(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 114818682 A
(43)申请公布日 2022.07.29

(21)申请号 202210749823.9

(22)申请日 2022.06.29

(71)申请人 中国人民解放军国防科技大学
地址 410073 湖南省长沙市开福区德雅路109号

(72)发明人 蒋林承 张俊丰 张维琦 赵超 邓劲生 曾道建 谭真 李硕豪 乔凤才

(74)专利代理机构 长沙国科天河知识产权代理有限公司 43225
专利代理师 邱轶

(51)Int.Cl.
G06F 40/279(2020.01)
G06N 3/04(2006.01)
G06N 3/08(2006.01)

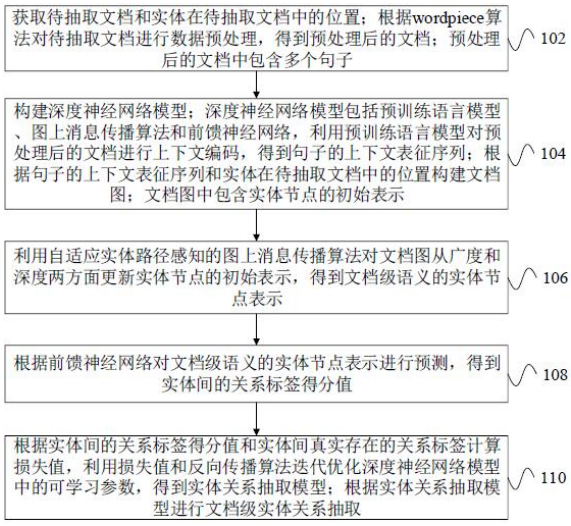
权利要求书3页 说明书15页 附图3页

(54)发明名称

基于自适应实体路径感知的文档级实体关系抽取方法

(57)摘要

本申请涉及一种基于自适应实体路径感知的文档级实体关系抽取方法。所述方法包括：根据句子的上下文表征序列和实体在待抽取文档中的位置构建文档图；利用自适应实体路径感知的图上消息传播算法对文档图从广度和深度两方面更新实体节点的初始表示，得到文档级语义的实体节点表示；根据前馈神经网络对文档级语义的实体节点表示进行预测，得到实体间的关系标签得分值；根据实体间的关系标签得分值和实体间真实存在的关系标签计算损失值，利用损失值和反向传播算法迭代优化深度神经网络模型中的可学习参数，得到实体关系抽取模型；根据实体关系抽取模型进行文档级实体关系抽取。采用本方法能够提高文档级实体关系抽取的准确率。



CN 114818682 A

1.一种基于自适应实体路径感知的文档级实体关系抽取方法,其特征在于,所述方法包括:

获取待抽取文档和实体在所述待抽取文档中的位置;

根据wordpiece算法对所述待抽取文档进行数据预处理,得到预处理后的文档;所述预处理后的文档中包含多个句子;

构建深度神经网络模型;所述深度神经网络模型包括预训练语言模型、自适应实体路径感知的图上消息传播算法和前馈神经网络;

利用预训练语言模型对所述预处理后的文档进行上下文编码,得到句子的上下文表征序列;

根据所述句子的上下文表征序列和实体在所述待抽取文档中的位置构建文档图;所述文档图中包含实体节点的初始表示;

利用自适应实体路径感知的图上消息传播算法对所述文档图从广度和深度两方面更新所述实体节点的初始表示,得到文档级语义的实体节点表示;

根据前馈神经网络对所述文档级语义的实体节点表示进行预测,得到实体间的关系标签得分值;

根据所述实体间的关系标签得分值和实体间真实存在的关系标签计算损失值,利用所述损失值和反向传播算法迭代优化所述深度神经网络模型中的可学习参数,得到实体关系抽取模型;

根据所述实体关系抽取模型进行文档级实体关系抽取。

2.根据权利要求1所述的方法,其特征在于,根据所述句子的上下文表征序列和实体在所述待抽取文档中的位置构建文档图,包括:

根据所述句子的上下文表征序列和实体在所述待抽取文档中的位置计算提及节点、实体节点和句子节点的初始表示,利用提及节点、实体节点和句子节点的初始表示和提及节点、实体节点和句子节点在待抽取文档中的自然关联连接节点构建文档图。

3.根据权利要求2所述的方法,其特征在于,所述提及节点,实体节点和句子节点在待抽取文档中的自然关联包括提及节点与提及节点之间的相互连接、提及节点和句子节点之间的相互连接、提及节点和实体节点之间的相互连接以及实体节点和句子节点之间的相互连接;所述提及节点,实体节点和句子节点构成文档图的节点集合;所述提及节点,实体节点和句子节点在待抽取文档中的自然关联构成文档图的边集合;所述提及节点为提及在文档中对应的词的上下文表征的平均值;所述实体节点为实体对应的所有提及节点表示的平均值;所述句子节点为句子中所有词的上下文表征的平均值。

4.根据权利要求1所述的方法,其特征在于,利用自适应实体路径感知的图上消息传播算法对所述文档图从广度和深度两方面更新所述实体节点的初始表示,得到文档级语义的实体节点表示,包括:

利用自适应实体路径感知的图上消息传播算法聚合所述文档图中目标节点N跳内的邻居信息,建模实体对间的交互,从广度和深度两方面共同控制消息传播算法,通过在所述文档图上自动学习实体相关的自适应路径来筛选和聚合文档级信息,得到文档级语义的实体节点表示。

5.根据权利要求4所述的方法,其特征在于,利用自适应实体路径感知的图上消息传播

算法聚合所述文档图中目标节点N跳内的邻居信息，建模实体对间的交互，从广度和深度两方面共同控制消息传播算法，通过在所述文档图上自动学习实体相关的自适应路径来筛选和聚合文档级信息，得到文档级语义的实体节点表示，包括：

利用自适应实体路径感知的图上消息传播算法聚合所述文档图中目标节点N跳内的邻居信息，建模实体对间的交互，对于广度方面，在每一跳邻居信息聚合过程中，根据广度自适应方式得到节点的广度临时聚合表示；

在深度方面根据LSTM的长短记忆网络，利用多个门控机制对所述节点的广度临时聚合表示有选择的保存节点相关的文档级高阶信息，只选择一定跳数内邻居进行传播，得到文档级语义的实体节点表示。

6.根据权利要求5所述的方法，其特征在于，根据广度自适应方式得到节点的广度临时聚合，包括：

根据广度自适应方式得到节点的广度临时聚合为

$$V_u^{tmp} = FFN(W_r \sum_{v \in N_{(u)}} \alpha_{(u,v)} V_u^l)$$

其中，$\alpha_{(u,v)} = \frac{exp[QV_v^l(KV_u^l)^T]}{\sum_{v' \in N_{(u)}} exp[QV_{v'}^l(KV_u^l)^T]}$，$\alpha_{(u,v)}$指节点u和邻居v的权重参数，$\mathrm{W_r}$是对邻居特征进行线性转换的可学习参数，$V_u^l$表示节点$\mathrm{u}$的表示，$\mathrm{Q}$和$\mathrm{K}$指注意力机制中的query和key矩阵，$\mathrm{FFN}(\cdot)$指一个前馈神经网络，$N_{(u)}$是节点$\mathrm{u}$的邻居节点集合，$V_{v'}^l$表示节点$v'$的表示，$v'$表示节点$\mathrm{u}$的邻居节点，$T$表示转置运算。

7.根据权利要求5所述的方法，其特征在于，在深度方面根据LSTM的长短记忆网络，利用多个门控机制对所述节点的广度临时聚合表示有选择的保存节点相关的文档级高阶信息，只选择一定跳数内邻居进行传播，得到文档级语义的实体节点表示，包括：

对于所述节点的广度临时聚合表示，利用更新门将节点的广度临时聚合表示中的有效信息添加到记忆单元中，遗忘门则过滤掉上一层记忆单元中无效的信息，输出门控制记忆单元，输出文档级语义的实体节点表示。

