(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 111813931 B
(45)授权公告日 2021.03.16

(21)申请号 202010548917.0
(22)申请日 2020.06.16
(65)同一申请的已公布的文献号
申请公布号 CN 111813931 A
(43)申请公布日 2020.10.23
(73)专利权人 清华大学
地址 100084 北京市海淀区双清路30号清华大学
(72)发明人 许斌 仝美涵 李涓子 侯磊
(74)专利代理机构 北京路浩知识产权代理有限公司 11002
代理人 郭亮
(51)Int.Cl.
G06F 16/35(2019.01)
G06F 40/30(2020.01)
G06K 9/62(2006.01)
G06N 3/04(2006.01)
G06N 3/08(2006.01)

(56)对比文件
CN 106951438 A,2017.07.14
CN 110188172 A,2019.08.30
CN 111159336 A,2020.05.15
CN 110751260 A,2020.02.04
WO 2016189084 A1,2016.12.01
严浩.开放式中文事件检测研究.《广西师范大学学报(自然科学版)》.2020,第38卷(第2期),第64-71页.
Aakanksha Naik.Towards Open Domain Event Trigger Identification using Adversarial Domain Adaptation.《arXiv数据库》.2020,

审查员 徐晓孜

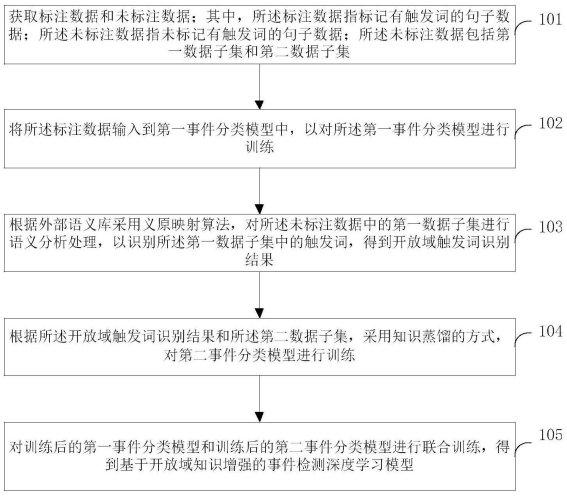
权利要求书3页 说明书13页 附图3页

(54)发明名称

事件检测模型的构建方法、装置、电子设备及存储介质

(57)摘要

本发明实施例提供了一种基于开放域知识增强的事件检测深度学习模型的构建方法、装置、电子设备及存储介质,方法包括:获取标注数据和未标注数据;将标注数据输入到第一事件分类模型中,进行训练;根据外部语义库采用义原映射算法,对未标注数据中的第一数据子集进行处理,得到开放域触发词识别结果;根据开放域触发词识别结果和第二数据子集,采用知识蒸馏的方式,对第二事件分类模型进行训练;对训练后的第一事件分类模型和训练后的第二事件分类模型进行联合训练,得到基于开放域知识增强的事件检测深度学习模型。本发明实施例得到的基于开放域知识增强的事件检测深度学习模型可以有效解决各类标注分布不均匀的长尾难题。



CN 111813931 B

1.一种基于开放域知识增强的事件检测深度学习模型的构建方法,其特征在于,包括:

获取标注数据和未标注数据;其中,所述标注数据指标记有触发词的句子数据;所述未标注数据指未标记有触发词的句子数据;所述未标注数据包括第一数据子集和第二数据子集;

将所述标注数据输入到第一事件分类模型中,以对所述第一事件分类模型进行训练;

根据外部语义库采用义原映射算法,对所述未标注数据中的第一数据子集进行语义分析处理,以识别所述第一数据子集中的触发词,得到开放域触发词识别结果;

根据所述开放域触发词识别结果和所述第二数据子集,采用知识蒸馏的方式,对第二事件分类模型进行训练;

对训练后的第一事件分类模型和训练后的第二事件分类模型进行联合训练,得到基于开放域知识增强的事件检测深度学习模型。

2.根据权利要求1所述的基于开放域知识增强的事件检测深度学习模型的构建方法,其特征在于,根据外部语义库采用义原映射算法,对所述未标注数据中的第一数据子集进行语义分析处理,以识别所述第一数据子集中的触发词,得到开放域触发词识别结果,包括:

基于外部语义库WordNet对第一数据子集进行词语消歧,将第一数据子集中的词对应到WordNet中单一语义的义原集中;

根据第一数据子集中每个词所属的义原集是否触发事件识别所述第一数据子集每个词是否为触发词,以得到开放域触发词识别结果。

3.根据权利要求1所述的基于开放域知识增强的事件检测深度学习模型的构建方法,其特征在于,所述第二事件分类模型包括学生模型和教师模型;

相应地,根据所述开放域触发词识别结果和所述第二数据子集,采用知识蒸馏的方式,对第二事件分类模型进行训练,包括:

以拥有开放域触发词识别结果的第一数据子集作为教师模型的输入,以所述第二数据子集作为学生模型的输入,以教师模型和学生模型的预测结果相同为训练目标,对所述教师模型和学生模型进行训练。

4.根据权利要求3所述的基于开放域知识增强的事件检测深度学习模型的构建方法,其特征在于,以拥有开放域触发词识别结果的第一数据子集作为教师模型的输入,以所述第二数据子集作为学生模型的输入,以教师模型和学生模型的预测结果相同为训练目标,对所述教师模型和学生模型进行训练,包括:

设定训练目标:

$p(Y|S^{+},\theta)=p(Y|S^{-},\theta)$

其中,$p(Y|S^{+},\theta)$与$p(Y|S^{-},\theta)$分别为教师模型和学生模型的预测结果;其中,教师模型的输入S^{+}是标记有开放域触发词知识的第一数据子集,学生模型的输入S^{-}则是未标记开放域触发词知识的第二数据子集;其中,θ表示教师模型和学生模型共享的参数群,Y表示事件类型预测结果,其中,S^{+}的构造过程包括:引入两个符号B-TRI和E-TRI,标记开放域触发词在句子中开始位置和结束位置,B-TRI表示开始位置,E-TRI表示结束位置,给定原始句子$S=<w_1,w_2,...,w_n>$以及开放域触发词w_i,编码进开放域触发词的句子表示为$S^{+}=<w_1,w_2,...,$B-TRI$,w_i,$E-TRI$,...,w_n>$;其中,S^{-}的构造过程包括:通过随机屏蔽开放域触发词

的事件性词语，扰乱学生模型的输入，给定原始句子S＝＜w_1，w_2，...，w_n＞以及开放域触发词w_i，构造S^-＝{w_1，w_2，...，[MASK]，...，w_n}；其中，[MASK]表示随机屏蔽了一部分开放域触发词；

通过将句子S^+以及S^-分别输入教师模型和学生模型，获得教师模型和学生模型两者的预测结果p(Y|S^+，θ)和p(Y|S^-，θ)；

若未标注数据为$U=\{S_i\}|_{i=1}^{N_U}$，则第二事件分类模型的优化函数为：

$$J_T(\theta) = KL(p(Y|S^+,\theta)||p(Y|S^-,\theta)$$

$$= \sum_i^{N_U} p\left(Y_{(i)}\middle|S_{(i)}^+,\theta\right)\frac{p\left(Y_{(i)}\middle|S_{(i)}^+,\theta\right)}{p\left(Y_{(i)}\middle|S_{(i)}^-,\theta\right)};$$

其中，$J_T(\theta)$表示衡量教师和学生模型预测分布差距的损失函数，KL表示信息增益散度，||表示分布相比运算符，N_U表示未标注数据的规模，$p\left(Y_{(i)}\middle|S_{(i)}^+,\theta\right)$表示教师模型的预测分布，$p\left(Y_{(i)}\middle|S_{(i)}^-,\theta\right)$表示学生模型的预测分布。

5.根据权利要求4所述的基于开放域知识增强的事件检测深度学习模型的构建方法，其特征在于，采用屏蔽词填写任务，利用周围的单词学习引入符号B-TRI和E-TRI的语义表示。

6.根据权利要求4所述的基于开放域知识增强的事件检测深度学习模型的构建方法，其特征在于，将所述标注数据输入到第一事件分类模型中，以对所述第一事件分类模型进行训练，包括：

