(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 111639171 A
(43)申请公布日 2020.09.08

(21)申请号 202010512399.7

(22)申请日 2020.06.08

(71)申请人 吉林大学
地址 130012 吉林省长春市前进大街2699号

(72)发明人 彭涛 崔海 刘露 包铁 王上 张雪松 梁琪

(74)专利代理机构 北京集佳知识产权代理有限公司 11227
代理人 张娜

(51)Int.Cl.
G06F 16/332(2019.01)
G06F 16/36(2019.01)
G06F 40/30(2020.01)

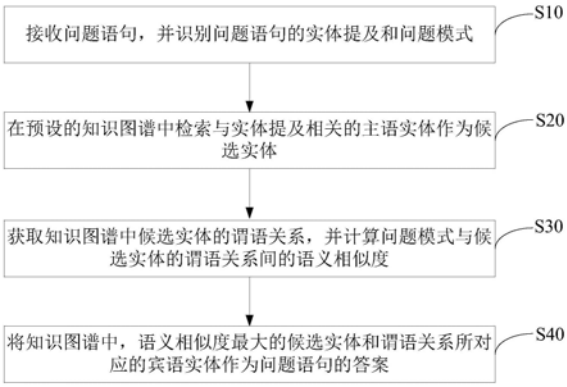
权利要求书3页 说明书12页 附图3页

(54)发明名称

一种知识图谱问答方法及装置

(57)摘要

本申请提供一种知识图谱问答方法及装置，该方法包括：接收问题语句，并识别问题语句的实体提及和问题模式；在预设的知识图谱中检索与实体提及相关的主语实体作为候选实体；获取知识图谱中候选实体的谓语关系，并计算问题模式与候选实体的谓语关系间的语义相似度；将知识图谱中，语义相似度最大的候选实体和谓语关系所对应的宾语实体作为问题语句的答案。本申请能够对问题语句的问题模式和知识图谱的谓语关系进行语义的联合分析，从而识别出知识图谱中语义最相关的宾语实体作为答案，从而提高问答结果的准确率。



CN 111639171 A

1.一种知识图谱问答方法,其特征在于,所述方法包括:

接收问题语句,并识别问题语句的实体提及和问题模式;

在预设的知识图谱中检索与所述实体提及相关的主语实体作为候选实体;

获取所述知识图谱中所述候选实体的谓语关系,并计算所述问题模式与所述候选实体的谓语关系间的语义相似度;

将所述知识图谱中,语义相似度最大的候选实体和谓语关系所对应的宾语实体作为所述问题语句的答案。

2.根据权利要求1所述的方法,其特征在于,所述在预设的知识图谱中检索与所述实体提及相关的主语实体作为候选实体,包括:

建立所述知识图谱中主语实体与所述主语实体的n-gram集合的反向映射索引,所述主语实体的n-gram集合中包含所述主语实体的所有组合方式;

生成所述实体提及的n-gram集合,所述实体提及的n-gram集合包含所述实体提及的所有组合方式;

采用启发式算法匹配所述实体提及的n-gram集合与所述主语实体的n-gram集合,基于所述反向映射索引将匹配到的主语实体作为候选实体。

3.根据权利要求2所述的方法,其特征在于,所述在预设的知识图谱中检索与所述实体提及相关的主语实体作为候选实体,还包括:

调用已训练的多标签分类模型,所述多标签分类模型是预先通过第一问题模式样本、以及为所述第一问题模式样本所标注的主题标签训练得到的;

将所述问题模式输入至所述多标签分类模型中,通过所述多标签分类模型获得所述问题模式所属主题的第一概率;

确定所述候选实体的主题,并从所述第一概率中获取所述问题模式属于所述候选实体的主题的第二概率;

计算所述候选实体与所述问题提及的编辑距离,并基于所述编辑距离和所述第二概率中的最大概率确定所述候选实体的评分;

筛选所述候选实体中评分符合预设排名的实体。

4.根据权利要求3所述的方法,其特征在于,所述多标签分类模型的训练过程,包括:

获取训练用的第一基础模型,所述第一基础模型为预设的文本分类模型;

基于所述知识图谱中的三元组生成所述第一问题模式样本,所述第一问题模式样本所标注的主题标签为所述三元组中的谓语关系;