8.根据权利要求1所述的方法，其特征在于，根据前馈神经网络对所述文档级语义的实体节点表示进行预测，得到实体间的关系标签得分值，包括：

根据前馈神经网络对所述文档级语义的实体节点表示进行预测，得到实体间的所有关系标签的得分为

$$logits = W_b\sigma(W_a\mathrm{r}_{(\mathrm{h,t})} + b_a) + b_b$$

其中$W_a \in \mathbb{R}^{2d*d}$，$W_b \in \mathbb{R}^{d*k}$，$b_a$和$b_b$是前馈神经网络中的分类器的可学习参数，$\sigma$指激活函数，d指前馈神经网络中的隐藏维度，k是标签的数量，$\mathrm{r}_{(\mathrm{h,t})}$表示不同的实体节点表示$\mathrm{e_h}$和$\mathrm{e_t}$拼接得到实体对的特征。

9.根据权利要求8所述的方法，其特征在于，根据所述实体间的关系标签得分值和实体

间真实存在的关系标签计算损失值，包括：

根据所述实体间的关系标签得分值和实体间真实存在的关系标签计算损失值为

$$L = -\sum_{r \in P_T} log\left(\frac{exp(logits_r)}{\sum_{r' \in P_T \cup \{TH\}} exp(logits_{r'})}\right) - log\left(\frac{exp(logits_{TH})}{\sum_{r' \in N_T \cup \{TH\}} exp(logits_{r'})}\right)$$

其中，TH表示阈值关系标签TH，P_T表示实体间真实存在的关系标签集合，N_T表示负样本关系标签集合，logits指实体对(e_h, e_t)中所有关系标签的得分，$logits_r$指关系标签r的得分值，r'表示关系标签，$logits_{r'}$表示关系标签r'的得分值，$logits_{TH}$表示阈值关系标签TH的得分值。

基于自适应实体路径感知的文档级实体关系抽取方法

技术领域

[0001] 本申请涉及数据处理技术领域，特别是涉及一种基于自适应实体路径感知的文档级实体关系抽取方法、装置、计算机设备和存储介质。

背景技术

[0002] 实体关系抽取是信息抽取领域的一个经典任务，其任务旨在识别给定非结构化文本中所包含的实体(概念)之间的语义关系，并将结果以关系三元组的结构化形式存储。如给定文本“2017年10月，今日头条宣布10亿美金估值收购音乐短视频平台Musical.ly”，通过实体关系抽取得到关系三元组《今日头条，收购，Musical.ly》。实体关系抽取作为信息抽取的关键技术，能够在自然语言处理的多个领域中发挥重要作用，特别是在互联网海量信息的时代背景下具有的重大研究意义和广阔的应用前景。从理论价值层面看，实体关系抽取涉及到机器学习、数据挖掘、自然语言处理等多个学科的理论和方法。从应用层面看，实体关系抽取可用于大规模知识库尤其是知识图谱的自动构建，为信息检索和自动问答系统的构建提供数据支持，也是自然语言理解的基础。现有的实体关系抽取工作主要聚焦于句子级抽取，局限于单句文本中的实体语义关系。然而，在真实应用场景中，实体语义关系的描述非常复杂，大量的实体间关系是通过多个句子表达的，并表现出多个实体之间复杂的关联性。根据从维基百科采样的人工标注数据的统计表明，至少40％的实体语义关系事实只能从多个句子中联合获取。因此，有必要将实体关系抽取推进至更符合真实场景的文档级别。相比于句子级实体关系抽取，文档级实体关系抽取更有挑战性，它需要更加复杂的推理技巧，如逻辑推理，共指推理，常识推理等。一篇文档中可包含多个实体，且每个实体拥有多个处于不同上下文的提及。为了识别出跨越句子的实体之间的关系，需要能够建模文档中多个实体之间的复杂交互以及综合利用实体的多个提及的上下文信息，这显然超出了句子级关系抽取方法的能力范围。

[0003] 目前，随着图神经网络研究的深入，研究者尝试使用文档图来建模文档内各类语义信息，其使用词，提及，实体或句子等作为节点，并利用启发式的规则连接成文档图。这些方法关注于如何构建更好的文档图以保留更多的语义信息及如何更好的在图上进行信息传播。借助于图神经网络强大的表示能力，这类方法取得了不错的效果，但也存在以下一些问题：a) 现有的工作在聚合实体表示时，往往是不加区分的聚合多个提及表示，然后再将提及表示组合成单个全局表示，来进行和其他所有实体的语义关系预测。实际上，由于实体的多个提及在文档中处于不同的上下文中，连接不同类型的节点时，每个节点起到的作用应是不同的。b) 图神经网络通过节点信息传播隐式地进行推理，为了捕获图中高阶信息的交互，往往会叠加使用多层图网络结构(如进行多次图卷积)，图中同一连通分量内的节点的表征会趋向于收敛到一个和输入无关的子空间，造成学习到的节点表示过平滑，不够准确。

发明内容

[0004] 基于此,有必要针对上述技术问题,提供一种能够提高文档级实体关系抽取的准确率的基于自适应实体路径感知的文档级实体关系抽取方法、装置、计算机设备和存储介质。

[0005] 一种基于自适应实体路径感知的文档级实体关系抽取方法,所述方法包括:

获取待抽取文档和实体在待抽取文档中的位置;

根据wordpiece算法对待抽取文档进行数据预处理,得到预处理后的文档;预处理后的文档中包含多个句子;

构建深度神经网络模型;深度神经网络模型包括预训练语言模型、自适应实体路径感知的图上消息传播算法和前馈神经网络;

利用预训练语言模型对预处理后的文档进行上下文编码,得到句子的上下文表征序列;

根据句子的上下文表征序列和实体在待抽取文档中的位置构建文档图;文档图中包含实体节点的初始表示;

利用自适应实体路径感知的图上消息传播算法对文档图从广度和深度两方面更新实体节点的初始表示,得到文档级语义的实体节点表示;

根据前馈神经网络对文档级语义的实体节点表示进行预测,得到实体间的关系标签得分值;

根据实体间的关系标签得分值和实体间真实存在的关系标签计算损失值,利用损失值和反向传播算法迭代优化深度神经网络模型中的可学习参数,得到实体关系抽取模型;

根据实体关系抽取模型进行文档级实体关系抽取。

[0006] 在其中一个实施例中,根据句子的上下文表征序列和实体在待抽取文档中的位置构建文档图,包括:

根据句子的上下文表征序列和实体在待抽取文档中的位置计算提及节点、实体节点和句子节点的初始表示,利用提及节点,实体节点和句子节点的初始表示和提及节点,实体节点和句子节点在待抽取文档中的自然关联连接节点构建文档图。

[0007] 在其中一个实施例中,提及节点,实体节点和句子节点在待抽取文档中的自然关联包括提及节点与提及节点之间的相互连接、提及节点和句子节点之间的相互连接、提及节点和实体节点之间的相互连接以及实体节点和句子节点之间的相互连接;提及节点,实体节点和句子节点构成文档图的节点集合;提及节点,实体节点和句子节点在待抽取文档中的自然关联构成文档图的边集合;提及节点为提及在文档中对应的词的上下文表征的平均值;实体节点为实体对应的所有提及节点表示的平均值;句子节点为句子中所有词的上下文表征的平均值。

[0008] 在其中一个实施例中,利用自适应实体路径感知的图上消息传播算法对文档图从广度和深度两方面更新实体节点的初始表示,得到文档级语义的实体节点表示,包括:

利用自适应实体路径感知的图上消息传播算法聚合文档图中目标节点N跳内的邻居信息,建模实体对间的交互,从广度和深度两方面共同控制消息传播算法,通过在文档图上自动学习实体相关的自适应路径来筛选和聚合文档级信息,得到文档级语义的实体节点

表示。

[0009] 在其中一个实施例中，利用自适应实体路径感知的图上消息传播算法聚合文档图中目标节点N跳内的邻居信息，建模实体对间的交互，从广度和深度两方面共同控制消息传播算法，通过在文档图上自动学习实体相关的自适应路径来筛选和聚合文档级信息，得到文档级语义的实体节点表示，包括：

利用自适应实体路径感知的图上消息传播算法聚合文档图中目标节点N跳内的邻居信息，建模实体对间的交互，对于广度方面，在每一跳邻居信息聚合过程中，根据广度自适应方式得到节点的广度临时聚合表示；

在深度方面根据LSTM的长短记忆网络，利用多个门控机制对节点的广度临时聚合表示有选择的保存节点相关的文档级高阶信息，只选择一定跳数内邻居进行传播，得到文档级语义的实体节点表示。