若S_i＝＜w_1，w_2，...，w_n＞和Y_i＝＜v_1，v_2，...，v_n＞分别表示第i个训练句子及其事件类型；

通过全连接层将句子的隐含表示H转化为中间表示O，以将表示维度和事件个数对齐来执行计算预测概率O_{ijc}＝δ(WH+b)的步骤；

其中，W和b是全连接层的参数，随机初始化并在训练过程中不断优化，O_{ijc}表示S_i中的第j个单词属于第c个事件类别的概率；

通过softmax函数归一化O，以获得条件概率；

$$p(Y_i|S_i,\theta) = \sum_{j=1}^{n}\frac{\exp(O_{ijc})}{\sum_{c=1}^{C}\exp(O_{ijc})}/n$$

若标注数据为$L = \{S_i, Y_i|_{i=1}^{N_L}\}$，则第一事件分类模型的优化函数为：

$$J_L(\theta) = -\sum_{i=1}^{N_L}\log p(Y_i|S_i,\theta)$$

其中，p(Y_i|S_i，θ)表示事件各分类上的条件概率，n表示句子中的词数，exp(O_{ijc})表示第i个训练句子中第j个词语在第c个事件类型上归一化的概率，C表示事件类型数，$J_L(\theta)$表示事件分类的损失函数，N_L表示标注数据的个数。

7.根据权利要求6所述的基于开放域知识增强的事件检测深度学习模型的构建方法，其特征在于，对训练后的第一事件分类模型和训练后的第二事件分类模型进行联合训练，得到基于开放域知识增强的事件检测深度学习模型，包括：

建立待进行联合训练的基于开放域知识增强的事件检测深度学习模型：

$J(\theta)=J_L(\theta)+\lambda J_T(\theta)$

对上述模型进行联合训练，并在计算J_T时，停止教师模型的梯度下降，以确保学习是从教师模型到学生模型的；其中，$J(\theta)$表示总体的损失函数，λ表示调和系数，$\lambda J_T(\theta)$表示调和系数λ乘以教师和学生模型预测差距损失函数$J_T(\theta)$；

其中，在训练过程中，采用训练信号退火TSA算法，线性地释放标注数据中的训练信号。

8.一种基于开放域知识增强的事件检测深度学习模型的构建装置，其特征在于，包括：

获取模块，用于获取标注数据和未标注数据；其中，所述标注数据指标记有触发词的句子数据；所述未标注数据指未标记有触发词的句子数据；所述未标注数据包括第一数据子集和第二数据子集；

第一训练模块，用于将所述标注数据输入到第一事件分类模型中，以对所述第一事件分类模型进行训练；

语义分析模块，用于根据外部语义库采用义原映射算法，对所述未标注数据中的第一数据子集进行语义分析处理，以识别所述第一数据子集中的触发词，得到开放域触发词识别结果；

第二训练模块，用于根据所述开放域触发词识别结果和所述第二数据子集，采用知识蒸馏的方式，对第二事件分类模型进行训练；

第三训练模块，用于对训练后的第一事件分类模型和训练后的第二事件分类模型进行联合训练，得到基于开放域知识增强的事件检测深度学习模型。

9.一种电子设备，包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序，其特征在于，所述处理器执行所述程序时实现如权利要求1至7任一项所述的基于开放域知识增强的事件检测深度学习模型的构建方法的步骤。

10.一种非暂态计算机可读存储介质，其上存储有计算机程序，其特征在于，该计算机程序被处理器执行时实现如权利要求1至7任一项所述的基于开放域知识增强的事件检测深度学习模型的构建方法的步骤。

事件检测模型的构建方法、装置、电子设备及存储介质

技术领域

[0001] 本发明涉及计算机技术领域，尤其涉及一种基于开放域知识增强的事件检测深度学习模型的构建方法、装置、电子设备及存储介质。

背景技术

[0002] 事件检测旨在从非结构化新闻报道中发现事件，目前，事件检测已经作为人工智能领域的一项基础核心技术，被广泛引入到阅读理解以及文本摘要任务。

[0003] 事件检测任务分为两步，第一步检测句子中的触发词，第二步将触发词分类为预定义的事件类型。现有的工作大都专注于第二步事件类型分类，例如提出了动态卷积网络、层级注意力机制等模型。然而，触发词识别同样非常的关键。触发词识别存在长尾问题，即训练样例只集中于少数类型上，剩余的大量的类只有极少的训练样例。以基准数据集ACE2005为例，频率小于5的触发词达到总数的78％以上。针对长尾问题，若采用有监督方法仅依赖标注语料进行训练，则容易过度拟合，在未出现/标注稀疏的触发器上表现不佳；若采用自迭代方法根据伪标签拓展训练实例，则会由于种子集数据本身分布不均，拓展数据集同样集中在多标注的触发词上，无法缓解长尾问题；若采用远程监督方法是依赖外部知识库扩展更多的数据，受限于知识库本身的领域局限以及覆盖率低的问题，无法缓解长尾问题。

发明内容

[0004] 针对现有技术中存在的问题，本发明实施例提供一种基于开放域知识增强的事件检测深度学习模型的构建方法、装置、电子设备及存储介质。

[0005] 第一方面，本发明实施例提供一种基于开放域知识增强的事件检测深度学习模型的构建方法，包括：

[0006] 获取标注数据和未标注数据；其中，所述标注数据指标记有触发词的句子数据；所述未标注数据指未标记有触发词的句子数据；所述未标注数据包括第一数据子集和第二数据子集；

[0007] 将所述标注数据输入到第一事件分类模型中，以对所述第一事件分类模型进行训练；

[0008] 根据外部语义库采用义原映射算法，对所述未标注数据中的第一数据子集进行语义分析处理，以识别所述第一数据子集中的触发词，得到开放域触发词识别结果；

[0009] 根据所述开放域触发词识别结果和所述第二数据子集，采用知识蒸馏的方式，对第二事件分类模型进行训练；

[0010] 对训练后的第一事件分类模型和训练后的第二事件分类模型进行联合训练，得到基于开放域知识增强的事件检测深度学习模型。

[0011] 进一步地，根据外部语义库采用义原映射算法，对所述未标注数据中的第一数据子集进行语义分析处理，以识别所述第一数据子集中的触发词，得到开放域触发词识别结

果，包括：

[0012] 基于外部语义库WordNet对第一数据子集进行词语消歧，将第一数据子集中的词对应到WordNet中单一语义的义原集中；

[0013] 根据第一数据子集中每个词所属的义原集是否触发事件识别所述第一数据子集每个词是否为触发词，以得到开放域触发词识别结果。

[0014] 进一步地，所述第二事件分类模型包括学生模型和教师模型；

[0015] 相应地，根据所述开放域触发词识别结果和所述第二数据子集，采用知识蒸馏的方式，对第二事件分类模型进行训练，包括：

[0016] 以拥有开放域触发词识别结果的第一数据子集作为教师模型的输入，以所述第二数据子集作为学生模型的输入，以教师模型和学生模型的预测结果相同为训练目标，对所述教师模型和学生模型进行训练。

[0017] 进一步地，以拥有开放域触发词识别结果的第一数据子集作为教师模型的输入，以所述第二数据子集作为学生模型的输入，以教师模型和学生模型的预测结果相同为训练目标，对所述教师模型和学生模型进行训练，包括：

[0018] 设定训练目标：

[0019] $p(Y|S^{+},\theta)=p(Y|S^{-},\theta)$

[0020] 其中，$p(Y|S^{+},\theta)$与$p(Y|S^{-},\theta)$分别为教师模型和学生模型的预测结果；其中，教师模型的输入S^{+}是标记有开放域触发词知识的第一数据子集，学生模型的输入S^{-}则是未标记开放域触发词知识的第二数据子集；其中，θ表示教师模型和学生模型共享的参数群，Y表示事件类型预测结果，其中，S^{+}的构造过程包括：引入两个符号B-TRI和E-TRI，标记开放域触发词在句子中开始位置和结束位置，B-TRI表示开始位置，E-TRI表示结束位置，给定原始句子$S=\langle w_1,w_2,\cdots,w_n\rangle$以及开放域触发词$w_i$，编码进开放域触发词的句子表示为$S^{+}=\langle w_1,w_2,\cdots,\text{B-TRI},w_i,\text{E-TRI},\cdots,w_n\rangle$；其中，$S^{-}$的构造过程包括：通过随机屏蔽开放域触发词的事件性词语，扰乱学生模型的输入，给定原始句子$S=\langle w_1,w_2,\cdots,w_n\rangle$以及开放域触发词$w_i$，构造$S^{-}=\{w_1,w_2,\cdots,[\text{MASK}],\cdots,w_n\}$；其中，[MASK]表示随机屏蔽了一部分开放域触发词；

