(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 118428471 B
(45)授权公告日 2024.09.24

(21)申请号 202410876214.9
(22)申请日 2024.07.02
(65)同一申请的已公布的文献号
申请公布号 CN 118428471 A
(43)申请公布日 2024.08.02
(73)专利权人 湖南董因信息技术有限公司
地址 410000 湖南省长沙市开福区清水塘街道芙蓉中路一段319号绿地中心新华保险大厦栋2401房
(72)发明人 关相承 修保新
(74)专利代理机构 东莞卓诚专利代理事务所(普通合伙) 44754
专利代理师 朱鹏
(51)Int.Cl.
G06N 5/025(2023.01)
G06F 40/284(2020.01)
(56)对比文件
CN 116313121 A,2023.06.23
审查员 尚晓娟

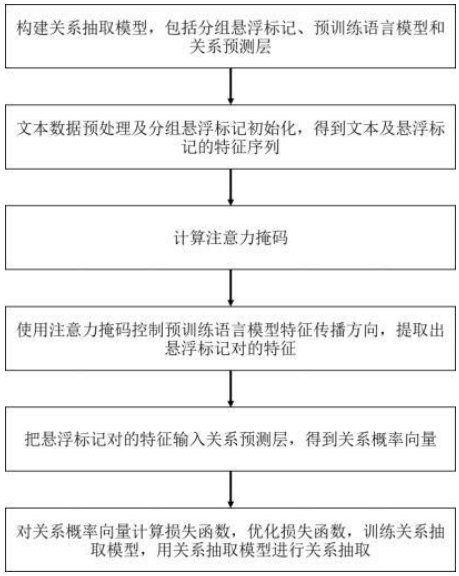
权利要求书3页 说明书8页 附图1页

(54)发明名称

基于预训练模型增强的图谱关系抽取方法

(57)摘要

本发明公开了基于预训练模型增强的图谱关系抽取方法,所述方法包括:构建关系抽取模型,包括分组悬浮标记、预训练语言模型和关系预测层;文本数据预处理及分组悬浮标记初始化,得到文本及悬浮标记的特征序列;计算注意力掩码;使用注意力掩码控制预训练语言模型特征传播方向,提取出悬浮标记对的特征;把悬浮标记对的特征输入关系预测层,得到关系概率向量;对关系概率向量计算损失函数,优化损失函数,训练关系抽取模型,用关系抽取模型进行关系抽取。本发明提出了分组悬浮标记的实体对表示方法,通过对悬浮标记进行分组,每个组复用头实体的特征,设计特定的注意力掩码,实现实体对特征的高效聚合,在较少计算量下实现了高精度的关系抽取。



CN 118428471 B

1.基于预训练模型增强的图谱关系抽取方法,其特征在于,所述方法包括:

步骤1,构建关系抽取模型,包括分组悬浮标记、预训练语言模型和关系预测层;

步骤2,文本数据预处理及分组悬浮标记初始化,得到文本及悬浮标记的特征序列;

步骤3,计算注意力掩码;

步骤4,使用注意力掩码控制预训练语言模型特征传播方向,提取出悬浮标记对的特征;

步骤5,把悬浮标记对的特征输入关系预测层,得到关系概率向量;

步骤6,对关系概率向量计算损失函数,优化损失函数,训练关系抽取模型,用关系抽取模型进行关系抽取;

所述的图谱为医疗知识图谱,所述图谱的实体包括疾病、症状、药品、手术,所述的图谱的关系包括疾病-症状关系、疾病-药品关系、疾病-疾病关系、症状-症状关系、疾病-手术关系;

所述的文本数据预处理及分组悬浮标记初始化,得到文本及悬浮标记的特征序列,包括以下步骤:

步骤201,对输入文本进行分词,得到分词序列;

步骤202,在分词序列的每个实体前插入“<e>”标记,在每个实体后插入“</e>”标记,用于标记出实体的位置,并在分词序列首部插入起始标记“<CLS>”,在尾部插入终止标记“<SEP>”;

步骤203,使用预训练语言模型Roberta-large的词嵌入模型把分词序列映射为词向量序列,对于总分词数为w_n,总实体数为m的分词序列,该分词序列映射得到的词向量序列数学表达式为:

$$S_{emb} = \left[E_{[\mathrm{CLS}]}, E_{w_1}, \ldots, E_{<\mathrm{e}>_1}, \ldots, E_{<\mathrm{e}>_m}, \ldots, E_{w_n}, E_{[\mathrm{SEP}]}\right];$$

其中,$E_{[\mathrm{CLS}]}$表示起始标记“<CLS>”的词向量,$E_{[\mathrm{SEP}]}$表示终止标记“<SEP>”的词向量,E_{w_i}表示第i个词的词向量,$E_{<\mathrm{e}>_i}$表示第i个“<e>”标记的词向量,每个“<e>”标记内容是固定的,因此每个“<e>”标记词向量是相同的;