[0010] 在其中一个实施例中，根据广度自适应方式得到节点的广度临时聚合，包括：

根据广度自适应方式得到节点的广度临时聚合为

$$V_u^{tmp} = FFN(W_r \sum_{v \in N_{(u)}} \alpha_{(u,v)} V_u^l)$$

其中，$\alpha_{(u,v)} = \frac{exp[QV_v^l(KV_u^l)^T]}{\sum_{v' \in N_{(u)}} exp[QV_{v'}^l(KV_u^l)^T]}$，$\alpha_{(u,v)}$指节点u和邻居v的权重参数，$\mathbf{W_r}$是对邻居特征进行线性转换的可学习参数，$V_u^l$表示节点$\mathbf{u}$的表示，$\mathbf{Q}$和$\mathbf{K}$指注意力机制中的query和key矩阵，$\mathbf{FFN(\cdot)}$指一个前馈神经网络，$N_{(u)}$是节点$\mathbf{u}$的邻居节点集合，$V_{v'}^l$表示节点$v'$的表示，$v'$表示节点$\mathbf{u}$的邻居节点，$T$表示转置运算。

[0011] 在其中一个实施例中，在深度方面根据LSTM的长短记忆网络，利用多个门控机制对节点的广度临时聚合表示有选择的保存节点相关的文档级高阶信息，只选择一定跳数内邻居进行传播，得到文档级语义的实体节点表示，包括：

对于节点的广度临时聚合表示，利用更新门将节点的广度临时聚合表示中的有效信息添加到记忆单元中，遗忘门则过滤掉上一层记忆单元中无效的信息，输出门控制记忆单元，输出文档级语义的实体节点表示。

[0012] 在其中一个实施例中，根据前馈神经网络对文档级语义的实体节点表示进行预测，得到实体间的关系标签得分值，包括：

根据前馈神经网络对文档级语义的实体节点表示进行预测，得到实体间的所有关系标签的得分为

$$logits = W_b\sigma(W_a\mathbf{r}_{(h,t)} + b_a) + b_b$$

其中$W_a \in \mathbb{R}^{2d*d}$，$W_b \in \mathbb{R}^{d*k}$，$b_a$和$b_b$是前馈神经网络中的分类器的可学习参数，$\sigma$指激活函数，d指前馈神经网络中的隐藏维度，k是标签的数量，$\mathbf{r}_{(h,t)}$表

示不同的实体节点表示$\mathbf{e_h}$和$\mathbf{e_t}$拼接得到实体对的特征。

[0013] 在其中一个实施例中，根据实体间的关系标签得分值和实体间真实存在的关系标签计算损失值，包括：

根据实体间的关系标签得分值和实体间真实存在的关系标签计算损失值为

$$L = -\sum_{\mathrm{r}\in P_T} log\left(\frac{exp(logits_r)}{\sum_{r'\in P_T\cup\{TH\}} exp(logits_{r'})}\right) - log\left(\frac{exp(logits_{\mathrm{TH}})}{\sum_{r'\in N_T\cup\{TH\}} exp(logits_{r'})}\right)$$

其中，TH表示阈值关系标签TH，P_T表示实体间真实存在的关系标签集合，N_T表示负样本关系标签集合，logits指实体对$(\mathbf{e_h}, \mathbf{e_t})$中所有关系标签的得分，$logits_r$指关系标签$\mathrm{r}$的得分值，$r'$表示关系标签，$logits_{r'}$表示关系标签$r'$的得分值，$logits_{\mathrm{TH}}$表示阈值关系标签$\mathrm{TH}$的得分值。

[0014] 一种基于自适应实体路径感知的文档级实体关系抽取装置，所述装置包括：

数据预处理模块，用于获取待抽取文档和实体在待抽取文档中的位置；根据wordpiece算法对待抽取文档进行数据预处理，得到预处理后的文档；预处理后的文档中包含多个句子；

构建文档图模块，用于构建深度神经网络模型；深度神经网络模型包括预训练语言模型、自适应实体路径感知的图上消息传播算法和前馈神经网络；利用预训练语言模型对预处理后的文档进行上下文编码，得到句子的上下文表征序列；根据句子的上下文表征序列和实体在待抽取文档中的位置构建文档图；文档图中包含实体节点的初始表示；

初始表示更新模块，用于利用自适应实体路径感知的图上消息传播算法对文档图从广度和深度两方面更新实体节点的初始表示，得到文档级语义的实体节点表示；

预测模块，用于根据前馈神经网络对文档级语义的实体节点表示进行预测，得到实体间的关系标签得分值；

文档级实体关系抽取模块，用于根据实体间的关系标签得分值和实体间真实存在的关系标签计算损失值，利用损失值和反向传播算法迭代优化深度神经网络模型中的可学习参数，得到实体关系抽取模型；根据实体关系抽取模型进行文档级实体关系抽取。

[0015] 一种计算机设备，包括存储器和处理器，所述存储器存储有计算机程序，所述处理器执行所述计算机程序时实现以下步骤：

获取待抽取文档和实体在待抽取文档中的位置；

根据wordpiece算法对待抽取文档进行数据预处理，得到预处理后的文档；预处理后的文档中包含多个句子；

构建深度神经网络模型；深度神经网络模型包括预训练语言模型、自适应实体路径感知的图上消息传播算法和前馈神经网络；

利用预训练语言模型对预处理后的文档进行上下文编码，得到句子的上下文表征序列；

根据句子的上下文表征序列和实体在待抽取文档中的位置构建文档图;文档图中包含实体节点的初始表示;

利用自适应实体路径感知的图上消息传播算法对文档图从广度和深度两方面更新实体节点的初始表示,得到文档级语义的实体节点表示;

根据前馈神经网络对文档级语义的实体节点表示进行预测,得到实体间的关系标签得分值;

根据实体间的关系标签得分值和实体间真实存在的关系标签计算损失值,利用损失值和反向传播算法迭代优化深度神经网络模型中的可学习参数,得到实体关系抽取模型;

根据实体关系抽取模型进行文档级实体关系抽取。

[0016] 一种计算机可读存储介质,其上存储有计算机程序,所述计算机程序被处理器执行时实现以下步骤:

获取待抽取文档和实体在待抽取文档中的位置;

根据wordpiece算法对待抽取文档进行数据预处理,得到预处理后的文档;预处理后的文档中包含多个句子;

构建深度神经网络模型;深度神经网络模型包括预训练语言模型、自适应实体路径感知的图上消息传播算法和前馈神经网络;

利用预训练语言模型对预处理后的文档进行上下文编码,得到句子的上下文表征序列;

根据句子的上下文表征序列和实体在待抽取文档中的位置构建文档图;文档图中包含实体节点的初始表示;

利用自适应实体路径感知的图上消息传播算法对文档图从广度和深度两方面更新实体节点的初始表示,得到文档级语义的实体节点表示;

根据前馈神经网络对文档级语义的实体节点表示进行预测,得到实体间的关系标签得分值;

根据实体间的关系标签得分值和实体间真实存在的关系标签计算损失值,利用损失值和反向传播算法迭代优化深度神经网络模型中的可学习参数,得到实体关系抽取模型;

根据实体关系抽取模型进行文档级实体关系抽取。

[0017] 上述基于自适应实体路径感知的文档级实体关系抽取方法、装置、计算机设备和存储介质,本发明采用预训练语言模型对不同层次信息之间的复杂交互进行建模,学习深层语境化的词汇表征,通过构建精细的文档图来建模文档内的语义信息,然后从广度和深度两方面控制消息传播算法,通过学习节点消息传播的自适应感知路径来筛选和聚合文档级信息,有选择性地聚合目标实体的有效文档级信息,解决目前实体关系抽取局限于句内实体关系的问题,也解决了基于文档图的文档级实体关系抽取方法中存在消息传播时不加区分地对待邻居节点和节点表示过平滑的问题,提高了文档级实体关系抽取的性能,实现了对实体语义关系的高效抽取,为大规模知识库构建、信息检索和自动问答系统以及自然语言理解的自然语言处理应用提供了数据支撑和核心算法技术。