[0021] 通过将句子S^{+}以及S^{-}分别输入教师模型和学生模型，获得教师模型和学生模型两者的预测结果$p(Y|S^{+},\theta)$和$p(Y|S^{-},\theta)$；

[0022] 若未标注数据为$U=\{S_i\}|_{i=1}^{N_U}$，则第二事件分类模型的优化函数为：

$$J_T(\theta)=KL(p(Y|S^{+},\theta)||p(Y|S^{-},\theta)$$

[0023]

$$=\sum_i^{N_U}p\left(Y_{(i)}\middle|S_{(i)}^{+},\theta\right)\frac{p\left(Y_{(i)}\middle|S_{(i)}^{+},\theta\right)}{p\left(Y_{(i)}\middle|S_{(i)}^{-},\theta\right)};$$

[0024] 其中，$J_T(\theta)$表示衡量教师和学生模型预测分布差距的损失函数，KL表示信息增益散度，||表示分布相比运算符，N_U表示未标注数据的规模，$p\left(Y_{(i)}\middle|S_{(i)}^{+},\theta\right)$表示教师模型的预测分布，$p\left(Y_{(i)}\middle|S_{(i)}^{-},\theta\right)$表示学生模型的预测分布。

[0025] 进一步地，采用屏蔽词填写任务，利用周围的单词学习引入符号B-TRI和E-TRI的语义表示。

[0026] 进一步地，将所述标注数据输入到第一事件分类模型中，以对所述第一事件分类模型进行训练，包括：

[0027] 若$S_i=\langle w_1,w_2,\cdots,w_n\rangle$和$Y_i=\langle v_1,v_2,\cdots,v_n\rangle$分别表示第i个训练句子及其事件类型；

[0028] 通过全连接层将句子的隐含表示H转化为中间表示O，以将表示维度和事件个数对齐来执行计算预测概率$O_{ijc}=\delta(WH+b)$的步骤；

[0029] 其中，W和b是全连接层的参数，随机初始化并在训练过程中不断优化，O_{ijc}表示S_i中的第j个单词属于第c个事件类别的概率；

[0030] 通过softmax函数归一化O，以获得条件概率；

[0031] $$p(Y_i|S_i,\theta)=\sum_{j=1}^{n}\frac{\exp(O_{ijc})}{\sum_{c=1}^{C}\exp(O_{ijc})}/n$$

[0032] 若标注数据为$L=\{S_i,Y_i|_{i=1}^{N_L}\}$，则第一事件分类模型的优化函数为：

[0033] $$J_L(\theta)=-\sum_{i=1}^{N_L}\log p(Y_i|S_i,\theta)$$

[0034] 其中，$p(Y_i|S_i,\theta)$表示事件各分类上的条件概率，n表示句子中的词数，$\exp(O_{ijc})$表示第i个训练句子中第j个词语在第c个事件类型上归一化的概率，C表示事件类型数，$J_L(\theta)$表示事件分类的损失函数，N_L表示标注数据的个数。

[0035] 进一步地，对训练后的第一事件分类模型和训练后的第二事件分类模型进行联合训练，得到基于开放域知识增强的事件检测深度学习模型，包括：

[0036] 建立待进行联合训练的基于开放域知识增强的事件检测深度学习模型：

[0037] $J(\theta)=J_L(\theta)+\lambda J_T(\theta)$

[0038] 对上述模型进行联合训练，并在计算J_T时，停止教师模型的梯度下降，以确保学习是从教师模型到学生模型的；其中，$J(\theta)$表示总体的损失函数，λ表示调和系数，$\lambda J_T(\theta)$表示调和系数λ乘以教师和学生模型预测差距损失函数$J_T(\theta)$；

[0039] 其中，在训练过程中，采用训练信号退火TSA算法，线性地释放标注数据中的训练信号。

[0040] 第二方面，本发明实施例提供了一种基于开放域知识增强的事件检测深度学习模型的构建装置，包括：

[0041] 获取模块，用于获取标注数据和未标注数据；其中，所述标注数据指标记有触发词的句子数据；所述未标注数据指未标记有触发词的句子数据；所述未标注数据包括第一数据子集和第二数据子集；

[0042] 第一训练模块，用于将所述标注数据输入到第一事件分类模型中，以对所述第一事件分类模型进行训练；

[0043] 语义分析模块，用于根据外部语义库采用义原映射算法，对所述未标注数据中的第一数据子集进行语义分析处理，以识别所述第一数据子集中的触发词，得到开放域触发词识别结果；

[0044] 第二训练模块，用于根据所述开放域触发词识别结果和所述第二数据子集，采用

知识蒸馏的方式，对第二事件分类模型进行训练；

[0045] 第三训练模块，用于对训练后的第一事件分类模型和训练后的第二事件分类模型进行联合训练，得到基于开放域知识增强的事件检测深度学习模型。

[0046] 第三方面，本发明实施例还提供了一种电子设备，包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序，所述处理器执行所述程序时实现如上第一方面所述的文字识别模型训练方法的步骤。

[0047] 第四方面，本发明实施例还提供了一种非暂态计算机可读存储介质，其上存储有计算机程序，该计算机程序被处理器执行时实现如上第一方面所述的文字识别模型训练方法的步骤。

[0048] 由上述技术方案可知，本发明实施例提供的基于开放域知识增强的事件检测深度学习模型的构建方法、装置、电子设备及存储介质，通过获取标注数据和海量未标注数据，从而扩大训练实例，避免训练样例只集中在有标注的语料中，摆脱“标注数据”知识库受限于本身的领域局限，进而使知识库的覆盖率更高；本实施例根据外部语义库采用义原映射算法得到开放域触发词识别结果，从而可以通过开放域触发词知识去解决触发词识别中的长尾问题，开放域触发词能够从语义角度给出哪些单词可以做触发事件，从而不受预定义事件类型的约束以及文本域的限制，开放域触发词知识能够从未标注的大规模语料中发现从未出现或者出现极少的触发词，这将极大的改善已标注语料中各类触发词分布不均衡的问题；本发明实施例利用开放域触发词知识采用知识蒸馏的方式，能够有效地从已标注的和未标注的语料库中提取开放域触发知识，从而增强事件检测性能；本发明实施例经过联合训练得到基于开放域知识增强的事件检测深度学习模型，可以在事件检测中提供优质的结构化知识信息，能够指导我们的智能模型具备更深层的事物理解、更精准的任务查询以及一定程度上的逻辑推理能力，从而对海量的信息分析、情报获取时起到至关重要的作用。

附图说明

[0049] 为了更清楚地说明本发明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，下面描述中的附图是本发明的一些实施例，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。

[0050] 图1为本发明一实施例提供的基于开放域知识增强的事件检测深度学习模型的构建方法的流程示意图；

[0051] 图2为本发明一实施例提供的基于开放域知识增强的事件检测深度学习模型的构建方法的检测框架结构示意图；

[0052] 图3为本发明一实施例提供的联合训练的结构示意图；

[0053] 图4为本发明一实施例提供的基于开放域知识增强的事件检测深度学习模型的构建装置的结构示意图；

[0054] 图5为本发明一实施例中电子设备的实体结构示意图。

具体实施方式

[0055] 为使本发明实施例的目的、技术方案和优点更加清楚，下面将结合本发明实施例

中的附图，对本发明实施例中的技术方案进行清楚地描述，显然，所描述的实施例是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，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，都属于本发明保护的范围。

