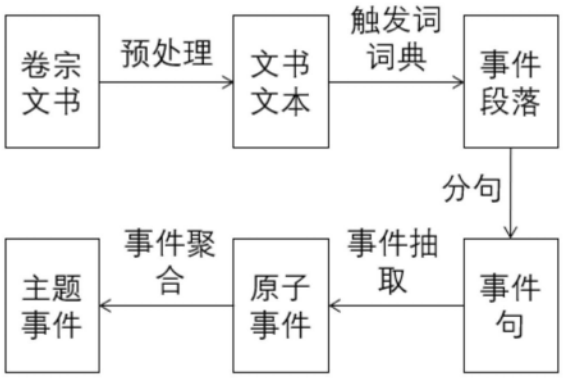
权利要求书2页 说明书5页 附图1页

(54)发明名称

一种针对随案电子卷宗的事件抽取及处理方法

(57)摘要

本发明公开了一种针对随案电子卷宗的事件抽取及处理方法，包括以下步骤：步骤一，从随案电子卷宗流转处理平台上获取需要的卷宗数据并存入数据库；步骤二，构建事件触发词词典，匹配电子卷宗事件描述段落，再进行分句、分词、词性标注等文本预处理方法；步骤三，事件属性抽取，结合依存句法分析与语义角色标注方法获得事件的施加者、承受者、行为、时间、地点、方式这6个事件属性；步骤四，事件聚合，将原子事件聚合为主题事件，合并相似主题事件，将主题事件存入事件数据库。本发明解决了随案电子卷宗快速获取犯罪事实信息的难题，可以更加准确地对犯罪事实进行事件抽取和组织，是高效、高质量阅卷的基础。



CN 111723564 B

1.一种针对随案电子卷宗的事件抽取及处理方法,其特征在于,包括以下步骤:

步骤1,获取需要的电子卷宗,提取电子卷宗的文本文书和格式文书的正文内容,对所述正文内容进行分段,并存入数据库建立索引,完成卷宗数据的定位;

步骤2,对电子卷宗进行分析,总结电子卷宗特定格式,构建触发词词典,依据触发词词典识别电子卷宗事件描述段落,再对事件描述段落进行分句、分词和词性标注,分句后得到事件句;

步骤3,对步骤2产生的每一句事件句进行事件属性抽取,得到多个原子事件,事件属性包括:施加者、承受者、行为、时间、地点和行为方式,具体为:根据步骤2中产生的分词和词性标注进行句法依存分析,得到核心词、核心词主谓关系以及核心词的动宾关系,即依存句法分析结果,通过语义角色标注技术得到语义角色分析结果,再用启发式规则结合依存句法分析结果和语义角色分析结果得到对应事件属性;

步骤4,对步骤3产生的每一个原子事件按所在事件句聚合为主题事件;计算各个事件句之间的相似性,将相似的事件句所在主题事件进行合并。

2.根据权利要求1所述的一种针对随案电子卷宗的事件抽取及处理方法,其特征在于,所述步骤1中,用规则匹配对文书中标题和结尾进行过滤,获取文本文书和格式文书的正文内容。

3.根据权利要求1所述的一种针对随案电子卷宗的事件抽取及处理方法,其特征在于,所述步骤2中,根据触发词词典中的开始词匹配段落首句,若开始词所在段落只有一句话,则自开始词所在段落的下一段至触发词词典中的停止词所在段落均为事件描述段落;否则,则自开始词所在段落至触发词词典中的停止词所在段落均为事件描述段落。

4.根据权利要求1所述的一种针对随案电子卷宗的事件抽取及处理方法,其特征在于,所述步骤2中,分句、分词和词性标注的具体过程为:利用标点符号对事件描述段落进行分句,然后利用汉语分词系统进行分词和词性标注。

5.根据权利要求1所述的一种针对随案电子卷宗的事件抽取及处理方法,其特征在于,所述步骤3包括以下步骤:

步骤3.1、句法分析:对步骤2的中产生的分词和词性标注进行依存句法解析,获得事件句中核心词、核心词的主谓关系和核心词的动宾关系,获取句子中的主语和宾语;

步骤3.2、语义角色抽取:对步骤2得到的事件句和事件句词性标注进行语义角色抽取,获得语义角色参数,语义角色参数包括:谓语动词参数、持有者参数、被持有者参数、时间参数、地点参数和方向参数;