将所述第一问题模式样本中用于本次训练的样本和其所标注的主题标签输入至所述文本分类模型中,并计算所述文本分类模型的交叉熵损失函数值;

在所述交叉熵损失函数值不符合预设的第一结束条件的情况下,调整所述文本分类模型的权重参数,并返回执行所述将所述第一问题模式样本中用于本次训练的样本和其所标注的主题标签输入至所述文本分类模型中;

在所述交叉熵损失函数值符合所述第一结束条件的情况下,将本次训练后的所述文本分类模型作为所述多标签分类模型。

5.根据权利要求1所述的方法,其特征在于,所述计算所述问题模式与所述候选实体的谓语关系间的语义相似度,包括:

调用已训练的关系检测模型，所述关系检测模型是预先通过第二问题模式样本、以及为所述第二问题模式样本所标注的关系标签训练得到的；

将所述问题模式与所述候选实体的谓语关系输入至所述关系检测模型中，通过所述关系检测模型获得所述问题模式与所述候选实体的谓语关系的语义相似度。

6.根据权利要求5所述的方法，其特征在于，所述关系检测模型的训练过程，包括：

获取训练用的第二基础模型，所述第二基础模型包括第一编码层、第二编码层、分类模型和输出层；

基于所述知识图谱中的三元组生成所述第二问题模式样本，所述第二问题模式样本包括正样本和负样本，所述正样本所标注的关系标签为所述三元组中的谓语关系，所述负样本所标注的关系标签非所述三元组中的谓语关系；

按照预设比例分别对所述正样本和所述负样本进行样本采集，得到用于本次训练的样本；

针对所述用于本次训练的样本，通过所述第一编码层生成该样本所标注的关系标签的嵌入向量，并将该样本所标注的关系标签的嵌入向量作为该样本所标注的关系标签的第一低维向量；

通过所述第二编码层生成该样本中词组的嵌入向量；

通过所述分类模型采用注意力机制处理所述词组的嵌入向量得到该样本的第二低维向量；

通过所述输出层计算所述第一低维向量与所述第二低维向量的关联程度，并基于所述关联程度确定折页损失函数值；

在所述折页损失函数值不符合预设的第二结束条件的情况下，基于所述折页损失函数值分别调整所述第一编码层、所述第二编码层和所述分类模型的权重参数，并返回执行所述按照预设比例分别对所述正样本和所述负样本进行样本采集，得到用于本次训练的样本；

在所述折页损失函数值符合所述第二结束条件的情况下，将本次训练后的所述第二基础模型作为所述关系检测模型。

7.一种知识图谱问答装置，其特征在于，所述装置包括：

实体检测模块，用于接收问题语句，并识别问题语句的实体提及和问题模式；

实体链接模块，用于在预设的知识图谱中检索与所述实体提及相关的主语实体作为候选实体；

关系检测模块，用于获取所述知识图谱中所述候选实体的谓语关系，并计算所述问题模式与所述候选实体的谓语关系间的语义相似度；

答案生成模块，用于将所述知识图谱中，语义相似度最大的候选实体和谓语关系所对应的宾语实体作为所述问题语句的答案。

8.根据权利要求7所述的装置，其特征在于，所述实体链接模块，具体用于：

建立所述知识图谱中主语实体与所述主语实体的n-gram集合的反向映射索引，所述主语实体的n-gram集合中包含所述主语实体的所有组合方式；生成所述实体提及的n-gram集合，所述实体提及的n-gram集合包含所述实体提及的所有组合方式；采用启发式算法匹配所述实体提及的n-gram集合与所述主语实体的n-gram集合，基于所述反向映射索引将匹配

到的主语实体作为候选实体。

9.根据权利要求8所述的装置,其特征在于,所述实体链接模块,还用于:

调用已训练的多标签分类模型,所述多标签分类模型是预先通过第一问题模式样本、以及为所述第一问题模式样本所标注的主题标签训练得到的;将所述问题模式输入至所述多标签分类模型中,通过所述多标签分类模型获得所述问题模式所属主题的第一概率;确定所述候选实体的主题,并从所述第一概率中获取所述问题模式属于所述候选实体的主题的第二概率;计算所述候选实体与所述问题提及的编辑距离,并基于所述编辑距离和所述第二概率中的最大概率确定所述候选实体的评分;筛选所述候选实体中评分符合预设排名的实体。