[0056] 图1为本发明一实施例提供的基于开放域知识增强的事件检测深度学习模型的构建方法的流程示意图；如图1所示，该方法包括：

[0057] 步骤101：获取标注数据和未标注数据；其中，所述标注数据指标记有触发词的句子数据；所述未标注数据指未标记有触发词的句子数据；所述未标注数据包括第一数据子集和第二数据子集。

[0058] 在本步骤中，标注数据指标记有触发词的句子数据，举例来说“美军正在向伊拉克军队开火”，此句中将“开火”标记为触发词，为标记有触发词的句子数据。

[0059] 在本步骤中，未标注数据指未标记有触发词的句子数据，举例来说“A man was hacked to death by the criminal”，此句中“hacked”是一个罕见词，为未标记有触发词的句子数据。

[0060] 本步骤通过获取标注数据和海量未标注数据，从而扩大训练实例，避免训练样例只集中在有标记的标注数据中，摆脱“有标记的标注数据”知识库受限于本身的领域局限，进而使知识库的覆盖率更高。

[0061] 步骤102：将所述标注数据输入到第一事件分类模型中，以对所述第一事件分类模型进行训练。

[0062] 在本步骤中，举例来说，标注数据为“Troops were trying to break up stone-throwing protests,but not use live fire”，其中“fire”为标记触发词，将上述句子输入到第一事件分类模型经特征编码器进行训练，分析得出事件预测结果为触发了“攻击”事件。

[0063] 在本步骤中，需要说明的是，

[0064] 步骤103：根据外部语义库采用义原映射算法，对所述未标注数据中的第一数据子集进行语义分析处理，以识别所述第一数据子集中的触发词，得到开放域触发词识别结果。

[0065] 在本步骤中，需要说明的是，“义原映射算法”使用外部语义库作为基础，例如外部语义库WordNet。

[0066] 在本步骤中，举例来说，利用外部语义库WordNet对句子中的词汇进行语义消歧，即进行语义分析处理；设计映射算法将无歧义的词汇原语中具有事件性的词汇标记出来，即识别出触发词，得到开放域触发词识别结果。

[0067] 在本步骤中引入开放域触发词知识去解决触发词识别中的长尾问题，由于开放域触发知识是一种先验知识，它能够从语义角度给出哪些单词可以做触发事件，而不受预定义事件类型的约束以及文本域的限制。例如，在训练语料中，从未出现“开火”触发事件的情况，但是开放域触发词知识告诉我们从语义角度“开火”是一个事件触发词，有了这个先验知识，我们的模型就能够极大的提高触发词的召回率。开放域触发词知识能够从未标注的大规模语料中发现从未出现或者出现极少的触发词，这将极大的改善已标注语料中各类触发词分布不均衡的问题，从而解决触发词识别中的长尾问题。

[0068] 本步骤提供的“义原映射算法”不仅能够发现足够多的开放域事件触发词，而且具有很高的效率，可以应用于大规模的知识收集。

[0069] 步骤104:根据所述开放域触发词识别结果和所述第二数据子集,采用知识蒸馏的方式,对第二事件分类模型进行训练。

[0070] 在本步骤中,首先获取知识,将开放域触发词识别结果和第二数据子集,采用知识蒸馏的方式,知识蒸馏到第二事件分类模型中进行训练,完成知识融合。

[0071] 本步骤提出的模型利用开放域触发词知识,将开放域触发词知识由输入端蒸馏到模型的参数中,进行训练完成知识融合,能够有效地从已标注的和未标注的语料库中提取开放域触发知识,从而增强事件检测性能。

[0072] 步骤105:对训练后的第一事件分类模型和训练后的第二事件分类模型进行联合训练,得到基于开放域知识增强的事件检测深度学习模型。

[0073] 在本步骤中,例如对训练后的第一事件分类模型和训练后的第二事件分类模型采用权重因子来综合两种模型,并用梯度下降算法进行联合学习,得到基于开放域知识增强的事件检测深度学习模型。

[0074] 由上面技术方案可知,本发明实施例提供的基于开放域知识增强的事件检测深度学习模型的构建方法,通过获取标注数据和海量未标注数据,从而扩大训练实例,避免训练样例只集中在有标注的语料中,摆脱“标注数据”知识库受限于本身的领域局限,进而使知识库的覆盖率更高;本实施例根据外部语义库采用义原映射算法得到开放域触发词识别结果,从而可以通过开放域触发词知识去解决触发词识别中的长尾问题,开放域触发词能够从语义角度给出哪些单词可以做触发事件,从而不受预定义事件类型的约束以及文本域的限制,开放域触发词知识能够从未标注的大规模语料中发现从未出现或者出现极少的触发词,这将极大的改善已标注语料中各类触发词分布不均衡的问题;本发明实施例利用开放域触发词知识采用知识蒸馏的方式,能够有效地从已标注的和未标注的语料库中提取开放域触发知识,从而增强事件检测性能;本发明实施例经过联合训练得到基于开放域知识增强的事件检测深度学习模型,可以在事件检测中提供优质的结构化知识信息,能够指导我们的智能模型具备更深层的事物理解、更精准的任务查询以及一定程度上的逻辑推理能力,从而对海量的信息分析、情报获取时起到至关重要的作用。

[0075] 在上述实施例的基础上,在本实施例中,根据外部语义库采用义原映射算法,对所述未标注数据中的第一数据子集进行语义分析处理,以识别所述第一数据子集中的触发词,得到开放域触发词识别结果,包括:

[0076] 基于外部语义库WordNet对第一数据子集进行词语消歧,将第一数据子集中的词对应到WordNet中单一语义的义原集中;

[0077] 根据第一数据子集中每个词所属的义原集是否触发事件识别所述第一数据子集每个词是否为触发词,以得到开放域触发词识别结果。

[0078] 在本实施例中,需要说明的是,义原映射算法为一种轻量级管道方法,义原映射算法使用外部语义库WordNet作为基础,以收集开放域触发知识,它有两个步骤,第一步,对词语消歧,将词对应到WordNet中单一语义的义原中。具体来说,例如我们首先通过斯坦福的自然语言处理工具(Stanford CoreNLP)获取句子中的词性标注和句法解析标注,利用这些句法标注作为特征输入,采用语言模型算法将句子中词消歧到WordNet中的义原集中。第二步,我们根据当前词所从属的义原集是否触发事件,来判定当前词是否触发事件。由此可见,获取知识时,我们设计了一个分步算法Tigger from WordNet(TFW),利用WordNet语义

库获得开放域触发词知识。

[0079] 在本实施例中，需要说明的是，开放域触发词知识可以从语义端，告诉我们句子中可以作为事件触发词的词语有哪些，该知识不限定于特定的领域，是我们做事件检测，尤其是做长尾低频词事件检测时一个重要的外部知识。

[0080] 由上面技术方案可知，本发明实施例提供的基于开放域知识增强的事件检测深度学习模型的构建方法，在词汇数据库的支持下，凭借外部语义资源WordNet，从已标注数据和大规模未标注数据中获取开放域触发词知识，能够有效地在各类数据分布不均情景下增强事件触发词的识别能力；我们提出的义原映射算法不仅能够发现足够多的开放域事件触发词，而且具有很高的效率，可以应用于大规模的知识收集。

[0081] 在上述实施例的基础上，在本实施例中，所述第二事件分类模型包括学生模型和教师模型；

[0082] 相应地，根据所述开放域触发词识别结果和所述第二数据子集，采用知识蒸馏的方式，对第二事件分类模型进行训练，包括：

[0083] 以拥有开放域触发词识别结果的第一数据子集作为教师模型的输入，以所述第二数据子集作为学生模型的输入，以教师模型和学生模型的预测结果相同为训练目标，对所述教师模型和学生模型进行训练。

[0084] 在本实施例中，采用知识蒸馏方式的目的是将开放域触发词识别结果蒸馏到第二事件分类模型中，进行知识融合得到包含开放域触发词识别结果的参数，利用该参数进行训练，可以提升事件检测结果。