步骤3.3、利用启发式规则确定事件属性:

若持有者参数和被持有者参数都不为空,则将持有者参数输入待抽取事件的施加者属性中,被持有者参数输入待抽取事件的承受者属性中;

若持有者参数为空且主语不为空,则将主语输入待抽取事件的施加者属性中;

若被持有者参数为空且宾语不为空,则将主语输入待抽取事件的承受者属性中;

若当前语义角色参数是时间参数且不为空,则将当前语义角色参数的对应词输入至待抽取事件的时间属性中;

若当前语义角色参数是地点参数且不为空,则将当前语义角色参数的对应词输入待抽取事件的地点属性中;

若当前语义角色参数是方向参数且不为空,则将当前语义角色参数的对应词输入待抽取事件的方式属性中;

若当前语义角色参数是谓语动词参数且不为空,则将当前语义角色参数的对应词输入待抽取事件的行为属性中;

若当前语义角色参数为空则不动作。

6.根据权利要求1所述的一种针对随案电子卷宗的事件抽取及处理方法,其特征在于,所述步骤4的具体过程为:对步骤3产生的每一个原子事件按所在事件句聚合为主题事件,再对每一句事件句进行语义相似性分析,对于判定结果为相似的事件句的主题事件进行合并,将合并后的主题事件存入事件数据库。

7.根据权利要求1所述的一种针对随案电子卷宗的事件抽取及处理方法,其特征在于,所述步骤4中,对于不同的主题事件,其所在的不同事件句间两两通过预训练词向量模型进行语义相似性计算,将待计算的两个事件句中所有的单词输入预训练词向量模型,获得各个单词对应的词向量,将每个句子的词向量分别叠加,得到两个句子向量,计算句子向量间的余弦相似度,若相似度超过阈值,则认为两句话描述的是同一事件,两个事件句所包含的主题事件为相似事件。

8.根据权利要求7所述的一种针对随案电子卷宗的事件抽取及处理方法,其特征在于,所述预训练词向量模型为Word2vec。

一种针对随案电子卷宗的事件抽取及处理方法

技术领域

[0001] 本发明属于自然语言处理技术领域，具体涉及一种针对随案电子卷宗的事件抽取及处理方法。

背景技术

[0002] 随着科学的不断发展，技术的不断进步，司法信息化建设也在不断深入，当前各级司法部门(检察院、法院、司法部)存储的案件电子卷宗每年以千万级的数量递增，海量卷宗文档蕴含着我国司法从业者多年的集体智慧，其中包含的数据信息是以现有技术尚无法充分利用的重要数据资产。根据第三方评估报告显示，2016年全国各级法院已全部实现案件信息数字化。目前司法信息化建设初步解决了卷宗的电子化和存储问题，但由于随案电子卷宗涉及文件多、内容复杂等情况常会给阅卷人带来“信息过载”和“认知迷航”的问题。在以审判为中心的判案体系中，检察官、法官的阅卷过程是非常关键的一环，因此，提高阅卷效率和避免重要信息遗漏是目前急需解决的问题。在此基础上，从随案电子卷宗中抽取出有用信息并有效地可视化展示的方法研究与实现具有重要意义。

[0003] 信息抽取是指结构化地提取出自然语言文本中的信息。信息的自动抽取及相关处理技术对于从数量多内容不统一的文档中提取出有用的信息是十分有意义的。有效的信息抽取技术能显著地降低信息获取难度，减少阅读负担。事件抽取是指从原始文本中提取事件信息，是特定的人、物、事在特定时间、特定地点，主要是抽取文本中的事件实例，并为每个抽取的事件实例抽取论元赋予相应的角色。

[0004] 现有的事件抽取方法主要集中在深度学习领域的研究上，深度学习的方法主要依赖大规模的语料库，准确率和数据质量相关度很高。中国专利CN106951438等一些方法通过卷积神经网络模型进行触发词识别，通过图模型获取事件参数，需要大规模标注语料，但由于司法领域随案电子卷宗数据的保密性和敏感性，想要获取大规模高质量数据是难以实现的；中国专利CN10892044采用基于句法分析的事件参数抽取方法，中国专利CN106951438采用基于规则的事件属性抽取的方法，这两类方法不需要标注语料但部分参数识别准确率不高。因此，需要提出一种新的事件属性抽取的方法。