10.根据权利要求7所述的装置,其特征在于,所述关系检测模块,具体用于:

调用已训练的关系检测模型,所述关系检测模型是预先通过第二问题模式样本、以及为所述第二问题模式样本所标注的关系标签训练得到的;将所述问题模式与所述候选实体的谓语关系输入至所述关系检测模型中,通过所述关系检测模型获得所述问题模式与所述候选实体的谓语关系的语义相似度。

一种知识图谱问答方法及装置

技术领域

[0001] 本发明涉及人工智能技术领域，更具体地说，涉及一种知识图谱问答方法及装置。

背景技术

[0002] 近年来，随着知识图谱的发展，人们正在探索如何获取知识图谱中的有效知识。虽然诸如SPARQL、GraphQL等查询语言被设计用于知识图谱的检索，但由于查询语言的语法细节无法被终端用户所理解，因此基于知识图谱的问答系统应运而生，即当用户提出自然语言形式的问题时，该系统会通过检索知识图谱给出答案。

[0003] 针对单关系事实型问题，即给出一个自然语言形式的问题，仅需要知识图谱中的一个三元组<主语实体，谓语关系，宾语实体>便可以回答该问题，例如对于问题“苹果公司的创始人是谁?”，就可以通过知识图谱中的三元组“<苹果公司，创始人，乔布斯>”进行回答。目前，处理单关系事实型问题的方法主要是端到端的神经网络方法，但该方法仅仅考虑了主语实体在字面上是否相同，而没有考虑语义上是否相关，因此如果知识图谱中出现多个重名的主语实体，则无法准确区分，导致问答结果准确率很低。

发明内容

[0004] 有鉴于此，为解决上述问题，本发明提供知识图谱问答方法及装置，技术方案如下：

[0005] 一种知识图谱问答方法，所述方法包括：

[0006] 接收问题语句，并识别问题语句的实体提及和问题模式；

[0007] 在预设的知识图谱中检索与所述实体提及相关的主语实体作为候选实体；

[0008] 获取所述知识图谱中所述候选实体的谓语关系，并计算所述问题模式与所述候选实体的谓语关系间的语义相似度；

[0009] 将所述知识图谱中，语义相似度最大的候选实体和谓语关系所对应的宾语实体作为所述问题语句的答案。

[0010] 优选的，所述在预设的知识图谱中检索与所述实体提及相关的主语实体作为候选实体，包括：

[0011] 建立所述知识图谱中主语实体与所述主语实体的n-gram集合的反向映射索引，所述主语实体的n-gram集合中包含所述主语实体的所有组合方式；

[0012] 生成所述实体提及的n-gram集合，所述实体提及的n-gram集合包含所述实体提及的所有组合方式；

[0013] 采用启发式算法匹配所述实体提及的n-gram集合与所述主语实体的n-gram集合，基于所述反向映射索引将匹配到的主语实体作为候选实体。

[0014] 优选的，所述在预设的知识图谱中检索与所述实体提及相关的主语实体作为候选实体，还包括：

[0015] 调用已训练的多标签分类模型，所述多标签分类模型是预先通过第一问题模式样

本、以及为所述第一问题模式样本所标注的主题标签训练得到的；

[0016] 将所述问题模式输入至所述多标签分类模型中，通过所述多标签分类模型获得所述问题模式所属主题的第一概率；

[0017] 确定所述候选实体的主题，并从所述第一概率中获取所述问题模式属于所述候选实体的主题的第二概率；

[0018] 计算所述候选实体与所述问题提及的编辑距离，并基于所述编辑距离和所述第二概率中的最大概率确定所述候选实体的评分；

[0019] 筛选所述候选实体中评分符合预设排名的实体。

[0020] 优选的，所述多标签分类模型的训练过程，包括：

[0021] 获取训练用的第一基础模型，所述第一基础模型为预设的文本分类模型；

[0022] 基于所述知识图谱中的三元组生成所述第一问题模式样本，所述第一问题模式样本所标注的主题标签为所述三元组中的谓语关系；

[0023] 将所述第一问题模式样本中用于本次训练的样本和其所标注的主题标签输入至所述文本分类模型中，并计算所述文本分类模型的交叉熵损失函数值；

[0024] 在所述交叉熵损失函数值不符合预设的第一结束条件的情况下，调整所述文本分类模型的权重参数，并返回执行所述将所述第一问题模式样本中用于本次训练的样本和其所标注的主题标签输入至所述文本分类模型中；