[0085] 由上面技术方案可知，本发明实施例提供的基于开放域知识增强的事件检测深度学习模型的构建方法，在第二事件分类模型中设计了学生模型和教师模型，通过以拥有开放域触发词识别结果的第一数据子集作为教师模型的输入，以所述第二数据子集作为学生模型的输入，以教师模型和学生模型的预测结果相同为训练目标，对所述教师模型和学生模型进行训练，从而实现让没有“开放域触发词知识”协助的学生模型达到有“开放域触发词知识”增强的教师模型的分类能力，从而把开放域触发词知识融入到模型的参数中，通过以教师模型和学生模型的预测结果相同为训练目标，即使用学生模型来模仿教师模型的决策，完成知识融合，通过知识蒸馏将开放域触发词知识融合到模型的参数中辅助事件检测的决策，提升事件检测在各类标注分布不均的长尾场景下的表现，提升事件检测结果。

[0086] 由上面技术方案可知，本发明实施例提供的基于开放域知识增强的事件检测深度学习模型的构建方法，通过采用知识蒸馏的方式将开放域触发词识别结果和第二数据子集融合，由于模型中包含开放域触发知识，从而不局限于已有的标注语料，而能利用伪标签从海量的无标注语料中进一步提升事件检测能力，从而能够有效的避免过度拟合的问题。

[0087] 在上述实施例的基础上，在本实施例中，以拥有开放域触发词识别结果的第一数据子集作为教师模型的输入，以所述第二数据子集作为学生模型的输入，以教师模型和学生模型的预测结果相同为训练目标，对所述教师模型和学生模型进行训练，包括：

[0088] 设定训练目标：

[0089] $p(Y|S^{+},\theta)=p(Y|S^{-},\theta)$

[0090] 其中，$p(Y|S^{+},\theta)$与$p(Y|S^{-},\theta)$分别为教师模型和学生模型的预测结果；其中，教师模型的输入S^{+}是标记有开放域触发词知识的第一数据子集，学生模型的输入S^{-}则是未标记

开放域触发词知识的第二数据子集;其中,θ表示教师模型和学生模型共享的参数群,Y表示事件类型预测结果,其中,S^+的构造过程包括:引入两个符号B-TRI和E-TRI,标记开放域触发词在句子中开始位置和结束位置,B-TRI表示开始位置,E-TRI表示结束位置,给定原始句子$S=<w_1,w_2,\cdots,w_n>$以及开放域触发词w_i,编码进开放域触发词的句子表示为$S^+=<w_1,w_2,\cdots,$B-TRI$,w_i,$E-TRI$,\cdots,w_n>$;其中,S^-的构造过程包括:通过随机屏蔽开放域触发词的事件性词语,扰乱学生模型的输入,给定原始句子$S=<w_1,w_2,\cdots,w_n>$以及开放域触发词w_i,构造$S^-=\{w_1,w_2,\cdots,$[MASK]$,\cdots,w_n\}$;其中,[MASK]表示随机屏蔽了一部分开放域触发词;

[0091] 通过将句子S^+以及S^-分别输入教师模型和学生模型,获得教师模型和学生模型两者的预测结果$p(Y|S^+,\theta)$和$p(Y|S^-,\theta)$;

[0092] 若未标注数据为$U=\{S_i\}|_{i=1}^{N_U}$,则第二事件分类模型的优化函数为:

[0093] $J_T(\theta)=KL(p(Y|S^+,\theta)||p(Y|S^-,\theta)$

[0094] $$=\sum_i^{N_U} p\left(Y_{(i)}\middle|S_{(i)}^+,\theta\right)\frac{p\left(Y_{(i)}\middle|S_{(i)}^+,\theta\right)}{p\left(Y_{(i)}\middle|S_{(i)}^-,\theta\right)};$$

[0095] 其中,$J_T(\theta)$表示衡量教师和学生模型预测分布差距的损失函数,KL表示信息增益散度,||表示分布相比运算符,N_U表示未标注数据的规模,$p\left(Y_{(i)}\middle|S_{(i)}^+,\theta\right)$表示教师模型的预测分布,$p\left(Y_{(i)}\middle|S_{(i)}^-,\theta\right)$表示学生模型的预测分布。

[0096] 为了更好的理解本实施例,举例来说:

[0097] 教师模型以拥有开放域触发词知识的文本作为输入,学生模型以原始文本作为输入,通过迫使学生模型在未标记的数据上生成与教师模型一样好的伪标签,使得模型能够将开放域触发词知识从输入端蒸馏到模型参数中。

[0098] 给定的学习目标为:

[0099] $p(Y|S^+,\theta)=p(Y|S^-,\theta)$

[0100] 其中$p(Y|S^+,\theta)$与$p(Y|S^-,\theta)$分别为教师模型和学生模型的预测结果。可以看到教师模型的输入S^+拥有开放域触发知识,而学生模型的输入S^-则不拥有开放域触发知识。

[0101] 下面给出S^+和S^-的详细构造过程:

[0102] 1)拥有开放域触发词知识的句子(S^+)

[0103] 采用了“标记机制”将收集好的开放域触发知识编码进输入端。具体来说,引入了两个符号:B-TRI和E-TRI,来标记开放域触发词在句子中开始和结束的位置。

[0104] 给定原始文本$S=<w_1,w_2,\cdots,w_n>$以及开放域触发知识识别的事件性词语w_i,编码进开放域触发词知识的句子表示为$S^+=<w_1,w_2,\cdots,$B-TRI$,w_i,$E-TRI$,\cdots,w_n>$。需要说明的是,“标记机制”对于特征提取器BERT很好用,并且在嵌入知识方面非常灵活,无需进行大量的工程工作即可方便地适应其他类型的知识。

[0105] “标记机制”会面临一个问题:新添加的符号缺少在BERT中的预训练嵌入。随机初始化会导致引入符号的语义无法表达,其中B-TRI符号需要表达触发词起始位置的语义,而E-TRI需要表达触发词结束位置的语义。通过对特征提取器进行微调来解决此问题。基于哈里斯分布假设,采用屏蔽词填写(Mask LM)任务,利用周围的单词来学习的引入符号(B-TRI和E-TRI)的语义表示。屏蔽字率设置为0.15,微调后,屏蔽字的填写准确度达到92.3%。

[0106] 2)不拥有开放域触发知识的句子(S-)

[0107] 为了增加学生模型的学习难度,通过随机屏蔽开放域触发器知识识别的事件性词语,进一步扰乱了学生模型的输入。通过扰乱输入,学生模型必须要在关键词缺失的情境下,仍然能够根据周围语境来判断触发词的事件类型。给定原始句子$S=\langle w_1,w_2,\cdots,w_n\rangle$以及开放域触发词知识识别的事件性词语$w_i$,构造的$S^-=\{w_1,w_2,\cdots,[MASK],\cdots,w_n\}$。

[0108] 3)一致性训练

[0109] 通过将句子S^+以及S^-分别输入句子编码模块,获得教师模型和学生模型两者的预测结果$p(Y|S^+,\theta)$和$p(Y|S^-,\theta)$。为了保证词汇之间的严格对齐,在计算KL散度之前,我们将插入的标记B-TRI,E-TRI重排到句尾。

[0110] 给定海量的未标注数据$U=\{S_i\}|_{i=1}^{N_U}$,知识蒸馏的优化函数定义为:

$$J_T(\theta)=KL(p(Y|S^+,\theta)||p(Y|S^-,\theta)$$

[0111]
$$=\sum_i^{N_U} p\left(Y_{(i)}|S_{(i)}^+,\theta\right)\frac{p\left(Y_{(i)}|S_{(i)}^+,\theta\right)}{p\left(Y_{(i)}|S_{(i)}^-,\theta\right)}$$

[0112] 需要说明的是KL散度是一个非对称的指标,可以让学生模型预测分布尽可能的接近教师模型预测分布,而不是相反,从而将开放域知识提炼到模型参数中。

[0113] 在本发明实施例中,需要说明的是,开放域触发知识从词义的角度阐述了单词是否触发事件。有了这个外部知识,我们训练的模型就可以避免对标注数据的过度依赖,从而提升模型在未被标记或者稀疏标记的触发词上的识别性能。例如,在“A man was hacked to death by the criminal”中,“hacked”是一个罕见词,从未在标记语料中出现过,如果仅从标记数据出发进行有监督的学习,模型很容易把它分成负例,即认为它不是一个事件触发词。但有了开放域触发词知识之后,我们可以知道“hacked”具有攻击的语义,应该会触发攻击事件,从而提升触发词识别的召回率。换句话说,开放域触发词知识作为一个常识,能够引导模型的注意力,使其关注到和事件有关的重点词上。