步骤204,使用预训练语言模型Roberta-large的位置嵌入模型获得分词序列的位置嵌入序列,对于所述的分词序列获得的位置嵌入序列的数学表达式为:

$$S_{pos} = \left[P_{[\mathrm{CLS}]}, P_{w_1}, \ldots, P_{<\mathrm{e}>_1}, \ldots, P_{<\mathrm{e}>_m}, \ldots, P_{w_n}, P_{[\mathrm{SEP}]}\right];$$

其中,$P_{[\mathrm{CLS}]}$表示起始标记“<CLS>”的位置嵌入,$P_{[\mathrm{SEP}]}$表示终止标记“<SEP>”的位置嵌入,P_{w_i}表示第i个词的位置嵌入,$P_{<\mathrm{e}>_i}$表示第i个“<e>”标记的位置嵌入,每个“<e>” 标记位置是不同的,因此每个“<e>”标记的位置嵌入是不同的;

步骤205,把分词序列映射得到的词向量序列S_{emb}和分词序列的位置嵌入序列S_{pos}按元素相加,得到分词序列的特征嵌入序列S_{text},数学表达式为:

$$S_{text} = S_{emb} + S_{pos};$$

步骤206,生成悬浮标记特征;第i个悬浮标记的特征为第i个“<e>”标记的词向量$E_{<e>_i}$加上第i个“<e>”标记的位置嵌入$P_{<e>_i}$,数学表达式为:

$$M_i = E_{<e>_i} + P_{<e>_i};$$

其中,M_i表示第i个悬浮标记的特征;

步骤207,生成悬浮标记特征序列;实体数为m,则有m个悬浮标记,目标是生成包含m组悬浮标记的悬浮标记特征序列,第i组悬浮标记的生成方式为:把第i个悬浮标记的特征M_i放在第i组悬浮标记序列的开头,其他悬浮标记按在文本中出现的顺序从小到大排在第i组悬浮标记序列的后面,其中i=1,2,3,…,m;将m组悬浮标记特征序列按顺序拼接在一起,得到长度为m^2的悬浮标记特征序列S_{mark};

步骤208,把分词序列的特征嵌入序列S_{text}和悬浮标记特征序列S_{mark}拼接在一起,数学表达式为:

$$S_{total} = [S_{text}, S_{mark}];$$

其中,S_{total}表示文本及悬浮标记的特征序列;

所述的计算注意力掩码,包括以下步骤:

所述的分词序列的特征嵌入序列S_{text}序列长度为L_t,悬浮标记特征序列S_{mark}序列长度为L_m,实体数$m = \sqrt{L_m}$,生成一个大小为$(L_t + L_m) \times (L_t + L_m)$的矩阵$A$,矩阵中元素赋值的数学表达式为:

$$A_{ij} = \begin{cases} 1, \{ij|j < L_t\} \cup \{ij|i = j\} \cup \{ij|i = L_t + k \cdot m, i < j < i + m - 1, k \in [0, m)\} \\ 0, \text{其他} \end{cases};$$

其中,A是注意力掩码,A_{ij}表示A第i行第j列的元素;

所述的使用注意力掩码控制预训练语言模型特征传播方向,提取出悬浮标记对的特征,包括以下步骤:

步骤401,把所述的文本及悬浮标记的特征序列S_{total}输入预训练语言模型Roberta-large中,并用所述的注意力掩码A作为Roberta-large前向传播的掩码,数学表达式为:

$$H = Roberta - large(S_{total}, A);$$

其中,$H \in \mathbb{R}^{(L_t+L_m)\times d}$是$Roberta - large$输出的最后一层隐藏层的特征,d是$Roberta - large$的隐藏层维度,$L_t$为所述的分词序列的特征嵌入序列$S_{text}$序列长度,$L_m$为所述的悬浮标记特征序列$S_{mark}$序列长度;

步骤402,从$Roberta - large$输出的最后一层隐藏层的特征H选取出每个实体对的特征,数学表达式如下:

$$F_{ij} = H[L_t + i \cdot \sqrt{L_m} + j];$$

其中,F_{ij}表示第i个实体和第j个实体的悬浮标记对的特征,[]表示从目标张量的第0

维度进行索引的操作。

2.根据权利要求1所述的基于预训练模型增强的图谱关系抽取方法,其特征在于,所述的把悬浮标记对的特征输入关系预测层,得到关系概率向量,包括以下步骤:

把所述的第i个实体和第j个实体的悬浮标记对的特征F_{ij},输入全连接层,获得第i个实体和第j个实体的关系预测向量,数学表达式为:

$$R_{ij} = softmax(W_r F_{ij} + b_r);$$

其中,$R_{ij} \in R^C$表示第i个实体和第j个实体的关系预测向量,$W_r \in R^{C \times d}$表示全连接层的权重矩阵,$b_r \in R^C$表示全连接层的偏置向量,C表示关系类别的数量,d表示悬浮标记对特征的维度,悬浮标记对特征的维度与$Roberta - large$的隐藏层维度相等,$softmax$是激活函数,用于把向量归一化为概率分布。

3.根据权利要求2所述的基于预训练模型增强的图谱关系抽取方法,其特征在于,所述的对关系概率向量计算损失函数,优化损失函数,训练关系抽取模型,用关系抽取模型进行关系抽取,包括以下步骤:

计算第i个实体和第j个实体的关系预测向量R_{ij}与真实关系标签Y_{ij}之间的交叉熵损失,数学表达式为:

$$L_{ij} = -\sum_{k=1}^{C} Y_{ij}(k)\log(R_{ij}(k));$$

其中,$Y_{ij}(k)$表示第i个实体和第j个实体的真实关系标签,若第i个实体和第j个实体具有第k类关系,则$Y_{ij}(k) = 1$,否则$Y_{ij}(k) = 0$,$R_{ij}(k)$表示模型预测的第i个实体和第j个实体具有第k类的概率,为所述的关系预测向量R_{ij}的索引值;

计算所有实体对的交叉熵损失,数学表达式如下:

$$L_{total} = \frac{1}{m(m-1)} \sum_{i=0}^{m-1} \sum_{j \neq i}^{m-1} L_{ij};$$

其中,L_{total}表示总的交叉熵损失;

使用Adam优化算法对L_{total}进行优化,训练关系抽取模型。

基于预训练模型增强的图谱关系抽取方法

技术领域

[0001] 本发明涉及深度学习和自然语言处理领域,尤其涉及一种基于预训练模型增强的图谱关系抽取方法。

背景技术

[0002] 关系抽取是自然语言处理中的一项任务,旨在从文本中识别和提取出实体之间的关系。给定一段文本和已标注的实体对,任务的目标是确定这些实体之间的关系类型或关系类别。关系抽取在自然语言处理和信息抽取领域具有重要的应用和价值,包括但不限于以下方面:知识图谱构建、信息检索与推荐、事件抽取与情报分析、社交网络分析和自动问答和智能助理等。

[0003] 当前的医学知识图谱的关系抽取方法大部分需要设计复杂的关系抽取模块,对语言模型输出的文本特征进行复杂的处理,计算量大,计算效率低。少部分方法通过设计悬浮标记,能够在一定程度上减少计算量,然而,现有的悬浮标记方法存在表示效率低的问题,这阻碍了算法的研究和落地。因此,如何设计一个实体抽取方法,通过改进其实体的表示法,使其能够高效表示实体特征,有其学术研究意义及产业应用意义。

发明内容

[0004] 本发明旨在至少解决现有技术中存在的技术问题之一。为此,本发明公开了基于预训练模型增强的图谱关系抽取方法。所述方法能够实现高效的关系抽取,相比现有方法,本方法创新性得提出了分组悬浮标记的实体对表示方法,通过对悬浮标记进行分组,每个组复用一个头实体的特征,并对分组悬浮标记设计特定的注意力掩码,实现了实体对特征的高效聚合,在计算量较少的情况下实现了高精度的关系抽取。

[0005] 本发明的目的是通过如下技术方案实现的,基于预训练模型增强的图谱关系抽取方法,所述方法包括:

[0006] 步骤1,构建关系抽取模型,包括分组悬浮标记、预训练语言模型和关系预测层;

[0007] 步骤2,文本数据预处理及分组悬浮标记初始化,得到文本及悬浮标记的特征序列;

[0008] 步骤3,计算注意力掩码;

[0009] 步骤4,使用注意力掩码控制预训练语言模型特征传播方向,提取出悬浮标记对的特征;

[0010] 步骤5,把悬浮标记对的特征输入关系预测层,得到关系概率向量;

[0011] 步骤6,对关系概率向量计算损失函数,优化损失函数,训练关系抽取模型,用关系抽取模型进行关系抽取。

[0012] 所述的文本数据预处理及分组悬浮标记初始化,得到文本及悬浮标记的特征序列,包括以下步骤:

[0013] 步骤201,对输入文本进行分词,得到分词序列;

[0014] 步骤202,在分词序列的每个实体前插入“<e>”标记,在每个实体后插入“</e>”标记,用于标记出实体的位置,并在分词序列首部插入起始标记“<CLS>”,在尾部插入终止标记“<SEP>”;

[0015] 步骤203,使用预训练语言模型Roberta-large的词嵌入模型把分词序列映射为词向量序列,对于总分词数为w_n,总实体数为m的分词序列,该分词序列映射得到的词向量序列数学表达式为:

[0016] $$S_{emb} = \left[E_{[CLS]}, E_{w_1}, \ldots, E_{<e>_1}, \ldots, E_{<e>_m}, \ldots, E_{w_n}, E_{[SEP]}\right]$$

[0017] 其中,$E_{[CLS]}$表示起始标记“<CLS>”的词向量,$E_{[SEP]}$表示终止标记“<SEP>”的词向量,E_{w_i}表示第i个词的词向量,$E_{<e>_i}$表示第i个“<e>”标记的词向量,每个“<e>”标记内容是固定的,因此每个“<e>”标记词向量是相同的;

[0018] 步骤204,使用预训练语言模型Roberta-large的位置嵌入模型获得分词序列的位置嵌入序列,对于步骤203所述的分词序列,获得的位置嵌入序列的数学表达式为:

[0019] $$S_{pos} = \left[P_{[CLS]}, P_{w_1}, \ldots, P_{<e>_1}, \ldots, P_{<e>_m}, \ldots, P_{w_n}, P_{[SEP]}\right]$$

[0020] 其中,$P_{[CLS]}$表示起始标记“<CLS>”的位置嵌入,$P_{[SEP]}$表示终止标记“<SEP>”的位置嵌入,P_{w_i}表示第i个词的位置嵌入,$P_{<e>_i}$表示第i个“<e>”标记的位置嵌入,每个“<e>”标记位置是不同的,因此每个“<e>”标记的位置嵌入是不同的;

[0021] 步骤205,把分词序列映射得到的词向量序列S_{emb}和分词序列的位置嵌入序列S_{pos}按元素相加,得到分词序列的特征嵌入序列S_{text},数学表达式为:

[0022] $$S_{text} = S_{emb} + S_{pos}$$

[0023] 步骤206,生成悬浮标记特征;第i个悬浮标记的特征为第i个“<e>”标记的词向量$E_{<e>_i}$加上第i个“<e>”标记的位置嵌入$P_{<e>_i}$,数学表达式为:

[0024] $$M_i = E_{<e>_i} + P_{<e>_i}$$

[0025] 其中,M_i表示第i个悬浮标记的特征;

[0026] 步骤207,生成悬浮标记特征序列;实体数为m,则有m个悬浮标记,目标是生成包含m组悬浮标记的悬浮标记特征序列,第i组悬浮标记的生成方式为:把第i个悬浮标记的特征M_i放在第i组悬浮标记序列的开头,其他悬浮标记按在文本中出现的顺序从小到大排在第i组悬浮标记序列的后面,其中i=1,2,3,…,m;将m组悬浮标记特征序列按顺序拼接在一起,得到长度为m^2的悬浮标记特征序列S_{mark};

[0027] 步骤208,把分词序列的特征嵌入序列S_{text}和悬浮标记特征序列S_{mark}拼接在一起,数学表达式为:

[0028] $$S_{total} = [S_{text}, S_{mark}]$$

[0029] 其中,S_{total}表示文本及悬浮标记的特征序列。

[0030] 所述的计算注意力掩码,包括以下步骤:

[0031] 所述的分词序列的特征嵌入序列S_{text}序列长度为L_t，悬浮标记特征序列S_{mark}序列长度为L_m，实体数$m=\sqrt{L_m}$，生成一个大小为$(L_t+L_m)\times(L_t+L_m)$的矩阵A，矩阵中元素赋值的数学表达式为：

[0032] $$A_{ij}=\begin{cases}1, \{ij|j<L_t\}\cup\{ij|i=j\}\cup\{ij|i=L_t+k\cdot m, i<j<i+m-1, k\in[0,m)\} \\ 0, \text{其他}\end{cases}$$

[0033] 其中，A是注意力掩码，A_{ij}表示A第i行第j列的元素。

[0034] 所述的使用注意力掩码控制预训练语言模型特征传播方向，提取出悬浮标记对的特征，包括以下步骤：

[0035] 步骤401，把所述的文本及悬浮标记的特征序列S_{total}输入预训练语言模型Roberta-large中，并用所述的注意力掩码A作为Roberta-large前向传播的掩码，数学表达式为：

[0036] $$H=Roberta-large(S_{total},A)$$

[0037] 其中，$H\in\mathbb{R}^{(L_t+L_m)\times d}$是$Roberta-large$输出的最后一层隐藏层的特征，d是$Roberta-large$的隐藏层维度，$L_t$为所述的分词序列的特征嵌入序列$S_{text}$序列长度，$L_m$为所述的悬浮标记特征序列$S_{mark}$序列长度；