发明内容

[0005] 本发明通过启发式规则结合句法分析和语义分析的方法抽取事件，该方法主要针对随案电子卷宗进行事件抽取，通过对司法领域案卷特定格式的分析，能够高效准确地对电子卷宗中的事件进行抽取。

[0006] 为达到上述目的，本发明一种针对随案电子卷宗的事件抽取及处理方法，包括以下步骤：

[0007] 步骤1，获取需要的电子卷宗，提取电子卷宗的文本文书和格式文书的正文内容，对正文内容进行分段，并存入数据库建立索引，完成卷宗数据的定位；

[0008] 步骤2，对电子卷宗进行分析，总结电子卷宗特定格式，构建触发词词典，依据触发

词词典识别电子卷宗事件描述段落，再对事件描述段落进行分句、分词和词性标注，分句后得到事件句；

[0009] 步骤3，对步骤2产生的每一句事件句进行事件属性抽取，得到多个原子事件，事件属性包括：施加者、承受者、行为、时间、地点和行为方式，具体为：根据步骤2中产生的分词和词性标注进行句法依存分析，得到核心词、核心词主谓关系以及核心词的动宾关系，即依存句法分析结果，通过语义角色标注技术得到语义角色分析结果，再用启发式规则结合依存句法分析结果和语义角色分析结果得到对应事件属性；

[0010] 步骤4，对步骤3产生的每一个原子事件按所在事件句聚合为主题事件；计算各个事件句之间的相似性，将相似的事件句所在主题事件进行合并。

[0011] 进一步的，步骤1中，用规则匹配对文书中标题和结尾进行过滤，获取文本文书和格式文书的正文内容。

[0012] 进一步的，步骤2中，根据触发词词典中的开始词匹配段落首句，若开始词所在段落只有一句话，则自开始词所在段落的下一段至触发词词典中的停止词所在段落均为事件描述段落；否则，则自开始词所在段落至触发词词典中的停止词所在段落均为事件描述段落。

[0013] 进一步的，步骤2中，分句、分词和词性标注的具体过程为：利用标点符号对事件描述段落进行分句，然后利用汉语分词系统进行分词和词性标注。

[0014] 进一步的，步骤3包括以下步骤：

[0015] 步骤3.1、句法分析：对步骤2的中产生的分词和词性标注进行依存句法解析，获得事件句中核心词、核心词的主谓关系和核心词的动宾关系，获取句子中的主语和宾语；

[0016] 步骤3.2、语义角色抽取：对步骤2得到的事件句和事件句词性标注进行语义角色抽取，获得语义角色参数，语义角色参数包括：谓语动词参数、持有者参数、被持有者参数、时间参数、地点参数和方向参数；

[0017] 步骤3.3、利用启发式规则确定事件属性：

[0018] 若持有者参数和被持有者参数都不为空，则将持有者参数输入待抽取事件的施加者属性中，被持有者参数输入待抽取事件的承受者属性中；若持有者参数为空且主语不为空，则将主语输入待抽取事件的施加者属性中；若被持有者参数为空且宾语不为空，则将主语输入待抽取事件的承受者属性中；

[0019] 若当前语义角色参数是时间参数且不为空，则将当前语义角色参数的对应词输入至待抽取事件的时间属性中；若当前语义角色参数是地点参数且不为空，则将当前语义角色参数的对应词输入待抽取事件的地点属性中；若当前语义角色参数是方向参数且不为空，则将当前语义角色参数的对应词输入待抽取事件的方式属性中；若当前语义角色参数是谓语动词参数且不为空，则将当前语义角色参数的对应词输入待抽取事件的行为属性中；

[0020] 若当前语义角色参数为空则不动作。

[0021] 进一步的，步骤4的具体过程为：对步骤3产生的每一个原子事件按所在事件句聚合为主题事件，再对每一句事件句进行语义相似性分析，对于判定结果为相似的事件句的主题事件进行合并，将合并后的主题事件存入事件数据库。

[0022] 进一步的，步骤4中，对于不同的主题事件，其所在的不同事件句间两两通过预训

练词向量模型进行语义相似性计算，将待计算的两个事件句中所有的单词输入预训练词向量模型，获得各个单词对应的词向量，将每个句子的词向量分别叠加，得到两个句子向量，计算句子向量间的余弦相似度，若相似度超过阈值，则认为两句话描述的是同一事件，两个事件句所包含的主题事件为相似事件。