[0114] 在本发明实施例中,需要说明的是,本发明实施例提出了丰富知识蒸馏Enrichment Knowledge Distillation(EKD)模型来利用开放域触发词知识。该模型能够有效地从已标注的和未标注的语料库中提取开放域触发知识,从而增强事件检测性能。丰富知识蒸馏模型分为两步:获取知识和融合知识。获取知识时,我们设计了一个分步算法Tigger from WordNet(TFW),利用WordNet语义库获得开放域触发词知识。融合知识时,我们设计了一个拥有开放域触发词知识的教师模型以及一个不拥有开放域触发词知识的学生模型。通过使用学生模型来模仿教师模型的决策,将开放域触发词知识由输入端蒸馏到模型的参数当中,完成知识的融合。

[0115] 由上面技术方案可知,本发明实施例提供的基于开放域知识增强的事件检测深度学习模型的构建方法,以拥有开放域触发词识别结果的第一数据子集作为教师模型的输入,以所述第二数据子集作为学生模型的输入,从而可以从标注和未标注的数据中学习,从而通过减少注释中的内置偏差来提高事件检测的性能;同时,本发明实施例,不仅限于利用开放域触发词知识,还可以方便地迁移到提炼其他知识上,例如命名实体知识,句法结构知识和事件角色知识。

[0116] 本发明实施例提出的模型利用了开放域触发词知识这样的语义信息，指导模型从大规模未标注的语料中发现未标注/标注稀疏的触发词，从而解决了标注语料中各类标注分布不均匀的长尾难题。除此之外，本发明实施例提出的模型可以灵活的迁移到提炼句法或者实体知识上，能够辅助更多的自然语言处理任务，有很广阔的应用前景。

[0117] 在上述实施例的基础上，在本实施例中，采用屏蔽词填写任务，利用周围的单词学习引入符号B-TRI和E-TRI的语义表示。

[0118] 在上述实施例的基础上，在本实施例中，将所述标注数据输入到第一事件分类模型中，以对所述第一事件分类模型进行训练，包括：

[0119] 若$S_i=\langle w_1,w_2,\cdots,w_n\rangle$和$Y_i=\langle v_1,v_2,\cdots,v_n\rangle$分别表示第i个训练句子及其事件类型；

[0120] 通过全连接层将句子的隐含表示H转化为中间表示O，以将表示维度和事件个数对齐来执行计算预测概率$O_{ijc}=\delta(WH+b)$的步骤；

[0121] 其中，W和b是全连接层的参数，随机初始化并在训练过程中不断优化，O_{ijc}表示S_i中的第j个单词属于第c个事件类别的概率；

[0122] 通过softmax函数归一化O，以获得条件概率；

[0123] $$p(Y_i|S_i,\theta)=\sum_{j=1}^{n}\frac{\exp(O_{ijc})}{\sum_{c=1}^{C}\exp(O_{ijc})}/n$$

[0124] 若标注数据为$L=\{S_i,Y_i|_{i=1}^{N_L}\}$，则第一事件分类模型的优化函数为：

[0125] $$J_L(\theta)=-\sum_{i=1}^{N_L}\log p(Y_i|S_i,\theta)$$

[0126] 其中，$p(Y_i|S_i,\theta)$表示事件各分类上的条件概率，n表示句子中的词数，$\exp(O_{ijc})$表示第i个训练句子中第j个词语在第c个事件类型上归一化的概率，C表示事件类型数，$J_L(\theta)$表示事件分类的损失函数，N_L表示标注数据的个数。

[0127] 为了更好地理解本实施例，举例来说：

[0128] $S_i=\langle w_1,w_2,\cdots,w_n\rangle$和$Y_i=\langle v_1,v_2,\cdots,v_n\rangle$分别表示第i个训练句子及其事件类型。我们首先通过全连接层将句子的隐含表示H转化为中间表示O，这一步是将表示维度和事件个数对齐，方便后续求预测概率

[0129] $O_{ijc}=\delta(WH+b)$

[0130] 其中W和b是全连接层的参数，随机初始化并在训练过程中不断优化，O_{ijc}表示S_i中的第j个单词属于第c个事件类别的概率。之后，我们通过softmax函数归一化O，以获得条件概率。

[0131] $$p(Y_i|S_i,\theta)=\sum_{j=1}^{n}\frac{\exp(O_{ijc})}{\sum_{c=1}^{C}\exp(O_{ijc})}/n$$

[0132] 给定标注数据$L=\{S_i,Y_i|_{i=1}^{N_L}\}$，有监督的优化函数定义为：

[0133] $$J_L(\theta) = -\sum_{i=1}^{N_L} \log p(Y_i|S_i,\theta)$$

[0134] 由上面技术方案可知，本发明实施例提供的基于开放域知识增强的事件检测深度学习模型的构建方法，能够引导模型的预测结果和标注结果一致。

[0135] 在上述实施例的基础上，在本实施例中，对训练后的第一事件分类模型和训练后的第二事件分类模型进行联合训练，得到基于开放域知识增强的事件检测深度学习模型，包括：

[0136] 建立待进行联合训练的基于开放域知识增强的事件检测深度学习模型：

[0137] $J(\theta) = J_L(\theta) + \lambda J_T(\theta)$

[0138] 对上述模型进行联合训练，并在计算J_T时，停止教师模型的梯度下降，以确保学习是从教师模型到学生模型的；其中，$J(\theta)$表示总体的损失函数，λ表示调和系数，$\lambda J_T(\theta)$表示调和系数λ乘以教师和学生模型预测差距损失函数$J_T(\theta)$；

[0139] 其中，在训练过程中，采用训练信号退火TSA算法，线性地释放标注数据中的训练信号。

[0140] 由上述技术方案可知，在计算J_T时，通过停止教师模型的梯度下降，从而确保学习是从教师模型到学生模型的，从而使事件预测结果更加准确。

[0141] 在本实施例中，由于未标记数据比标记数据的规模大很多，联合训练会使模型已经完全拟合标记数据时，仍不足以拟合未标记数据。为了解决这个问题，我们采用训练信号退火(TSA)技术，来线性地释放标注数据中的训练信号，解决了联合训练仍不足以拟合未标记数据的问题。

[0142] 为了更好的理解本发明实施例，下面结合图2和图3进一步阐述发明实施例的内容，但本发明不仅仅局限于本发明实施例。

[0143] 图2为本发明一实施例提供的基于开放域知识增强的事件检测深度学习模型的构建方法的检测框架结构示意图，将标注数据经过句子编码输入第一事件分类模型进行训练；将未标注数据分为第一数据子集和第二数据子集，其中第一数据通过知识收集得到开放域触发词识别结果，将开放域触发词识别结果和第二数据子集经句子编码，采用知识蒸馏的方式一并输入到第二事件分类模型进行训练；参见图3，将训练后的第一事件分类模型和训练后的第二事件分类模型进行联合训练，输出事件检测结果。

[0144] 在本发明实施例中，需要说明的是，句子编码：对于输入的句子，该模块通过多头注意力交互机制以及多层的交互编码，获得句子中各个词的嵌入表示，即将离散的词语映射到可计算的隐含表示空间。具体的，例如采用BERT模型来获得标记和未标记句子的隐藏表示。BERT是一种经过预先训练的语言表示模型，它采用的多层堆叠的多头注意力机制(multi-layer multi-head attention mechanism)，该机制不仅有更好的并行性，而且能够有效的提升神经网络对句子整体语义的表示能力，近年来其强大的表示能力在许多任务(例如问题回答和语言推断)上都取得了优异的性能。事件检测场景中也证明了BERT的强大功能。

[0145] 举例来说，输入句子$S=\langle w_1, w_2, \cdots, w_n\rangle$到BERT中，采用最后一层的序列输出作为S中每个单词隐含表示$H=\langle h1, h2, \ldots, hn\rangle$。