[0038] 步骤402，从$Roberta-large$输出的最后一层隐藏层的特征H选取出每个实体对的特征，数学表达式如下：

[0039] $$F_{ij}=H[L_t+i\cdot\sqrt{L_m}+j]$$

[0040] 其中，F_{ij}表示第i个实体和第j个实体的悬浮标记对的特征，[]表示从目标张量的第0维度进行索引的操作。

[0041] 所述的把悬浮标记对的特征输入关系预测层，得到关系概率向量，包括以下步骤：

[0042] 把所述的第i个实体和第j个实体的悬浮标记对的特征F_{ij}，输入全连接层，获得第i个实体和第j个实体的关系预测向量，数学表达式为：

[0043] $$R_{ij}=softmax(W_rF_{ij}+b_r)$$

[0044] 其中，$R_{ij}\in R^C$表示第i个实体和第j个实体的关系预测向量，$W_r\in R^{C\times d}$表示全连接层的权重矩阵，$b_r\in R^C$表示全连接层的偏置向量，C表示关系类别的数量，d表示悬浮标记对特征的维度，$softmax$是激活函数，用于把向量归一化为概率分布。

[0045] 所述的对关系概率向量计算损失函数，优化损失函数，训练关系抽取模型，用关系抽取模型进行关系抽取，包括以下步骤：

[0046] 计算第i个实体和第j个实体的关系预测向量R_{ij}与真实关系标签Y_{ij}之间的交叉熵损失，数学表达式为：

[0047] $L_{ij} = -\sum_{k=1}^{C} Y_{ij}(k)\log(R_{ij}(k))$

[0048] 其中，$Y_{ij}(k)$表示第i个实体和第j个实体的真实关系标签，若第i个实体和第j个实体具有第k类关系，则$Y_{ij}(k) = 1$，否则$Y_{ij}(k) = 0$，$R_{ij}(k)$表示模型预测的第i个实体和第j个实体具有第k类的概率，为所述的关系预测向量R_{ij}的索引值；

[0049] 计算所有实体对的交叉熵损失，数学表达式如下：

[0050] $$L_{total} = \frac{1}{m(m-1)} \sum_{i=0}^{m-1} \sum_{j \neq i}^{m-1} L_{ij}$$

[0051] 其中，L_{total}表示总的交叉熵损失；

[0052] 使用Adam优化算法对L_{total}进行优化，训练关系抽取模型。

[0053] 与现有方法相比，本发明方法的优点在于：本技术提供了基于预训练模型增强的图谱关系抽取方法。本方法创新性得提出了分组悬浮标记的实体对表示方法，通过对悬浮标记进行分组，每个组复用一个头实体的特征，并对分组悬浮标记设计特定的注意力掩码，实现了实体对特征的高效聚合，在计算量较少的情况下实现了高精度的关系抽取。

附图说明

[0054] 图1示出了本发明实施例的流程示意图。

具体实施方式

[0055] 为了使本发明的目的、技术方案和优点更加清楚，下面将结合附图对本发明作进一步地详细描述，显然，所描述的实施例仅仅是本发明一部份实施例，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它实施例，都属于本发明保护的范围。

[0056] 应当理解，此处所描述的具体实施例仅仅用以解释本发明，并不用于限定本发明。

[0057] 知识图谱是以结构化形式描述的知识及其联系的集合，是一种将实体、属性和关系组织成图形结构的知识表示方式，旨在更好地描述和理解世界的知识和概念，可以用于存储、查询、推理和分析知识。知识图谱的设计不仅仅是将知识以图形结构的形式表示，还需要考虑如何定义其属性和关系，如何建立它们之间的联系以及如何进行查询、推理和分析等操作。这样的设计使知识图谱成为一个强大的工具，用于存储和操作大量复杂的知识。

[0058] 本实施例中，互联网中存在海量的医疗领域知识可以用于医疗病情咨询、健康养生，但传统的搜索引擎并无法根据病人的实际情况做出合理的判断，无法满足使用需求。假设搭建大规模医疗知识图谱，需要在互联网上爬取海量的文本数据并将其结构化，关系抽取作为文本结构化中的一个关键环节，在对文本进行关系抽取的过程中，基于预训练模型增强的图谱关系抽取方法用于医疗领域，抽取医疗相关的关系。构建了一个可靠的中文医疗知识系统，该系统能够帮助解决人们对日常疾病相关知识的需求，具有很高的应用价值。

[0059] 医疗知识图谱(Medical Knowledge Graph)作为医疗人工智能的核心，本质上是

一种揭示医疗实体之间关系的语义网络,可以对现实世界的事物及其相互关系进行形式化描述。通常情况下,医疗知识图谱是在人工构建的专业知识基础上,通过算法以及专家审核的方式不断扩充实体及关系来构建的,包括疾病、症状、药品、手术等医学概念和多种医学关系。在广泛的医疗场景中,医疗知识图谱已经被证明可以有效地为算法提供医学知识支撑并为算法的预测结果提供医学解释。在可预见的未来,知识图谱都将在医疗这个强知识属性的领域发挥至关重要的作用。所以本发明实施例的基于预训练模型增强的图谱关系抽取方法能够为医学知识图谱的关系抽取提供非常重要的支撑。