[0023] 进一步的，预训练词向量模型为Word2vec。

[0024] 与现有技术相比，本发明至少具有以下有益的技术效果：

[0025] 本发明利用基于语义和句法分析的事件属性抽取的方法，针对每一个事件标注事件属性，分别为事件的施加者、承受者、发生时间、地点、行为以及方式，可以更加准确地对随案电子卷宗中的犯罪事实进行事件抽取，明确事件内容。通过事件聚合的方式实现了对随案电子卷宗中的犯罪事实的重新组织，组织后的事件更符合人的直观思维，避免了信息冗余，是高效、高质量阅卷的基础。且本发明不需要提前针对卷宗数据进行标注、训练，同时保证了卷宗数据的私密性和事件抽取结果的可用性。通过针对卷宗特性和中文语义特点设计的匹配模式，本发明能够高效准确地对随案电子卷宗的事件进行抽取，具有一定的工业实用价值。

[0026] 合并后的事件更直观，合理组织了原子事件。事件聚合利用中文一句话表达一个主题的习惯对原子事件进行组织，聚合了原子事件的分散信息，信息抽取的结果更便于理解。同时主题事件合并避免了多文书描述同一事件带来的信息冗余，有利于快速发现多文书描述的哪些事件是相似的。

附图说明

[0027] 图1为本发明的一种针对随案电子卷宗的事件抽取及处理方法的流程示意图；

[0028] 图2为本发明所述事件属性抽取流程图。

具体实施方式

[0029] 为了使本发明的目的和技术方案更加清晰和便于理解。以下结合附图和实施例，对本发明进行进一步的详细说明，此处所描述的具体实施例仅用于解释本发明，并非用于限定本发明。

[0030] 本发明针对随案电子卷宗领域的中文犯罪事实信息，提供一种针对随案电子卷宗的事件抽取及处理方法。对于随案电子卷宗中存在的犯罪事实，利用自然语言处理相关技术，对犯罪事实进行结构化的事件信息抽取，为犯罪事实智能聚合，高效阅卷判案提供了基础。

[0031] 一种针对随案电子卷宗的事件抽取及处理方法如图1所示，主要包括以下步骤：

[0032] 步骤一，文本数据预处理，从随案电子卷宗存储系统上获取需要的卷宗数据，卷宗数据包括文本文书和格式文书，去除卷宗数据中的多余信息，以提取文本文书和格式文书的正文内容，对正文内容进行分段，并存入数据库建立索引，完成卷宗数据的定位。

[0033] 具体包括以下内容：

[0034] 1)针对随案电子卷宗包含特定事件信息的文书选取。随案电子卷宗的犯罪事实信息分散在起诉书和裁判文书等文本文书以及证人证言等格式文书中，针对这两类特定文书内容进行事件抽取，可以找到犯罪事实信息。

[0035] 2)对于文本文书,用规则匹配过滤文书中标题和结尾等无关信息获取正文内容:随案电子卷宗中犯罪事实信息的文书通常除了正文还包括标题和结尾签名,标题和结尾签名为多余信息,根据文书固定结构筛选过滤出正文内容,并以分段符作为标志进行切割分段,存入数据库。

[0036] 3)对于格式文书,用固定格式匹配文书进行过滤获取正文内容。以证人证言为例,根据文书结构过滤掉标题和结尾的签字格式,得到正文内容。再依据段落格式过滤掉第一句的个人身份介绍,获取证人证言中关于犯罪事实的有效描述段落,存入数据库。

[0037] 步骤二,事件触发词识别,依靠人工对电子卷宗数据进行分析,总结电子卷宗特定格式,构建触发词词典,依据触发词词典识别电子卷宗事件描述段落,再对该段落进行分句、分词和词性标注文本预处理方法,得到事件句和事件句词性标注结果。

[0038] 其中步骤二的事件触发词识别具体包括依次进行的以下内容:

[0039] 1)触发词词典构建。对于随案电子卷宗的裁判文书、起诉书特定格式进行分析,构建触发词词典,首先对事件段落的开始词进行归纳总结,得到开始词“审理查明”、“审查查明”、“复核确认”、“判决认定”、“事实和理由”5个,用来表示开始犯罪事实描述的词语,然后对随案电子卷宗标志犯罪事实结束的文本段落中的停止词进行总结,得到停止词“上述事实”1个,用来表示结束犯罪事实描述的词语,形成触发词词典。