附图说明

[0018] 图1为一个实施例中一种基于自适应实体路径感知的文档级实体关系抽取方法的流程示意图；

图2为一个实施例中自适应实体路径感知的示意图；

图3为一个实施例中一种基于自适应实体路径感知的文档级实体关系抽取装置的结构框图；

图4为一个实施例中计算机设备的内部结构图。

具体实施方式

[0019] 为了使本申请的目的、技术方案及优点更加清楚明白，以下结合附图及实施例，对本申请进行进一步详细说明。应当理解，此处描述的具体实施例仅仅用以解释本申请，并不用于限定本申请。

[0020] 在一个实施例中，如图1所示，提供了一种基于自适应实体路径感知的文档级实体关系抽取方法，包括以下步骤：

步骤102，获取待抽取文档和实体在待抽取文档中的位置；根据wordpiece算法对待抽取文档进行数据预处理，得到预处理后的文档；预处理后的文档中包含多个句子。

[0021] 本发明的步骤102中将待抽取文档用$D = \{S_i\}_{i=1}^{N}$表示，文档D由N个句子组成，其中$S_i = \{x_j\}_{j=1}^{M}$指第i句话包含M个词。文档中标注了一个包含P个实体的实体集合$E = \{e_i\}_{i=1}^{P}$，其中$e_i = \{m_j\}_{j=1}^{Q}$指第i个实体在文档中对应Q个共指的实体提及，每一个实体提及出现在不同的上下文中。将文档中以句子为单位，分别输入到wordpiece分词器进行分词，如第i句话分词后为$S_i = \{w_j\}_{j=1}^{k}$，其中k<=M，得到预处理后的文档。将待抽取文档进行预处理后有利于预训练语言模型进行上下文编码。

[0022] 步骤104，构建深度神经网络模型；深度神经网络模型包括预训练语言模型、自适应实体路径感知的图上消息传播算法和前馈神经网络，利用预训练语言模型对预处理后的文档进行上下文编码，得到句子的上下文表征序列；根据句子的上下文表征序列和实体在待抽取文档中的位置构建文档图；文档图中包含实体节点的初始表示。

[0023] 为了更好地建模输入文档的语义，将分词后的预处理后的文档输入至预训练语言模型BERT，通过预训练语言模型BERT将文档分词后的序列映射为包含上下文语义的低维实数向量，其中第i句话对应的输入序列$S_i = \{w_j\}_{j=1}^{k}$映射为上下文表征序列$S_i = \{h_j\}_{j=1}^{k}$，其中$h_j \in \mathbb{R}^{\mathrm{d}}$，d为隐藏维度，一般为768。采用预训练语言模型BERT，BERT能对不同层次信息之间的复杂交互进行建模，学习深层语境化的词汇表征。

[0024] 根据句子的上下文表征序列和实体在待抽取文档中的位置计算提及节点、实体节点和句子节点的初始表示，利用提及节点，实体节点和句子节点的初始表示和提及节点，实体节点和句子节点在待抽取文档中的自然关联连接节点构建文档图，解决了基于文档图的文档级实体关系抽取方法中存在消息传播时不加区分地对待邻居节点和节点表示过平滑

的问题。通过构建精细的文档图来建模文档内的语义信息。自适应实体路径感知的图上消息传播算法是对图上消息传播算法的改进，以往的图上消息传播算法在进行节点聚合时是不做选择，聚合目标节点的所有邻居信息，并且没有从广度和深度对消息对消息传播算法进行控制，本申请提出来的自适应实体路径感知的图上消息传播算法在进行节点聚合时，是聚合文档图中目标节点N跳内的邻居信息，建模实体对间的交互，从广度和深度两方面共同控制消息传播算法，通过在文档图上自动学习实体相关的自适应路径来筛选和聚合文档级信息来得到文档级语义的实体节点表示。

[0025] 步骤106，利用自适应实体路径感知的图上消息传播算法对文档图从广度和深度两方面更新实体节点的初始表示，得到文档级语义的实体节点表示。

[0026] 利用自适应实体路径感知的图上消息传播算法聚合目标节点N跳内的邻居信息，建模实体对间的交互，学习实体在文档图上的自适应路径以提升实体节点的表示，进而得到文档级语义的实体节点表示，通过学习节点消息传播的自适应感知路径来筛选和聚合文档级信息，有选择性地聚合目标实体的有效文档级信息，以捕获更有效的关系语义信息，解决目前实体关系抽取局限于句内实体关系的问题。

[0027] 步骤108，根据前馈神经网络对文档级语义的实体节点表示进行预测，得到实体间的关系标签得分值。

[0028] 实体一般指人名、地名、机构名等专有名词或概念，关系指实体之间的语义联系，比如：“今日头条宣布10亿美金估值收购音乐短视频平台Musical.ly”，通过实体关系抽取得到关系三元组《今日头条，收购，Musical.ly》。

[0029] 为了预测文档级语义的实体节点对之间所包含的语义关系，将文档级语义的实体节点表示中包含的头尾实体表示拼接得到包含语义关系的实体对，根据前馈神经网络对包含语义关系的实体对进行预测，利用预测结果，即实体间的关系标签得分值和实体间真实存在的关系标签计算损失值，有利于进行深度神经网络模型的训练，得到准确的实体关系抽取模型。

[0030] 步骤110，根据实体间的关系标签得分值和实体间真实存在的关系标签计算损失值，利用损失值和反向传播算法迭代优化深度神经网络模型中的可学习参数，得到实体关系抽取模型；根据实体关系抽取模型进行文档级实体关系抽取。

[0031] 根据实体间的关系标签得分值和实体间真实存在的关系标签计算损失值，实体间真实存在的关系标签是人工预先标注的实体间的真实关系标签，随机梯度下降最小化该损失值，根据误差反向传播逐层更新深度神经网络模型中的可学习参数。当优化过程中损失函数收敛后，得到实体关系抽取模型，将实体关系抽取模型保存后可以用于文档级实体关系抽取，本申请中的wordpiece算法和反向传播算法是利用的现有的，并未对算法进行改进。

[0032] 上述基于自适应实体路径感知的文档级实体关系抽取方法中，本发明采用预训练语言模型对不同层次信息之间的复杂交互进行建模，学习深层语境化的词汇表征，通过构建精细的文档图来建模文档内的语义信息，然后从广度和深度两方面控制消息传播算法，通过学习节点消息传播的自适应感知路径来筛选和聚合文档级信息，有选择性地聚合目标实体的有效文档级信息，解决目前实体关系抽取局限于句内实体关系的问题，也解决了基于文档图的文档级实体关系抽取方法中存在消息传播时不加区分地对待邻居节点和节点

表示过平滑的问题，提高了文档级实体关系抽取的性能，实现了对实体语义关系的高效抽取，为大规模知识库构建、信息检索和自动问答系统以及自然语言理解的自然语言处理应用提供了数据支撑和核心算法技术。

[0033] 在其中一个实施例中，根据句子的上下文表征序列和实体在待抽取文档中的位置构建文档图，包括：

根据句子的上下文表征序列和实体在待抽取文档中的位置计算提及节点、实体节点和句子节点的初始表示，利用提及节点，实体节点和句子节点的初始表示和提及节点，实体节点和句子节点在待抽取文档中的自然关联连接节点构建文档图。

[0034] 在其中一个实施例中，提及节点，实体节点和句子节点在待抽取文档中的自然关联包括提及节点与提及节点之间的相互连接、提及节点和句子节点之间的相互连接、提及节点和实体节点之间的相互连接以及实体节点和句子节点之间的相互连接；提及节点，实体节点和句子节点构成文档图的节点集合；提及节点，实体节点和句子节点在待抽取文档中的自然关联构成文档图的边集合；提及节点为提及在文档中对应的词的上下文表征的平均值；实体节点为实体对应的所有提及节点表示的平均值；句子节点为句子中所有词的上下文表征的平均值。

[0035] 在具体实施例中，提及节点旨在表示每个实体在文档中的不同提及。提及节点的表示形式为提及中包含的词对应的隐藏表示的平均值，假设一篇文档中共包含N个提及，则提及节点的表示形式为$N_{m_j} = [avg_{h_i \in e_j}(h_i); t_m], \mathrm{j} = 1,2,\ldots,\mathrm{N}$，其中$t_m$是提及节点的类型嵌入。实体节点与提及节点的形式类似，实体节点的表示形式为实体对应的所有提及表示的平均值，假设一篇文档中包含P个实体，则实体节点的表示形式为$N_{e_j} = [avg_{m_i \in e_j}(m_i); t_e], \mathrm{j} = 1,2,\ldots,\mathrm{P}$，其中$t_e$是实体节点的类型嵌入。句子节点的表示形式为句子序列中所有包含的词对应的隐藏表示的平均值，假设一篇文档中包含T个句子，则句子节点的表示形式为$N_{s_j} = [avg_{h_i \in s_j}(h_i); t_s], \mathrm{j} = 1,2,\ldots,\mathrm{T}$，其中$t_s$是句子节点的类型嵌入。