[0025] 在所述交叉熵损失函数值符合所述第一结束条件的情况下，将本次训练后的所述文本分类模型作为所述多标签分类模型。

[0026] 优选的，所述计算所述问题模式与所述候选实体的谓语关系间的语义相似度，包括：

[0027] 调用已训练的关系检测模型，所述关系检测模型是预先通过第二问题模式样本、以及为所述第二问题模式样本所标注的关系标签训练得到的；

[0028] 将所述问题模式与所述候选实体的谓语关系输入至所述关系检测模型中，通过所述关系检测模型获得所述问题模式与所述候选实体的谓语关系的语义相似度。

[0029] 优选的，所述关系检测模型的训练过程，包括：

[0030] 获取训练用的第二基础模型，所述第二基础模型包括第一编码层、第二编码层、分类模型和输出层；

[0031] 基于所述知识图谱中的三元组生成所述第二问题模式样本，所述第二问题模式样本包括正样本和负样本，所述正样本所标注的关系标签为所述三元组中的谓语关系，所述负样本所标注的关系标签非所述三元组中的谓语关系；

[0032] 按照预设比例分别对所述正样本和所述负样本进行样本采集，得到用于本次训练的样本；

[0033] 针对所述用于本次训练的样本，通过所述第一编码层生成该样本所标注的关系标签的嵌入向量，并将该样本所标注的关系标签的嵌入向量作为该样本所标注的关系标签的第一低维向量；

[0034] 通过所述第二编码层生成该样本中词组的嵌入向量；

[0035] 通过所述分类模型采用注意力机制处理所述词组的嵌入向量得到该样本的第二低维向量；

[0036] 通过所述输出层计算所述第一低维向量与所述第二低维向量的关联程度，并基于所述关联程度确定折页损失函数值；

[0037] 在所述折页损失函数值不符合预设的第二结束条件的情况下，基于所述折页损失函数值分别调整所述第一编码层、所述第二编码层和所述分类模型的权重参数，并返回执行所述按照预设比例分别对所述正样本和所述负样本进行样本采集，得到用于本次训练的样本；

[0038] 在所述折页损失函数值符合所述第二结束条件的情况下，将本次训练后的所述第二基础模型作为所述关系检测模型。

[0039] 一种知识图谱问答装置，所述装置包括：

[0040] 实体检测模块，用于接收问题语句，并识别问题语句的实体提及和问题模式；

[0041] 实体链接模块，用于在预设的知识图谱中检索与所述实体提及相关的主语实体作为候选实体；

[0042] 关系检测模块，用于获取所述知识图谱中所述候选实体的谓语关系，并计算所述问题模式与所述候选实体的谓语关系间的语义相似度；

[0043] 答案生成模块，用于将所述知识图谱中，语义相似度最大的候选实体和谓语关系所对应的宾语实体作为所述问题语句的答案。

[0044] 优选的，所述实体链接模块，具体用于：

[0045] 建立所述知识图谱中主语实体与所述主语实体的n-gram集合的反向映射索引，所述主语实体的n-gram集合中包含所述主语实体的所有组合方式；生成所述实体提及的n-gram集合，所述实体提及的n-gram集合包含所述实体提及的所有组合方式；采用启发式算法匹配所述实体提及的n-gram集合与所述主语实体的n-gram集合，基于所述反向映射索引将匹配到的主语实体作为候选实体。

[0046] 优选的，所述实体链接模块，还用于：

[0047] 调用已训练的多标签分类模型，所述多标签分类模型是预先通过第一问题模式样本、以及为所述第一问题模式样本所标注的主题标签训练得到的；将所述问题模式输入至所述多标签分类模型中，通过所述多标签分类模型获得所述问题模式所属主题的第一概率；确定所述候选实体的主题，并从所述第一概率中获取所述问题模式属于所述候选实体的主题的第二概率；计算所述候选实体与所述问题提及的编辑距离，并基于所述编辑距离和所述第二概率中的最大概率确定所述候选实体的评分；筛选所述候选实体中评分符合预设排名的实体。