[0060] 由此,如图1所示,基于预训练模型增强的图谱关系抽取方法,所述方法包括:

[0061] 步骤1,构建关系抽取模型,包括分组悬浮标记、预训练语言模型和关系预测层;

[0062] 步骤2,文本数据预处理及分组悬浮标记初始化,得到文本及悬浮标记的特征序列;

[0063] 步骤3,计算注意力掩码;

[0064] 步骤4,使用注意力掩码控制预训练语言模型特征传播方向,提取出悬浮标记对的特征;

[0065] 步骤5,把悬浮标记对的特征输入关系预测层,得到关系概率向量;

[0066] 步骤6,对关系概率向量计算损失函数,优化损失函数,训练关系抽取模型,用关系抽取模型进行关系抽取。

[0067] 所述的图谱为医疗知识图谱,所述图谱的实体包括疾病、症状、药品、手术,所述图谱的关系包括疾病-症状关系、疾病-药品关系、疾病-疾病关系、症状-症状关系、疾病-手术关系。

[0068] 所述的文本数据预处理及分组悬浮标记初始化,得到文本及悬浮标记的特征序列,包括以下步骤:

[0069] 步骤201,对输入文本进行分词,得到分词序列;

[0070] 步骤202,在分词序列的每个实体前插入“<e>”标记,在每个实体后插入“</e>”标记,用于标记出实体的位置,并在分词序列首部插入起始标记“<CLS>”,在尾部插入终止标记“<SEP>”;

[0071] 步骤203,使用预训练语言模型Roberta-large的词嵌入模型把分词序列映射为词向量序列,对于总分词数为w_n,总实体数为m的分词序列,该分词序列映射得到的词向量序列数学表达式为:

[0072] $$S_{emb}=\left[E_{[\mathrm{CLS}]},E_{w_1},\ldots,E_{<\mathrm{e}>_1},\ldots,E_{<\mathrm{e}>_m},\ldots,E_{w_n},E_{[\mathrm{SEP}]}\right]$$

[0073] 其中,$E_{[\mathrm{CLS}]}$表示起始标记“<CLS>”的词向量,$E_{[\mathrm{SEP}]}$表示终止标记“<SEP>”的词向量,E_{w_i}表示第i个词的词向量,$E_{<\mathrm{e}>_i}$表示第i个“<e>”标记的词向量,每个“<e>”标记内容是固定的,因此每个“<e>”标记词向量是相同的;

[0074] 步骤204,使用预训练语言模型Roberta-large的位置嵌入模型获得分词序列的位置嵌入序列,对于步骤203所述的分词序列,获得的位置嵌入序列的数学表达式为:

[0075] $$S_{pos}=\left[P_{[\mathrm{CLS}]},P_{w_1},\ldots,P_{<\mathrm{e}>_1},\ldots,P_{<\mathrm{e}>_m},\ldots,P_{w_n},P_{[\mathrm{SEP}]}\right]$$

[0076] 其中,$P_{[\mathrm{CLS}]}$表示起始标记“<CLS>”的位置嵌入,$P_{[\mathrm{SEP}]}$表示终止标记“<SEP>”的位

置嵌入，P_{w_i}表示第i个词的位置嵌入，$P_{<e>_i}$表示第i个“<e>”标记的位置嵌入，每个“<e>”标记位置是不同的，因此每个“<e>”标记的位置嵌入是不同的；

[0077] 步骤205，把分词序列映射得到的词向量序列S_{emb}和分词序列的位置嵌入序列S_{pos}按元素相加，得到分词序列的特征嵌入序列S_{text}，数学表达式为：

[0078] $$S_{text} = S_{emb} + S_{pos}$$

[0079] 步骤206，生成悬浮标记特征；第i个悬浮标记的特征为第i个“<e>”标记的词向量$E_{<e>_i}$加上第i个“<e>”标记的位置嵌入$P_{<e>_i}$，数学表达式为：

[0080] $$M_i = E_{<e>_i} + P_{<e>_i}$$

[0081] 其中，M_i表示第i个悬浮标记的特征；

[0082] 步骤207，生成悬浮标记特征序列；实体数为m，则有m个悬浮标记，目标是生成包含m组悬浮标记的悬浮标记特征序列，第i组悬浮标记的生成方式为：把第i个悬浮标记的特征M_i放在第i组悬浮标记序列的开头，其他悬浮标记按在文本中出现的顺序从小到大排在第i组悬浮标记序列的后面，其中i=1,2,3,…,m；将m组悬浮标记特征序列按顺序拼接在一起，得到长度为m^2的悬浮标记特征序列S_{mark}；