[0036] 通过上述三种类型的节点构造，得到节点的表示集合$\mathrm{V} \in \mathbb{R}^{(\mathrm{N}+\mathrm{R}+\mathrm{T})*\mathrm{d}}$，其中d是隐藏维度，共计N+R+T个节点。

[0037] 节点构造完成之后，基于文档节点元素之间的自然关联来连接节点构成文档图：a)提及节点-提及节点边：同时共现在同一个句子中的提及之间相互连接。b)提及节点-句子节点边：提及与其所位于的句子相互连接。c)提及节点-实体节点边：提及与其对应的句实体相互连接。d)实体节点-句子节点边：实体与包含其提及的句子相连接。e)句子节点-句子节点边：相互连接所有的句子节点。值得注意的是，图中并没有直接连接两个实体节点，目的是利用下一步骤所述自适应实体路径感知的图上消息传播算法聚合实体节点之间的多跳中间节点建模实体对间的关系。

[0038] 综上，利用上述文档不同节点元素之间的自然关联将构建的N+R+T个提及、实体和句子节点连接成文档图$\mathrm{G} = (\mathrm{V}, \mathrm{E})$，其中V是节点集合，E是边集合。

[0039] 在其中一个实施例中，利用自适应实体路径感知的图上消息传播算法对文档图从

广度和深度两方面更新实体节点的初始表示，得到文档级语义的实体节点表示，包括：

利用自适应实体路径感知的图上消息传播算法聚合文档图中目标节点N跳内的邻居信息，建模实体对间的交互，从广度和深度两方面共同控制消息传播算法，通过在文档图上自动学习实体相关的自适应路径来筛选和聚合文档级信息，得到文档级语义的实体节点表示。

[0040] 在其中一个实施例中，利用自适应实体路径感知的图上消息传播算法聚合文档图中目标节点N跳内的邻居信息，建模实体对间的交互，从广度和深度两方面共同控制消息传播算法，通过在文档图上自动学习实体相关的自适应路径来筛选和聚合文档级信息，得到文档级语义的实体节点表示，包括：

利用自适应实体路径感知的图上消息传播算法聚合文档图中目标节点N跳内的邻居信息，建模实体对间的交互，对于广度方面，在每一跳邻居信息聚合过程中，根据广度自适应方式得到节点的广度临时聚合表示；

在深度方面根据LSTM的长短记忆网络，利用多个门控机制对节点的广度临时聚合表示有选择的保存节点相关的文档级高阶信息，只选择一定跳数内邻居进行传播，得到文档级语义的实体节点表示。

[0041] 在其中一个实施例中，根据广度自适应方式得到节点的广度临时聚合，包括：

根据广度自适应方式得到节点的广度临时聚合为

$$V_u^{tmp} = FFN(W_r \sum_{v \in N_{(u)}} \alpha_{(u,v)} V_u^l)$$

其中，$\alpha_{(u,v)} = \frac{exp[QV_v^l (KV_u^l)^T]}{\sum_{v' \in N_{(u)}} exp[QV_{v'}^l (KV_u^l)^T]}$，$\alpha_{(u,v)}$指节点u和邻居v的权重参数，$\mathbf{W_r}$是对邻居特征进行线性转换的可学习参数，$V_u^l$表示节点$\mathbf{u}$的表示，$\mathbf{Q}$和$\mathbf{K}$指注意力机制中的query和key矩阵，$\mathbf{FFN}(\cdot)$指一个前馈神经网络，$N_{(u)}$是节点$\mathbf{u}$的邻居节点集合，$V_{v'}^l$表示节点$v'$的表示，$v'$表示节点$\mathbf{u}$的邻居节点，$T$表示转置运算。

[0042] 在具体实施例中，对于广度方面，利用多层图注意力网络在每一跳邻居的信息聚合过程中，对于每一个节点在第l+1层的表示V_u^{l+1}，通过如下公式所示的广度自适应方式先得到节点的广度临时聚合表示V_u^{tmp}：

$$V_u^{tmp} = FFN(W_r \sum_{v \in N_{(u)}} \alpha_{(u,v)} V_u^l)$$

$$\alpha_{(u,v)} = \frac{exp[QV_v^l (KV_u^l)^T]}{\sum_{v' \in N_{(u)}} exp[QV_{v'}^l (KV_u^l)^T]}$$

$\alpha_{(u,v)}$通过赋予不同邻居不同权重来区别对待一阶邻居节点。

[0043] 在其中一个实施例中，在深度方面根据LSTM的长短记忆网络，利用多个门控机制对节点的广度临时聚合表示有选择的保存节点相关的文档级高阶信息，只选择一定跳数内邻居进行传播，得到文档级语义的实体节点表示，包括：

对于节点的广度临时聚合表示，利用更新门将节点的广度临时聚合表示中的有效信息添加到记忆单元中，遗忘门则过滤掉上一层记忆单元中无效的信息，输出门控制记忆单元，输出文档级语义的实体节点表示。

[0044] 在具体实施例中，引入LSTM的长短记忆，利用多个门控机制，保存并更新将每一跳邻居信息，有选择的保存节点相关的文档级高阶信息，只选择一定跳数内邻居进行传播，有效防止传播过载和过平滑问题。基于节点的广度临时聚合表示V_u^{tmp}，更新门i_i将新的有效信息添加到记忆单元C^{t+1}中，遗忘门f_i则过滤掉上一层记忆单元C^t中无效的信息，更新门和遗忘门互相配合在探索更远的邻居时可起到选择性抽取与过滤的作用。最后，输出门O_i控制记忆单元C^{t+1}，输出节点i的第$t+1$层的节点表示$v_i{}^{t+1}$。其计算过程如下所示。

[0045] $$f_u = \sigma(W_f{}^t V_u^{tmp}) \quad i_u = \sigma(W_i{}^t V_u^{tmp}) \quad o_u = \sigma(W_o{}^t V_u^{tmp})$$

$$\tilde{C} = tanh(W_c{}^t V_u^{tmp}) \quad C_u^{t+1} = f_u \cdot C_u^t + i_u \cdot \tilde{C} \quad V_u^{l+1} = o_u \cdot tanh(C_u^{t+1})$$

其中$W_f{}^t, W_i{}^t, W_o{}^t, W_c{}^t$分别指遗忘门、更新门、输出门和记忆单元对应的线性转换的可学习参数。

[0046] 如图2所示，该方法通过对每个节点的宽度(哪一跳邻居是重要的)和深度(第t跳邻居的重要性)展开来确定一个合适的子图，可以学习到实体在文档图上消息传播的自适应路径，有选择性地聚合目标实体的有效文档级信息，解决目前实体关系抽取局限于句内实体关系的问题。通过所述消息传播算法从多次迭代，得到一组包含文档级语义的实体节点表示$N_e = \{e_1, e_1, \ldots, e_P\}$。

[0047] 在其中一个实施例中，根据前馈神经网络对文档级语义的实体节点表示进行预测，得到实体间的关系标签得分值，包括：

根据前馈神经网络对文档级语义的实体节点表示进行预测，得到实体间的所有关系标签的得分为

$$logits = W_b \sigma(W_a \mathrm{r}_{(\mathrm{h,t})} + b_a) + b_b$$

其中$W_a \in \mathbb{R}^{2d*d}$，$W_b \in \mathbb{R}^{d*k}$，$b_a$和$b_b$是前馈神经网络中的分类器的可学习参数，$\sigma$指激活函数，d指前馈神经网络中的隐藏维度，k是标签的数量，$\mathrm{r}_{(\mathrm{h,t})}$表示不同的实体节点表示$\mathrm{e_h}$和$\mathrm{e_t}$拼接得到实体对的特征。