[0048] 优选的，所述关系检测模块，具体用于：

[0049] 调用已训练的关系检测模型，所述关系检测模型是预先通过第二问题模式样本、以及为所述第二问题模式样本所标注的关系标签训练得到的；将所述问题模式与所述候选实体的谓语关系输入至所述关系检测模型中，通过所述关系检测模型获得所述问题模式与所述候选实体的谓语关系的语义相似度。

[0050] 本申请提供的知识图谱问答方法及装置，识别问题语句的实体提及和问题模式，并在知识图谱中检索与实体提及相关的主语实体作为候选实体，进而通过计算问题模式与候选实体的谓语关系的语义相似度来确定问题语句的答案。本申请能够对问题语句的问题模式和知识图谱的谓语关系进行语义的联合分析，从而识别出知识图谱中语义最相关的宾

语实体作为答案，从而提高问答结果的准确率。

附图说明

[0051] 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，下面描述中的附图仅仅是本发明的实施例，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，还可以根据提供的附图获得其他的附图。

[0052] 图1为本申请实施例提供的知识图谱问答方法的方法流程图；

[0053] 图2为本申请实施例提供的知识图谱问答方法的部分方法流程图；

[0054] 图3为本申请实施例提供的知识图谱问答方法的另一部分方法流程图；

[0055] 图4为本申请实施例提供的知识图谱问答方法的另一部分方法流程图；

[0056] 图5为本申请实施例提供的场景实施例的示意图；

[0057] 图6为本申请实施例提供的知识图谱问答装置的结构示意图。

具体实施方式

[0058] 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、完整地描述，显然，所描述的实施例仅仅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。

[0059] 为使本发明的上述目的、特征和优点能够更加明显易懂，下面结合附图和具体实施方式对本发明作进一步详细的说明。

[0060] 为方便理解本申请，以下首先对本申请相关的概率进行解释说明：

[0061] 1)知识图谱：知识图谱是一种语义网络，网络中的节点代表真实世界中的实体，网络中的边代表一种语义关系。知识图谱通常由大量三元组构成；

[0062] 2)三元组：一条三元组代表一条知识，例如三元组<姚明，身高，2米26>表示“姚明的身高是2米26”。三元组由主语实体、谓语关系以及宾语实体组成；

[0063] 3)主语实体：即三元组中的第一项，例如上述三元组中的“姚明”是主语实体；

[0064] 4)谓语关系：即三元组中的第二项，例如上述三元组中的“身高”是谓语关系；

[0065] 5)宾语实体：即三元组中的第三项，例如上述三元组中的“2米26”是宾语实体；

[0066] 6)管道模型：也称流水线模型，是指通过多个子模块的级联来解决一个问题，与端到端模型相对；

[0067] 7)实体提及：是指在自然语言句子中出现的有关实体的部分。例如给出问句“姚明的身高是多少?”，那么“姚明”便是问句中的实体提及。

[0068] 8)实体链接：将问句中的实体提及与知识图谱中的主语实体对应的过程即为实体链接；

[0069] 9)深度神经网络：是机器学习的分支，以人工神经网络为架构，对数据进行表征学习的算法；

[0070] 10)最大池化：对邻域内的特征点取最大值；

[0071] 11)实体识别：是指识别文本中具有特定意义的实体，主要包括人名、地名、机构名

等。例如给出问句“姚明的身高是多少?”,那么从问句中识别出“姚明”的过程就是命名实体识别;

[0072] 12)注意力机制:当人们注意到某个场景时,对该场景内每一处空间位置上的注意力分布是不一样的。在自然语言处理领域中,可以将注意力机制看作是为每个单词分配不同的权重,越重要的单词的权重应该越高;

[0073] 13)序列标注:属于自然语言处理领域的一种任务,即对输入的语句进行标注。例如可以标注词性,或标注具有意义的实体等;

[0074] 14)BiGRU模型:即双向门控神经网络。由两个单向门控神经网络构成,常用于自然语言处理过程中,对文本进行表征学习;

[0075] 15)CRF模型:即条件随机场。是一种无向图模型,在命名实体识别等序列标注任务中应用广泛;

[0076] 16)多标签分类:即一个样本可以被分类到多个不同的类别中,样本和类别可以具有一对多的关系;

[0077] 17)“BIO”标注模式:对一个序列的每个元素标注一个标签,“B”表示片段的开头,“I”表示片段的中间位置,“O”表示不属于任何类型。例如给出问句“姚明的身高是多少?”,如果采用“BIO”标注模式对实体进行标注,那么标注结果为“BIOOOOOOO”;