[0083] 步骤208，把分词序列的特征嵌入序列S_{text}和悬浮标记特征序列S_{mark}拼接在一起，数学表达式为：

[0084] $$S_{total} = [S_{text}, S_{mark}]$$

[0085] 其中，S_{total}表示文本及悬浮标记的特征序列。

[0086] RoBERTa-large是基于BERT(Bidirectional Encoder Representations from Transformers)模型的变体之一，由Facebook AI(现称为Meta AI)开发。RoBERTa的全称是"A Robustly Optimized BERT Pretraining Approach"，其在BERT的基础上进行了多项改进和优化。以下是RoBERTa-large的一些关键特点：(1)模型规模：RoBERTa-large比BERT-large更大，拥有24层Transformer编码器，每层有1024个隐藏单元，总共有355M参数。相比之下，BERT-large有24层，每层有1024个隐藏单元，总共340M参数。(2)预训练数据量：RoBERTa使用了更大的预训练数据集，约160GB的数据，比BERT的16GB要多得多。这包括BookCorpus、English Wikipedia、CC-News、OpenWebText和Stories等数据集。(3)预训练策略：RoBERTa在预训练过程中进行了更多的优化。例如，去掉了BERT中的Next Sentence Prediction(NSP)任务，并使用更长的训练序列(更长的句子)。(4)训练时间：RoBERTa进行了更长时间的预训练，以确保模型更好地捕捉语言模式和上下文关系。(5)效果提升：由于上述优化，RoBERTa在多个自然语言处理任务上的表现优于BERT，包括文本分类、问答、文本生成等任务。

[0087] 所述的计算注意力掩码，包括以下步骤：

[0088] 所述的分词序列的特征嵌入序列S_{text}序列长度为L_t，悬浮标记特征序列S_{mark}序列长度为L_m，实体数$m = \sqrt{L_m}$，生成一个大小为$(L_t + L_m) \times (L_t + L_m)$的矩阵$A$，矩

阵中元素赋值的数学表达式为：

[0089] $$A_{ij}=\begin{cases}1, \{ij|j<L_t\}\cup\{ij|i=j\}\cup\{ij|i=L_t+k\cdot m, i<j<i+m-1, k\in[0,m)\} \\ 0, \text{其他}\end{cases}$$

[0090] 其中，A是注意力掩码，A_{ij}表示A第i行第j列的元素。

[0091] 注意力掩码在Transformer模型中的作用之一是控制信息传播，即决定哪些位置的信息可以相互影响。

[0092] 注意力机制在计算注意力权重时将每个位置与其他位置进行交互，并根据它们的相关性来分配权重。通过在注意力掩码中标记某些位置，我们可以控制模型在计算注意力权重时是否将这些位置考虑在内。

[0093] 所述的使用注意力掩码控制预训练语言模型特征传播方向，提取出悬浮标记对的特征，包括以下步骤：

[0094] 步骤401，把所述的文本及悬浮标记的特征序列S_{total}输入预训练语言模型Roberta-large中，并用所述的注意力掩码A作为Roberta-large前向传播的掩码，数学表达式为：

[0095] $$H=Roberta-large(S_{total},A)$$

[0096] 其中，$H\in\mathbb{R}^{(L_t+L_m)\times d}$是$Roberta-large$输出的最后一层隐藏层的特征，d是$Roberta-large$的隐藏层维度，$L_t$为所述的分词序列的特征嵌入序列$S_{text}$序列长度，$L_m$为所述的悬浮标记特征序列$S_{mark}$序列长度；

[0097] 步骤402，从$Roberta-large$输出的最后一层隐藏层的特征H选取出每个实体对的特征，数学表达式如下：

[0098] $$F_{ij}=H[L_t+i\cdot\sqrt{L_m}+j]$$

[0099] 其中，F_{ij}表示第i个实体和第j个实体的悬浮标记对的特征，[]表示从目标张量的第0维度进行索引的操作。

[0100] 所述的把悬浮标记对的特征输入关系预测层，得到关系概率向量，包括以下步骤：

[0101] 把所述的第i个实体和第j个实体的悬浮标记对的特征F_{ij}，输入全连接层，获得第i个实体和第j个实体的关系预测向量，数学表达式为：

[0102] $$R_{ij}=softmax(W_rF_{ij}+b_r)$$

[0103] 其中，$R_{ij}\in R^C$表示第i个实体和第j个实体的关系预测向量，$W_r\in R^{C\times d}$表示全连接层的权重矩阵，$b_r\in R^C$表示全连接层的偏置向量，C表示关系类别的数量，d表示悬浮标记对特征的维度，$softmax$是激活函数，用于把向量归一化为概率分布。