[0146] H＝BERT(S)。

[0147] 图4为本发明一实施例提供的基于开放域知识增强的事件检测深度学习模型的构建装置的结构示意图，如图4所示，该装置包括：获取模块201、第一训练模块202、语义分析模块203、第二训练模块204，第三训练模块205，其中：

[0148] 其中，获取模块201，用于获取标注数据和未标注数据；其中，所述标注数据指标记有触发词的句子数据；所述未标注数据指未标记有触发词的句子数据；所述未标注数据包括第一数据子集和第二数据子集；

[0149] 第一训练模块202，用于将所述标注数据输入到第一事件分类模型中，以对所述第一事件分类模型进行训练；

[0150] 语义分析模块203，用于根据外部语义库采用义原映射算法，对所述未标注数据中的第一数据子集进行语义分析处理，以识别所述第一数据子集中的触发词，得到开放域触发词识别结果；

[0151] 第二训练模块204，用于根据所述开放域触发词识别结果和所述第二数据子集，采用知识蒸馏的方式，对第二事件分类模型进行训练；

[0152] 第三训练模块205，用于对训练后的第一事件分类模型和训练后的第二事件分类模型进行联合训练，得到基于开放域知识增强的事件检测深度学习模型。

[0153] 本发明实施例提供的基于开放域知识增强的事件检测深度学习模型的构建装置具体可以用于执行上述实施例的基于开放域知识增强的事件检测深度学习模型的构建方法，其技术原理和有益效果类似，具体可参见上述实施例，此处不再赘述。

[0154] 基于相同的发明构思，本发明实施例提供一种电子设备，参见图5，电子设备具体包括如下内容：处理器310、通信接口320、存储器330和通信总线340；

[0155] 其中，处理器310、通信接口320、存储器330通过总线340完成相互间的通信；通信接口320用于实现各建模软件及智能制造装备模块库等相关设备之间的信息传输；处理310用于调用存储器330中的计算机程序，处理器执行计算机程序时实现上述各方法实施例所提供的方法，例如，处理器执行计算机程序时实现下述步骤：获取标注数据和未标注数据；其中，所述标注数据指标记有触发词的句子数据；所述未标注数据指未标记有触发词的句子数据；所述未标注数据包括第一数据子集和第二数据子集；将所述标注数据输入到第一事件分类模型中，以对所述第一事件分类模型进行训练；根据外部语义库采用义原映射算法，对所述未标注数据中的第一数据子集进行语义分析处理，以识别所述第一数据子集中的触发词，得到开放域触发词识别结果；根据所述开放域触发词识别结果和所述第二数据子集，采用知识蒸馏的方式，对第二事件分类模型进行训练；对训练后的第一事件分类模型和训练后的第二事件分类模型进行联合训练，得到基于开放域知识增强的事件检测深度学习模型。

[0156] 基于相同的发明构思，本发明又一实施例还提供一种非暂态计算机可读存储介质，其上存储有计算机程序，该计算机程序被处理器执行时实现以执行上述各方法实施例提供的方法，例如，获取标注数据和未标注数据；其中，所述标注数据指标记有触发词的句子数据；所述未标注数据指未标记有触发词的句子数据；所述未标注数据包括第一数据子集和第二数据子集；将所述标注数据输入到第一事件分类模型中，以对所述第一事件分类模型进行训练；根据外部语义库采用义原映射算法，对所述未标注数据中的第一数据子集

进行语义分析处理，以识别所述第一数据子集中的触发词，得到开放域触发词识别结果；根据所述开放域触发词识别结果和所述第二数据子集，采用知识蒸馏的方式，对第二事件分类模型进行训练；对训练后的第一事件分类模型和训练后的第二事件分类模型进行联合训练，得到基于开放域知识增强的事件检测深度学习模型。

[0157] 以上所描述的装置实施例仅仅是示意性的，其中作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部模块来实现本实施例方案的目的。本领域普通技术人员在不付出创造性的劳动的情况下，即可以理解并实施。

[0158] 通过以上的实施方式的描述，本领域的技术人员可以清楚地了解到各实施方式可借助软件加必需的通用硬件平台的方式来实现，当然也可以通过硬件。基于这样的理解，上述技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来，该计算机软件产品可以存储在计算机可读存储介质中，如ROM/RAM、磁碟、光盘等，包括若干指令用以使得一台计算机设备(可以是个人计算机，服务器，或者网络设备等)执行各个实施例或者实施例的某些部分的方法。

[0159] 此外，在本发明中，诸如“第一”、“第二”仅用于描述目的，而不能理解为指示或暗示相对重要性或者隐含指明所指示的技术特征的数量。由此，限定有“第一”、“第二”的特征可以明示或者隐含地包括至少一个该特征。在本发明的描述中，“多个”的含义是至少两个，例如两个，三个等，除非另有明确具体的限定。

[0160] 此外，在本发明中，诸如第一和第二等之类的关系术语仅仅用来将一个实体或者操作与另一个实体或操作区分开来，而不一定要求或者暗示这些实体或操作之间存在任何这种实际的关系或者顺序。而且，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法、物品或者设备不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、方法、物品或者设备所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括要素的过程、方法、物品或者设备中还存在另外的相同要素。

[0161] 此外，在本说明书的描述中，参考术语“一个实施例”、“一些实施例”、“示例”、“具体示例”、或“一些示例”等的描述意指结合该实施例或示例描述的具体特征、结构、材料或者特点包含于本发明的至少一个实施例或示例中。在本说明书中，对上述术语的示意性表述不必须针对的是相同的实施例或示例。而且，描述的具体特征、结构、材料或者特点可以在任一个或多个实施例或示例中以合适的方式结合。此外，在不相互矛盾的情况下，本领域的技术人员可以将本说明书中描述的不同实施例或示例以及不同实施例或示例的特征进行结合和组合。

[0162] 最后应说明的是：以上实施例仅用以说明本发明的技术方案，而非对其限制；尽管参照前述实施例对本发明进行了详细的说明，本领域的普通技术人员应当理解：其依然可以对前述各实施例所记载的技术方案进行修改，或者对其中部分技术特征进行等同替换；而这些修改或替换，并不使相应技术方案的本质脱离本发明各实施例技术方案的精神和范围。

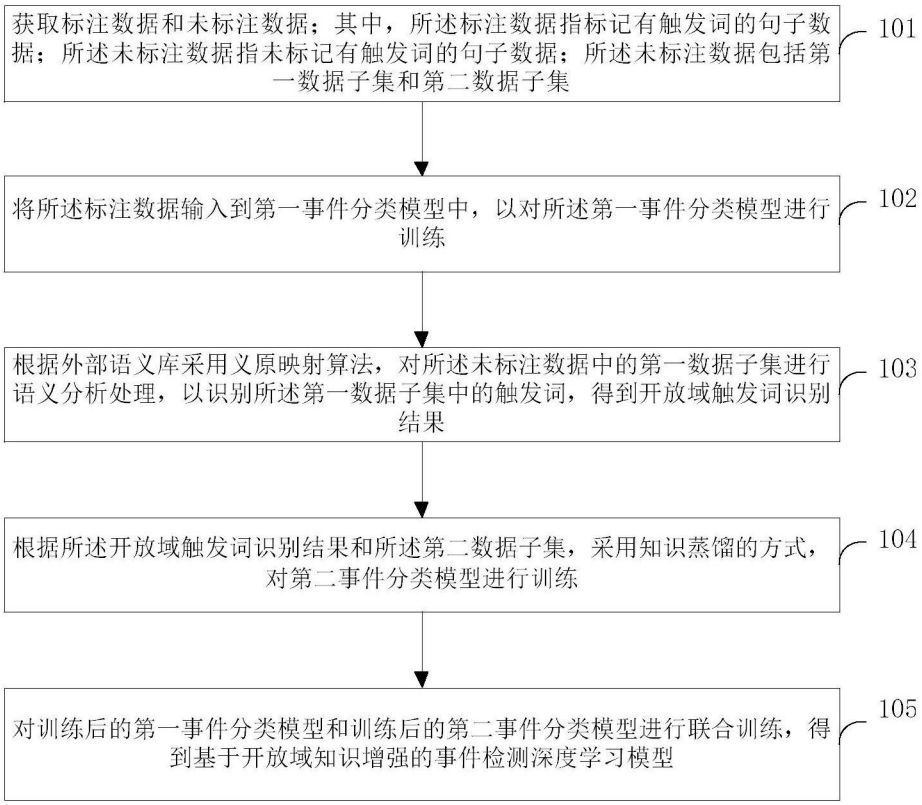

获取标注数据和未标注数据；其中，所述标注数据指标记有触发词的句子数据；所述未标注数据指未标记有触发词的句子数据；所述未标注数据包括第一数据子集和第二数据子集
101
将所述标注数据输入到第一事件分类模型中，以对所述第一事件分类模型进行训练
102
根据外部语义库采用义原映射算法，对所述未标注数据中的第一数据子集进行语义分析处理，以识别所述第一数据子集中的触发词，得到开放域触发词识别结果
103
根据所述开放域触发词识别结果和所述第二数据子集，采用知识蒸馏的方式，对第二事件分类模型进行训练
104
对训练后的第一事件分类模型和训练后的第二事件分类模型进行联合训练，得到基于开放域知识增强的事件检测深度学习模型
105