[0078] 18)n-gram模型:将一句话拆分成长度为n的连续片段。例如当n值为2时,按照字级别对问句“姚明的身高是多少?”进行拆分,就可以得到“姚明”、“明的”、“的身”、“身高”等片段;

[0079] 19)编辑距离:两个字符串之间,由一个转换成另外一个所需要的最少操作次数,允许的操作包括替换字符,增加字符,删除字符。两字符串的编辑距离越小,说明两字符串的相似度越高;反之,两字符串的编辑距离越大,说明两字符串的相似度越低;

[0080] 20)TextRNN、TextCNN、TextRCNN、TextRNN-Attention:均为基础的基于神经网络的文本分类算法;

[0081] 21)交叉熵损失函数:在分类模型中常见的衡量误差的函数;

[0082] 22)折页损失函数:用于衡量正样本和负样本之间差异的损失函数,如果正样本和负样本的差异越大说明模型效果越好。

[0083] 现阶段,单关系事实型问题面临着四方面的挑战:(1)一词多义现象,即相同的词汇或者短语在不同的上下文语境中会表达出完全不同的语义。例如“苹果”这个词即有可能表示一种水果,也可以表示科技公司;(2)歧义现象,即一个词汇或短语可以被链接到知识图谱中的不同实体。例如“芝加哥”这个词可以与知识图谱中的实体“芝加哥城市”相关联,同时也可以与实体“芝加哥公牛队”相关联;(3)语义鸿沟现象,即知识图谱中的一个谓语关系在问句中具有多种表述形式。例如知识图谱中的关系“出生于”可以被表述为“你的家乡在哪里?”或者“你出生于哪里?”;(4)实体重名现象,即随着知识图谱规模的增长,很多实体具有完全相同的名称,这使得无法从字面上对实体加以区分。

[0084] 对于端到端的神经网络方法处理单关系事实型问题,以下以申请号201910306552.8的专利《一种端到端的基于上下文的知识库问答方法与装置》来说明:

[0085] 其方法实施的主要步骤为:

[0086] (1)对自然语言问题进行预处理,过滤特殊字符;

[0087] (2)基于知识库构建与问题相关的候选主语实体集合,并根据候选实体在知识库中相关联的关系构建候选谓语关系集合;

[0088] (3)对于每个问题的候选主语实体集合中的每个实体,抽取实体在问题中的上下文;

[0089] (4)对于每个问题的候选谓语关系集合中的每个关系进行不同粒度的划分;

[0090] (5)基于CERM模型进行训练,通过训练数据学习主语实体的上下文表示和谓语关系的不同粒度的表示,使得正确的实体和正确的关系的相似度更高;在测试阶段,返回候选实体列表和候选关系列表中得分最高主语实体和谓语关系;

[0091] (6)利用预测的主语实体和谓语关系在知识库中找到宾语实体作为答案返回。

[0092] 其中,CERM模型包括:

[0093] 实体编码器单元:利用深度神经网络模型对实体的上下文进行序列建模,将候选实体转化为一个包含问题上下文语义的低维空间的分布式向量;

[0094] 关系编码器单元:将划分后的关系看作一个序列,利用深度神经网络将划分后的关系转化为包含关系语义的一个分布式向量;

[0095] 实体和关系得分列表单元:将一个自然语言问题的候选主语实体和候选谓语关系分别通过实体编码器和关系编码器得到的特征向量进行点积运算得到实体和关系的相似度矩阵,对矩阵分别进行行方向和列方向的最大池化操作得到关系相似度得分列表和实体相似度得分列表;

[0096] 实体和关系预测单元:在训练阶段,通过最小化对数归一化指数损失,使得候选实体和候选关系相似度得分列表中正确的实体和关系的得分更高;测试阶段,返回主语实体和谓语关系得分列表中得分最高的实体和关系。

[0097] 但是,以上技术方案具有以下两点缺陷:

[0098] (1)在实体链接过程中,该方法首先收集知识库中的实体标签名,形成待检索实体库,然后在上述实体库中检索与问题中单词或词组相匹配的实体名称,从而形成实体候选集。该方法的缺点在于首先没有在问句中进行命名实体识别,因此如果问句中的实体名称与知识图谱中的标准实体名称不完全一致,则会导致正确实体不会出现在实体候选集中。另外,该方法仅仅考虑了实体名称在字面上是否相同,而没有考虑语义上是否相关,因此如果知识图谱中出现多个重名实体,则该方法无法区分正确实体。