[0104] 所述的对关系概率向量计算损失函数，优化损失函数，训练关系抽取模型，用关系抽取模型进行关系抽取，包括以下步骤：

[0105] 计算第i个实体和第j个实体的关系预测向量R_{ij}与真实关系标签Y_{ij}之间的交叉

熵损失,数学表达式为:

[0106] $$L_{ij} = -\sum_{k=1}^{C} Y_{ij}(k)\log(R_{ij}(k))$$

[0107] 其中,$Y_{ij}(k)$表示第i个实体和第j个实体的真实关系标签,若第i个实体和第j个实体具有第k类关系,则$Y_{ij}(k)=1$,否则$Y_{ij}(k)=0$,$R_{ij}(k)$表示模型预测的第i个实体和第j个实体具有第k类的概率,为所述的关系预测向量R_{ij}的索引值;

[0108] 计算所有实体对的交叉熵损失,数学表达式如下:

[0109] $$L_{total} = \frac{1}{m(m-1)} \sum_{i=0}^{m-1} \sum_{j \neq i}^{m-1} L_{ij}$$

[0110] 其中,L_{total}表示总的交叉熵损失;

[0111] 使用Adam优化算法对L_{total}进行优化,训练关系抽取模型。

[0112] 本领域技术人员应明白,本申请的实施例可提供为方法、系统或计算机程序产品。因此,本申请可采用完全硬件实施例、完全软件实施例或结合软件和硬件方面的实施例的形式。而且,本申请可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质(包括但不限于磁盘存储器、CD-ROM、光学存储器等)上实施的计算机程序产品的形式。

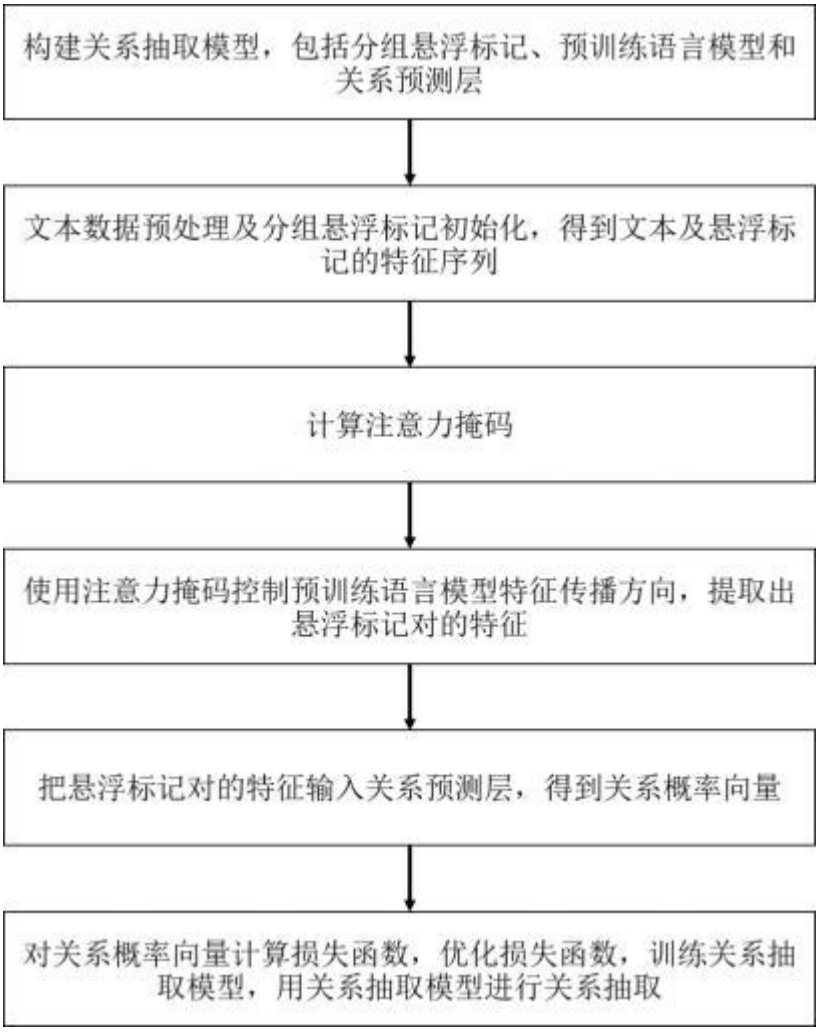

构建关系抽取模型，包括分组悬浮标记、预训练语言模型和关系预测层
文本数据预处理及分组悬浮标记初始化，得到文本及悬浮标记的特征序列
计算注意力掩码
使用注意力掩码控制预训练语言模型特征传播方向，提取出悬浮标记对的特征
把悬浮标记对的特征输入关系预测层，得到关系概率向量
对关系概率向量计算损失函数，优化损失函数，训练关系抽取模型，用关系抽取模型进行关系抽取

图 1