[0040] 2)触发词匹配。利用触发词词典中的开始词匹配卷宗文本段落,若该段落就一句话,则识别下一段落为事件描述段落,不断识别下一段落为事件描述段,直到到达触发词词典中的停止词匹配段落,识别段落都认为是事件段落,如果识别到文本末依然没有识别到停止词,则只认为开始词所在段落为事件段落。如果该段落不是一句话,则识别本段落,并不断识别下一段落为事件描述段,直到到达触发词词典中的停止词匹配段落,识别段落都认为是事件段落,如果识别到文本末依然没有识别到停止词,则只认为开始词所在段落为事件段落。

[0041] 3)分句、分词和词性标注。针对电子卷宗的事件抽取是逐句进行的,为更好地匹配事件抽取模式,需要对事件段落文本进行分句处理,将段落文本划分为具有较完整要素的句子即事件句。段落文本的分句实际上是对标点符号的处理,标点符号是句子中表示停顿与句调的辅助性符号,识别出如“。”和“;”等符号对文本进行切割。然后要对分句后的卷宗文本进行中文分词和词性标注,选用哈工大PYLTP汉语分词系统进行分词和词性标注。

[0042] 步骤三,事件属性抽取,对步骤二产生的每一句事件句进行事件属性抽取,得到多个原子事件,每一个原子事件包括施加者、承受者、行为、时间、地点和行为方式这6个属性。

[0043] 如图2所示,步骤三的事件属性抽取过程如下:

[0044] 1)句法分析。使用哈工大PYLTP依存句法分析系统对步骤二的中产生的分词和词性标注进行依存句法解析,获得句子中核心词HED及核心词的主谓关系SBV和动宾关系VOB,获取句子中的主语和宾语。

[0045] 2)语义角色抽取。使用哈工大PYLTP语义角色标注系统对步骤二的得到的事件句和事件句词性标注结果进行语义角色抽取,获得语义角色参数,语义角色参数包括:谓语动词参数(PRD)、持有者参数(PSR)、被持有者参数(PSE)、时间参数(TMP)、地点参数(LOC)和方向参数(DIR)。

[0046] 3)利用启发式规则确定事件属性。若持有者参数(PSR)和被持有者参数(PSE)都不

为空,则将持有者参数(PSR)输入待抽取事件的施加者属性中,被持有者参数(PSE)输入待抽取事件的承受者属性中。若持有者参数(PSR)为空且主语不为空,则将主语输入待抽取事件的施加者属性中。若被持有者参数为空且宾语不为空,则将主语输入待抽取事件的承受者属性中。若当前语义角色参数是时间(TMP)且不为空,则将当前语义角色参数的对应词输入至待抽取事件的时间属性中;若当前语义角色参数是地点(LOC)且不为空,则将当前语义角色参数的对应词输入待抽取事件的地点属性中;若当前语义角色参数是方向参数(DIR)且不为空,则将当前语义角色参数的对应词输入待抽取事件的方式属性中;若当前语义角色参数是谓语动词参数(PRD)且不为空,则将当前语义角色参数的对应词输入待抽取事件的行为属性中。若当前语义角色参数为空则不动作。

[0047] 步骤四,事件聚合,对步骤三产生的每一个原子事件按所在事件句聚合为主题事件;通过预训练词向量模型进行相似度计算,得到事件句之间的相似性,对于判定结果为相似的事件句所在主题事件进行合并,将合并后的主题事件存入事件数据库。

[0048] 事件聚合方法具体包括以下内容:

[0049] 1)对于步骤三中所获得的原子事件按其所在事件句分别聚合,认为在同一事件句的原子事件是在表达同一主题,因此将在同一事件句中的所有原子事件标志为同一主题事件,一个主题事件由一个事件句中所有原子事件的集合表示。即当某一主题事件中有一个事件句和另一个主题事件中的某一事件句相似时,则将这两个主题事件进行合并。

[0050] 2)对于不同的主题事件,其所在的不同事件句间两两通过预训练词向量模型如word2vec进行语义相似性计算。将待计算的两个事件句中所有的单词输入预训练词向量模型,获得各个单词对应的词向量,将每个句子的词向量分别叠加,得到两个句子向量,计算句子向量间的余弦相似度。设立相似度阈值为0.8,若相似度超过0.8的认为两句话描述的是同一事件,两个事件句所包含的主题事件为相似事件。因此将两个主题事件合并为一个主题事件,认为这个主题事件拥有两个事件句,其原子事件集合为合并前两个主题事件的原子事件的并集,即参数相同的原子事件只记一次,参数不同的原子事件分别保留,合并后的主题事件存入事件数据库。