图1

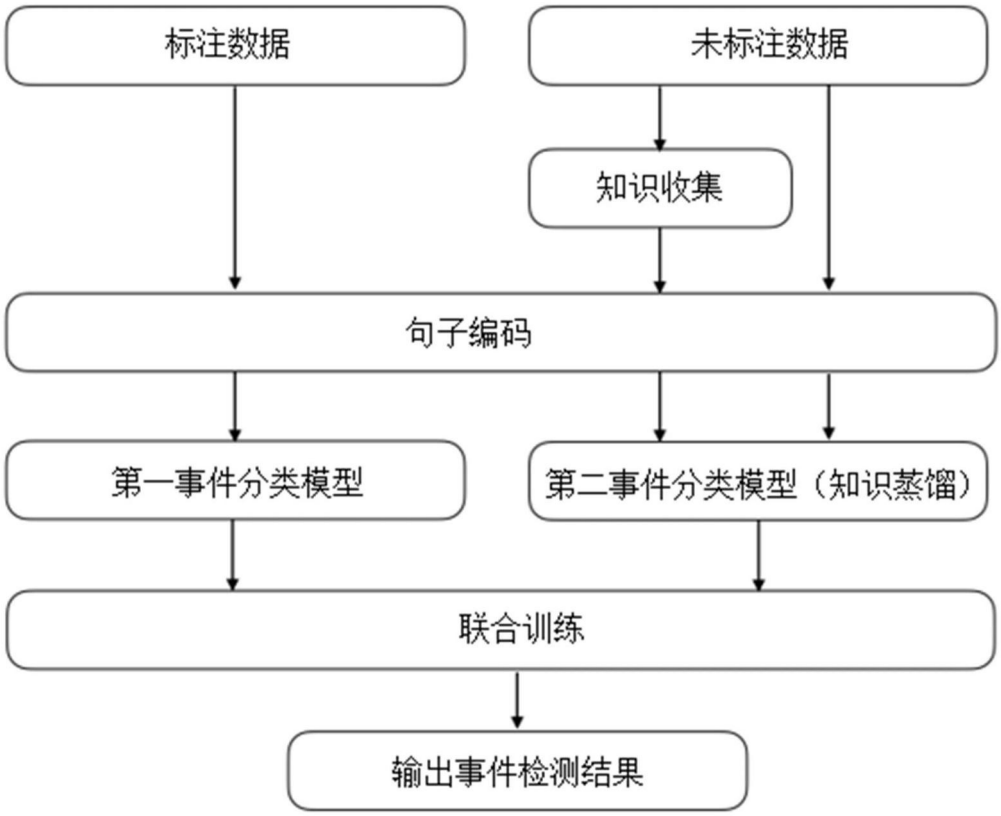

标注数据
未标注数据
知识收集
句子编码
第一事件分类模型
第二事件分类模型（知识蒸馏）
联合训练
输出事件检测结果

图2

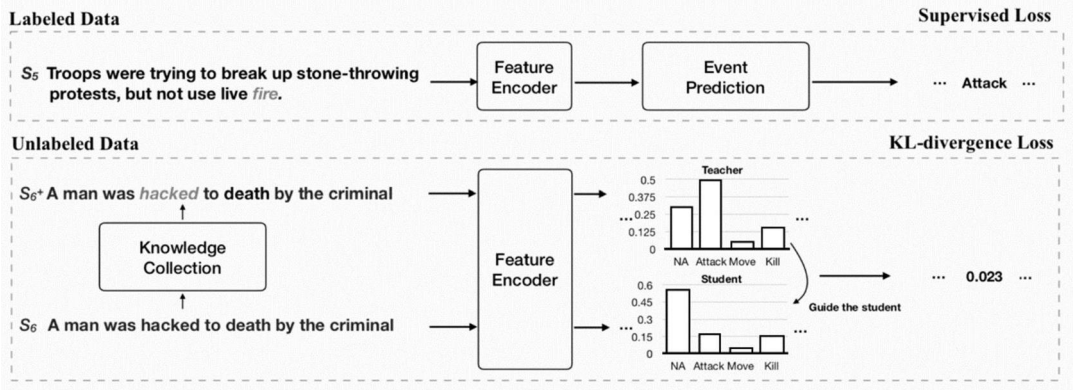

Labeled Data
Supervised Loss
S5 Troops were trying to break up stone-throwing protests, but not use live fire.
Feature Encoder
Event Prediction
... Attack ...
Unlabeled Data
KL-divergence Loss
S6+ A man was hacked to death by the criminal
Knowledge Collection
S6 A man was hacked to death by the criminal
Feature Encoder
Teacher
Student
NA Attack Move Kill
Guide the student
... 0.023 ...

图3


基于开放域知识增强的事件检测深度学习模型的构建装置
获取模块
201
第一训练模块
202
语义分析模块
203
第二训练模块
204
第三训练模块
205

图4

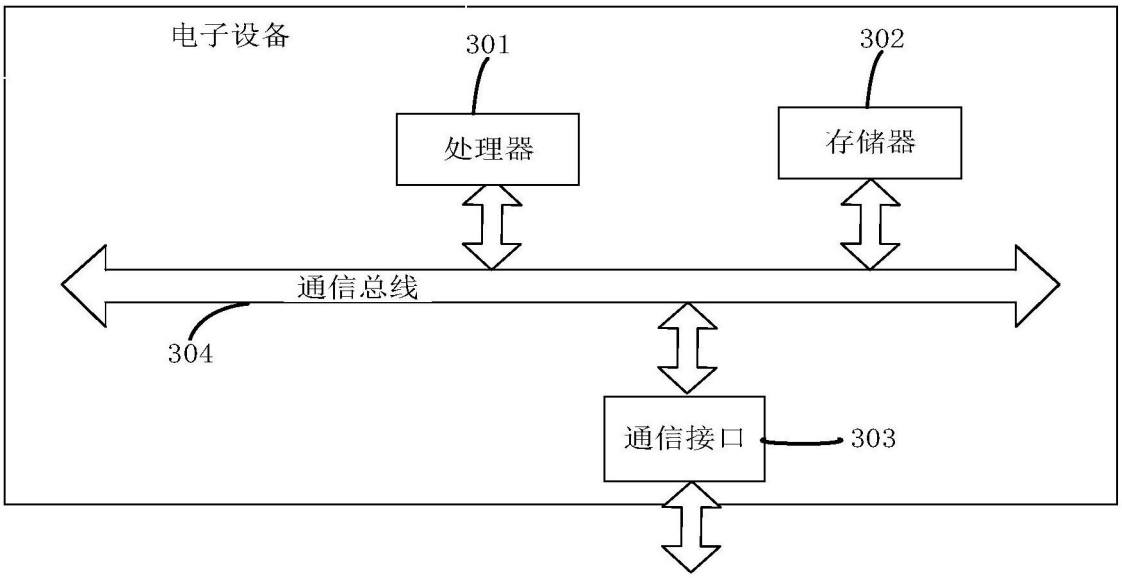

电子设备
301
处理器
302
存储器
通信总线
304
通信接口
303

图5