[0048] 在具体实施例中，首先，为了预测实体对$(\mathrm{e_h}, \mathrm{e_t})$所包含的语义关系，将实体对包含的头尾实体表示$\mathrm{e_h}$和$\mathrm{e_t}$拼接得到实体对的特征$\mathrm{r}_{(\mathrm{h,t})}$，

$$r_{(h,t)} = [e_h;\ e_t]。$$

[0049] 然后，利用前馈神经网络根据实体对的特征$r_{(h,t)}$计算实体对(e_h, e_t)中所有关系标签的得分$logits$：

$$logits = W_b\sigma(W_a r_{(h,t)} + b_a) + b_b$$

其中$W_a \in \mathbb{R}^{2d*d}$，$W_b \in \mathbb{R}^{d*k}$，$b_a$和$b_b$是前馈神经网络中的分类器的可学习参数，$\sigma$指激活函数，d指前馈神经网络中的隐藏维度，k是标签的数量。

[0050] 在预测阶段，利用非线性激活函数sigmoid进行归一化，可得到实体对(e_h, e_t)中具有关系标签r的概率值，

$$P(r|e_h, e_t) = sigmoid(logits_r)。$$

[0051] 其中$logits_r$指$logits$中关系标签r的得分值。sigmoid是把得分值转成一个0到1内的值，当成是模型给出的关系标签存在于目标实体对的概率，有利于增强关系标签得分的可解释性。因为模型是自己学习的，最后的输出范围不能提前获悉，比如不清楚标签的得分为100是高还是低，使用sigmoid可以把范围压缩在(0-1)以内，这样就知道一般大于0.5的就说明得分属于挺高了。

[0052] 在其中一个实施例中，根据实体间的关系标签得分值和实体间真实存在的关系标签计算损失值，包括：

根据实体间的关系标签得分值和实体间真实存在的关系标签计算损失值为

$$L = -\sum_{r\in P_T} log\left(\frac{exp(logits_r)}{\sum_{r'\in P_T\cup\{TH\}} exp(logits_{r'})}\right) -log\left(\frac{exp(logits_{TH})}{\sum_{r'\in N_T\cup\{TH\}} exp(logits_{r'})}\right)$$

其中，TH表示阈值关系标签TH，P_T表示实体间真实存在的关系标签集合，N_T表示负样本关系标签集合，logits指实体对(e_h, e_t)中所有关系标签的得分，$logits_r$指关系标签r的得分值，r'表示关系标签，$logits_{r'}$表示关系标签r'的得分值，$logits_{TH}$表示阈值关系标签TH的得分值。

[0053] 在具体实施例中，为了更高效的处理多标签问题，即同一实体对可能包含多种关系标签，和样本不平衡问题，即大部分的实体对为负样本，没有包含任何关系标签。本发明采用自适应阈值损失作为损失函数，以端到端的方式优化模型参数。自适应阈值损失引入了一个额外的阈值关系标签TH，其优化目标是实体间真实存在的正样本关系标签集合P_T的得分值高于阈值类标签TH，实体间不存在的负样本关系标签集合N_T的得分值低于阈值

类标签TH，其中，正样本标签指实体间存在的实体间真实存在的关系标签，负样本关系标签指实体间不包含的关系。该损失函数计算公式如下：

$$L = -\sum_{r\in P_T} log\left(\frac{exp(logits_r)}{\sum_{r'\in P_T\cup\{TH\}} exp(logits_{r'})}\right) -log\left(\frac{exp(logits_{TH})}{\sum_{r'\in N_T\cup\{TH\}} exp(logits_{r'})}\right)$$

其中logits指实体对(e_h, e_t)中所有关系标签的得分。为了得到最优的模型参数，本发明通过该损失函数计算实体间的语义关系和实体间真实存在的关系标签之间的损失值，并使用随机梯度下降最小化该损失值L，根据误差反向传播逐层更新模型中的可学习参数。当优化过程中损失函数收敛后，将模型保存后用于文档级实体关系抽取。

[0054] 应该理解的是，虽然图1的流程图中的各个步骤按照箭头的指示依次显示，但是这些步骤并不是必然按照箭头指示的顺序依次执行。除非本文中有明确的说明，这些步骤的执行并没有严格的顺序限制，这些步骤可以以其它的顺序执行。而且，图1中的至少一部分步骤可以包括多个子步骤或者多个阶段，这些子步骤或者阶段并不必然是在同一时刻执行完成，而是可以在不同的时刻执行，这些子步骤或者阶段的执行顺序也不必然是依次进行，而是可以与其它步骤或者其它步骤的子步骤或者阶段的至少一部分轮流或者交替地执行。

[0055] 在一个实施例中，如图3所示，提供了一种基于自适应实体路径感知的文档级实体关系抽取装置，包括：数据预处理模块302、构建文档图模块304、初始表示更新模块306、预测模块308和文档级实体关系抽取模块310，其中：

数据预处理模块302，用于获取待抽取文档和实体在待抽取文档中的位置；根据wordpiece算法对待抽取文档进行数据预处理，得到预处理后的文档；预处理后的文档中包含多个句子；

构建文档图模块304，用于构建深度神经网络模型；深度神经网络模型包括预训练语言模型、自适应实体路径感知的图上消息传播算法和前馈神经网络，利用预训练语言模型对预处理后的文档进行上下文编码，得到句子的上下文表征序列；根据句子的上下文表征序列和实体在待抽取文档中的位置构建文档图；文档图中包含实体节点的初始表示；

初始表示更新模块306，用于利用自适应实体路径感知的图上消息传播算法对文档图从广度和深度两方面更新实体节点的初始表示，得到文档级语义的实体节点表示；

预测模块308，用于根据前馈神经网络对文档级语义的实体节点表示进行预测，得到实体间的关系标签得分值；

文档级实体关系抽取模块310，用于根据实体间的关系标签得分值和实体间真实存在的关系标签计算损失值，利用损失值和反向传播算法迭代优化深度神经网络模型中的可学习参数，得到实体关系抽取模型；根据实体关系抽取模型进行文档级实体关系抽取。

[0056] 在其中一个实施例中，构建文档图模块304还用于根据句子的上下文表征序列和实体在待抽取文档中的位置构建文档图，包括：

根据句子的上下文表征序列和实体在待抽取文档中的位置计算提及节点、实体节点和句子节点的初始表示，利用提及节点，实体节点和句子节点的初始表示和提及节点，实

体节点和句子节点在待抽取文档中的自然关联连接节点构建文档图。

[0057] 在其中一个实施例中，提及节点，实体节点和句子节点在待抽取文档中的自然关联包括提及节点与提及节点之间的相互连接、提及节点和句子节点之间的相互连接、提及节点和实体节点之间的相互连接以及实体节点和句子节点之间的相互连接；提及节点，实体节点和句子节点构成文档图的节点集合；提及节点，实体节点和句子节点在待抽取文档中的自然关联构成文档图的边集合；提及节点为提及在文档中对应的词的上下文表征的平均值；实体节点为实体对应的所有提及节点表示的平均值；句子节点为句子中所有词的上下文表征的平均值。

[0058] 在其中一个实施例中，初始表示更新模块306还用于利用自适应实体路径感知的图上消息传播算法对文档图从广度和深度两方面更新实体节点的初始表示，得到文档级语义的实体节点表示，包括：

利用自适应实体路径感知的图上消息传播算法聚合文档图中目标节点N跳内的邻居信息，建模实体对间的交互，从广度和深度两方面共同控制消息传播算法，通过在文档图上自动学习实体相关的自适应路径来筛选和聚合文档级信息，得到文档级语义的实体节点表示。

[0059] 在其中一个实施例中，初始表示更新模块306还用于利用自适应实体路径感知的图上消息传播算法聚合文档图中目标节点N跳内的邻居信息，建模实体对间的交互，从广度和深度两方面共同控制消息传播算法，通过在文档图上自动学习实体相关的自适应路径来筛选和聚合文档级信息，得到文档级语义的实体节点表示，包括：

利用自适应实体路径感知的图上消息传播算法聚合文档图中目标节点N跳内的邻居信息，建模实体对间的交互，对于广度方面，在每一跳邻居信息聚合过程中，根据广度自适应方式得到节点的广度临时聚合表示；

在深度方面根据LSTM的长短记忆网络，利用多个门控机制对节点的广度临时聚合表示有选择的保存节点相关的文档级高阶信息，只选择一定跳数内邻居进行传播，得到文档级语义的实体节点表示。