[0051] 本发明未详细阐述部分属于本领域公知技术。

[0052] 以上所述,仅为本发明部分具体实施方式,但本发明的保护范围并不局限于此,任何熟悉本领域的人员在本发明揭露的技术范围内,可轻易想到的变化或替换,都应涵盖在本发明的保护范围之内。

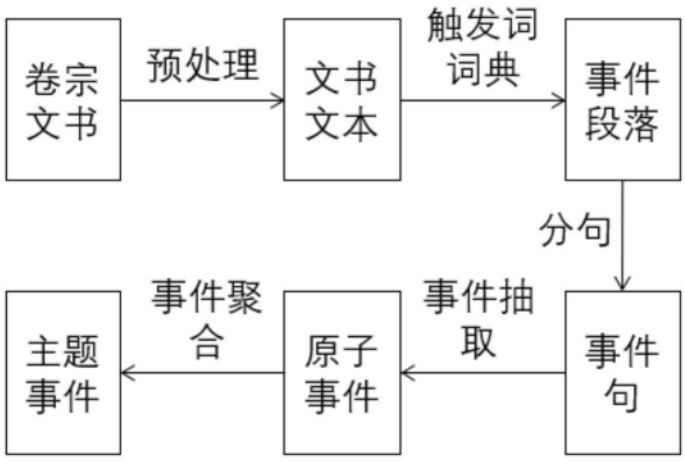

卷宗文书
预处理
文书文本
触发词词典
事件段落
分句
事件句
事件抽取
原子事件
事件聚合
主题事件

图1

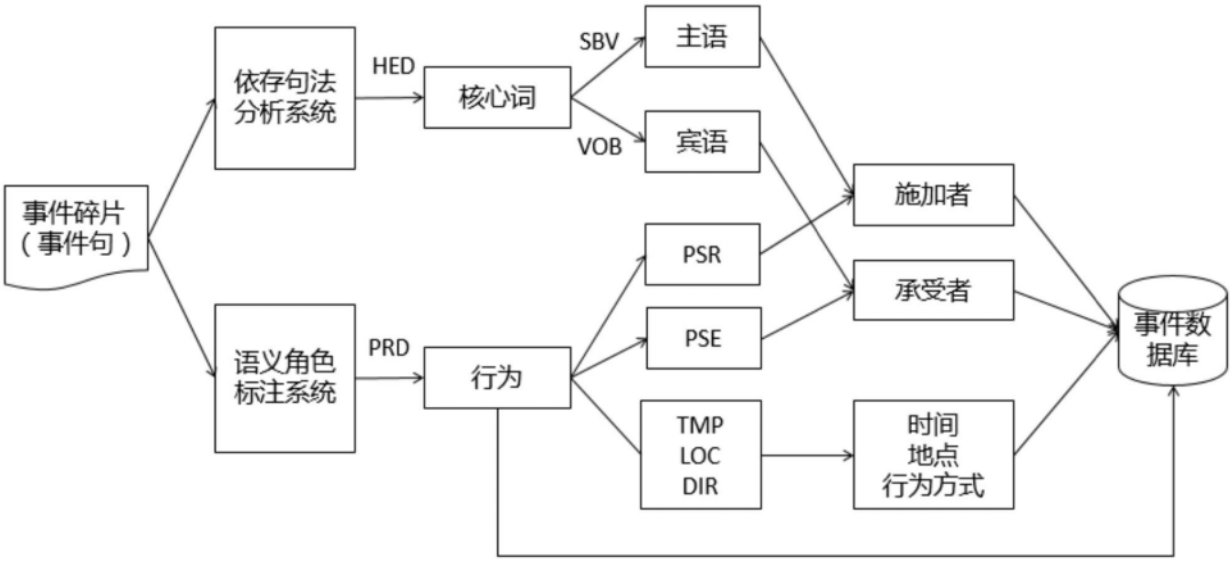

事件碎片（事件句）
依存句法分析系统
HED
核心词
SBV
主语
VOB
宾语
语义角色标注系统
PRD
行为
PSR
PSE
TMP
LOC
DIR
施加者
承受者
时间
地点
行为方式
事件数据库

图2