[0060] 在其中一个实施例中，初始表示更新模块306还用于根据广度自适应方式得到节点的广度临时聚合，包括：

根据广度自适应方式得到节点的广度临时聚合为

$$V_u^{tmp} = FFN(W_r \sum_{v \in N_{(u)}} \alpha_{(u,v)} V_u^l)$$

其中，$\alpha_{(u,v)} = \frac{exp[QV_v^l(KV_u^l)^T]}{\sum_{v' \in N_{(u)}} exp[QV_{v'}^l(KV_u^l)^T]}$，$\alpha_{(u,v)}$指节点u和邻居v的权重参数，$W_r$是对邻居特征进行线性转换的可学习参数，$V_u^l$表示节点$u$的表示，$Q$和$K$指注意力机制中的query和key矩阵，$FFN(\cdot)$指一个前馈神经网络，$N_{(u)}$是节点$u$的邻居节点集合$V_{v'}^l$表示节点$v'$的表示，$v'$表示节点$u$的邻居节点，$T$表示转置运算。

[0061] 在其中一个实施例中，初始表示更新模块306还用于在深度方面根据LSTM的长短

记忆网络，利用多个门控机制对节点的广度临时聚合表示有选择的保存节点相关的文档级高阶信息，只选择一定跳数内邻居进行传播，得到文档级语义的实体节点表示，包括：

对于节点的广度临时聚合表示，利用更新门将节点的广度临时聚合表示中的有效信息添加到记忆单元中，遗忘门则过滤掉上一层记忆单元中无效的信息，输出门控制记忆单元，输出文档级语义的实体节点表示。

[0062] 在其中一个实施例中，预测模块308还用于根据前馈神经网络对文档级语义的实体节点表示进行预测，得到实体间的关系标签得分值，包括：

根据前馈神经网络对文档级语义的实体节点表示进行预测，得到实体间的所有关系标签的得分为

$$logits = W_b\sigma(W_a\mathrm{r}_{(\mathrm{h,t})} + b_a) + b_b$$

其中$W_a \in \mathbb{R}^{2d*d}$，$W_b \in \mathbb{R}^{d*k}$，$b_a$和$b_b$是前馈神经网络中的分类器的可学习参数，$\sigma$指激活函数，d指前馈神经网络中的隐藏维度，k是标签的数量，$\mathrm{r}_{(\mathrm{h,t})}$表示不同的实体节点表示$\mathrm{e_h}$和$\mathrm{e_t}$拼接得到实体对的特征。

[0063] 在其中一个实施例中，文档级实体关系抽取模块310还用于根据实体间的关系标签得分值和实体间真实存在的关系标签计算损失值，包括：

根据实体间的关系标签得分值和实体间真实存在的关系标签计算损失值为

$$L = -\sum_{\mathrm{r}\in P_T} log\left(\frac{exp(logits_r)}{\sum_{r'\in P_T\cup\{TH\}} exp(logits_{r'})}\right) - log\left(\frac{exp(logits_{\mathrm{TH}})}{\sum_{r'\in N_T\cup\{TH\}} exp(logits_{r'})}\right)$$

其中，TH表示阈值关系标签TH，P_T表示实体间真实存在的关系标签集合，N_T表示负样本关系标签集合，logits指实体对$(\mathrm{e_h}, \mathrm{e_t})$中所有关系标签的得分，$logits_r$指关系标签$\mathrm{r}$的得分值，$r'$表示关系标签，$logits_{r'}$表示关系标签$r'$的得分值，$logits_{\mathrm{TH}}$表示阈值关系标签$\mathrm{TH}$的得分值。

[0064] 关于一种基于自适应实体路径感知的文档级实体关系抽取装置的具体限定可以参见上文中对于一种基于自适应实体路径感知的文档级实体关系抽取方法的限定，在此不再赘述。上述一种基于自适应实体路径感知的文档级实体关系抽取装置中的各个模块可全部或部分通过软件、硬件及其组合来实现。上述各模块可以硬件形式内嵌于或独立于计算机设备中的处理器中，也可以以软件形式存储于计算机设备中的存储器中，以便于处理器调用执行以上各个模块对应的操作。

[0065] 在一个实施例中，提供了一种计算机设备，该计算机设备可以是终端，其内部结构图可以如图4所示。该计算机设备包括通过系统总线连接的处理器、存储器、网络接口、显示屏和输入装置。其中，该计算机设备的处理器用于提供计算和控制能力。该计算机设备的存

储器包括非易失性存储介质、内存储器。该非易失性存储介质存储有操作系统和计算机程序。该内存储器为非易失性存储介质中的操作系统和计算机程序的运行提供环境。该计算机设备的网络接口用于与外部的终端通过网络连接通信。该计算机程序被处理器执行时以实现一种基于自适应实体路径感知的文档级实体关系抽取方法。该计算机设备的显示屏可以是液晶显示屏或者电子墨水显示屏，该计算机设备的输入装置可以是显示屏上覆盖的触摸层，也可以是计算机设备外壳上设置的按键、轨迹球或触控板，还可以是外接的键盘、触控板或鼠标等。

[0066] 本领域技术人员可以理解，图4中示出的结构，仅仅是与本申请方案相关的部分结构的框图，并不构成对本申请方案所应用于其上的计算机设备的限定，具体的计算机设备可以包括比图中所示更多或更少的部件，或者组合某些部件，或者具有不同的部件布置。

[0067] 在一个实施例中，提供了一种计算机设备，包括存储器和处理器，该存储器存储有计算机程序，该处理器执行计算机程序时实现上述实施例中方法的步骤。

[0068] 在一个实施例中，提供了一种计算机存储介质，其上存储有计算机程序，计算机程序被处理器执行时实现上述实施例中方法的步骤。

[0069] 本领域普通技术人员可以理解实现上述实施例方法中的全部或部分流程，是可以通过计算机程序来指令相关的硬件来完成，所述的计算机程序可存储于一非易失性计算机可读取存储介质中，该计算机程序在执行时，可包括如上述各方法的实施例的流程。其中，本申请所提供的各实施例中所使用的对存储器、存储、数据库或其它介质的任何引用，均可包括非易失性和/或易失性存储器。非易失性存储器可包括只读存储器(ROM)、可编程ROM(PROM)、电可编程ROM(EPROM)、电可擦除可编程ROM(EEPROM)或闪存。易失性存储器可包括随机存取存储器(RAM)或者外部高速缓冲存储器。作为说明而非局限，RAM以多种形式可得，诸如静态RAM(SRAM)、动态RAM(DRAM)、同步DRAM(SDRAM)、双数据率SDRAM(DDRSDRAM)、增强型SDRAM(ESDRAM)、同步链路(Synchlink) DRAM(SLDRAM)、存储器总线(Rambus)直接RAM(RDRAM)、直接存储器总线动态RAM(DRDRAM)、以及存储器总线动态RAM(RDRAM)等。

[0070] 以上实施例的各技术特征可以进行任意的组合，为使描述简洁，未对上述实施例中的各个技术特征所有可能的组合都进行描述，然而，只要这些技术特征的组合不存在矛盾，都应当认为是本说明书记载的范围。

[0071] 以上所述实施例仅表达了本申请的几种实施方式，其描述较为具体和详细，但并不能因此而理解为对发明专利范围的限制。应当指出的是，对于本领域的普通技术人员来说，在不脱离本申请构思的前提下，还可以做出若干变形和改进，这些都属于本申请的保护范围。因此，本申请专利的保护范围应以所附权利要求为准。

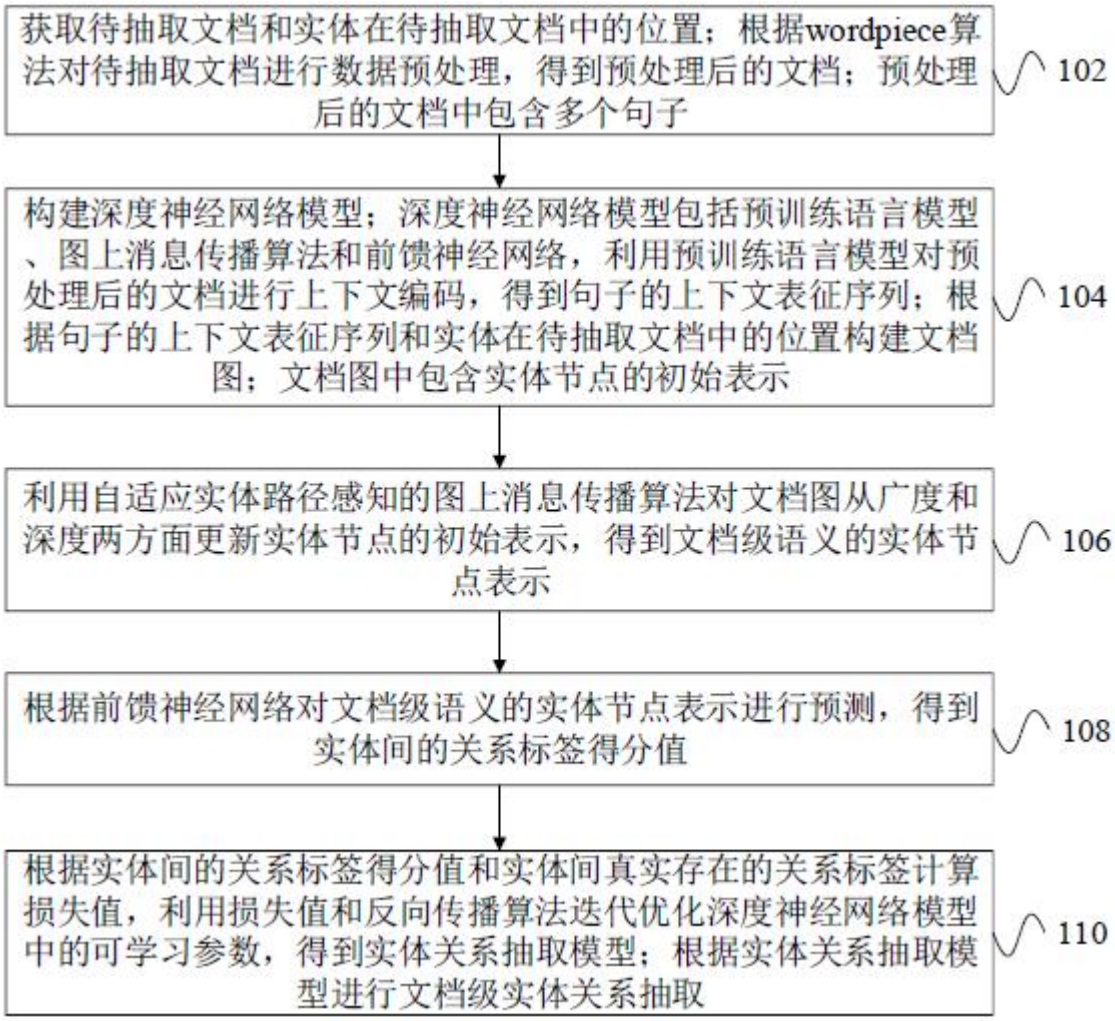

获取待抽取文档和实体在待抽取文档中的位置；根据wordpiece算法对待抽取文档进行数据预处理，得到预处理后的文档；预处理后的文档中包含多个句子
102
构建深度神经网络模型；深度神经网络模型包括预训练语言模型、图上消息传播算法和前馈神经网络，利用预训练语言模型对预处理后的文档进行上下文编码，得到句子的上下文表征序列；根据句子的上下文表征序列和实体在待抽取文档中的位置构建文档图；文档图中包含实体节点的初始表示
104
利用自适应实体路径感知的图上消息传播算法对文档图从广度和深度两方面更新实体节点的初始表示，得到文档级语义的实体节点表示
106
根据前馈神经网络对文档级语义的实体节点表示进行预测，得到实体间的关系标签得分值
108
根据实体间的关系标签得分值和实体间真实存在的关系标签计算损失值，利用损失值和反向传播算法迭代优化深度神经网络模型中的可学习参数，得到实体关系抽取模型；根据实体关系抽取模型进行文档级实体关系抽取
110

图1

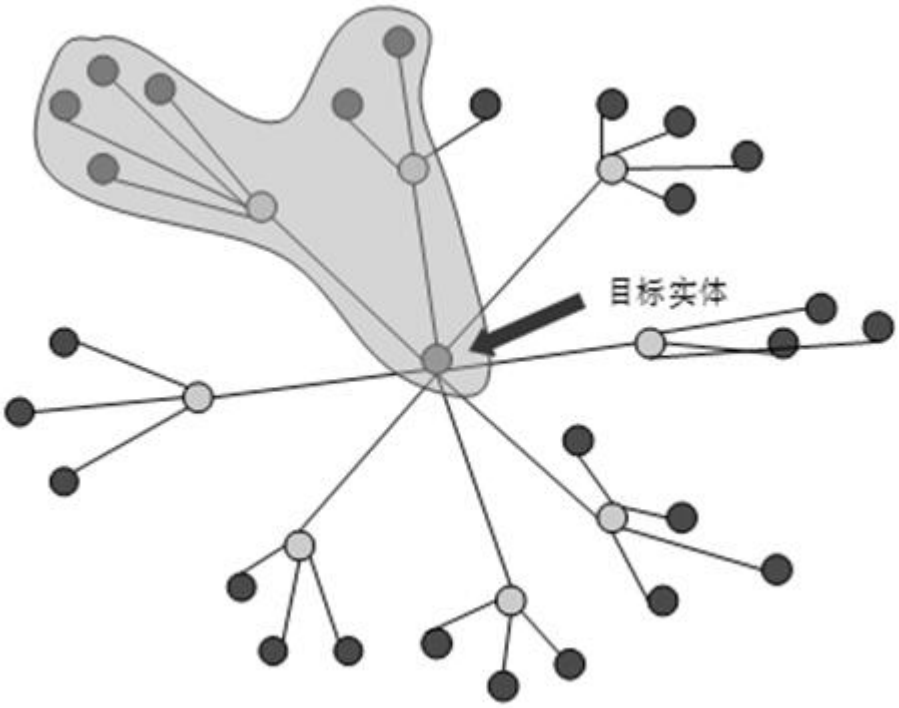

目标实体

图2

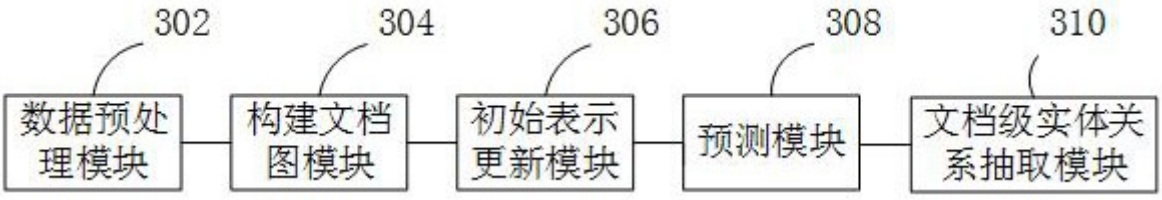

302
304
306
308
310
数据预处理模块
构建文档图模块
初始表示更新模块
预测模块
文档级实体关系抽取模块

图3

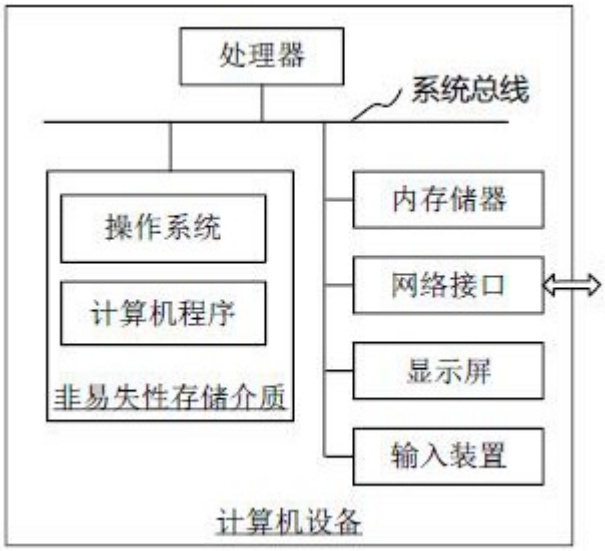


图4