[0099] (2)在该方法使用的CERM模型中,无论是实体编码器还是关系编码器仅仅通过深度神经网络转化为低维空间的分布式向量。由于深度神经网络具有黑箱效应,该方法不能为预测结果给出合理的解释。

[0100] 本申请提供一种基于管道模型的知识图谱问答方法,该方法的方法流程图如图1所示,包括如下步骤:

[0101] S10,接收问题语句,并识别问题语句的实体提及和问题模式。

[0102] 本申请实施例中,该步骤为实体检测,可以将实体提及的识别任务视为序列标注任务。具体的,首先从问句样本中标注出具有实体含义的词组、以及普通词组,标注时采用“BIO”模式;进而使用标注好的问句样本训练BiGRU-CRF模型,将该模型作为后续识别问题语句的实体提及的工具。

[0103] 而在识别出问题语句中的实体提及后,将实体提及用通用符号,比如“head”替换,

可以得到该问题语句的问题模式。例如，给出问题语句“撒哈拉以南非洲的时区是多少?”，该问题语句的实体提及为“撒哈拉以南非洲”，相应的，该问题语句的问题模式为“<head>的时区是多少?”。

[0104] S20，在预设的知识图谱中检索与实体提及相关的主语实体作为候选实体。

[0105] 本申请实施例中，该步骤为实体链接，可以通过字符匹配的方法从知识图谱中检索出与实体提及相关的主语实体。

[0106] 具体实现过程中，步骤S20“在预设的知识图谱中检索与实体提及相关的主语实体作为候选实体”可以采用如下步骤，方法流程图如图2所示：

[0107] S201，建立知识图谱中主语实体与主语实体的n-gram集合的反向映射索引，主语实体的n-gram集合中包含主语实体的所有组合方式。

[0108] 本申请实施例中，可以收集知识图谱中的所有主语实体，形式主语实体库。针对该主语实体库中的每个主语实体，获得该主语实体的n-gram集合，并建立该主语实体与n-gram集合的映射关系。

[0109] 比如给出主语实体“贝拉克·侯赛因·奥巴马”，则其n-gram集合为{贝拉克，侯赛因，奥巴马，贝拉克·侯赛因，侯赛因·奥巴马，贝拉克·侯赛因·奥巴马}。

[0110] 需要说明的是，以上举例的n-gram集合的组合基准为一个词组，但在实际应用中，n-gram集合的组合基准可以为一个词组，还可以为一个字符，还可为一个空格，本实施例对此不做限定，可以根据实际场景进行设置。

[0111] S202，生成实体提及的n-gram集合，实体提及的n-gram集合包含实体提及的所有组合方式。

[0112] 本申请实施例中，生成实体提及的n-gram集合可以参见步骤S202，本实施例对此不做限定。需要说明的是，步骤S201与步骤S202所采用的组合基准相同。

[0113] S203，采用启发式算法匹配实体提及的n-gram集合与主语实体的n-gram集合，基于反向映射索引将匹配到的主语实体作为候选实体。

[0114] 本申请实施例中，为减少候选实体的规模，采用启发式算法，按照n值由大到小对实体提及的n-gram集合中的组合方式进行排序，优先匹配字符长度较长的组合，如果能够匹配到，则不考虑字符长度较短的组合。

[0115] 在此基础上，为解决无法区分知识图谱中名称相同的主语实体，进一步，本申请将根据实体提及的上下文，即问题模式进行处理，使正确的主语实体出现在候选实体中排名靠前的位置，从而缓解歧义现象以及重名现象。

[0116] 步骤S203之后，还可以采用如下步骤，方法流程图如图3所示：

[0117] S204，调用已训练的多标签分类模型，多标签分类模型是预先通过第一问题模式样本、以及为第一问题模式样本所标注的主题标签训练得到的。

[0118] 本申请实施例中，由于问题模式与主题间具有对应关系，例如给出问题模式“<head>的时区是多少?”，显然该问题模式的主题关乎于“时区”而非“职业”。因此，我们可以利用基于深度神经网络的文本分类模型对问题模式与主题的映射关系进行建模。

[0119] 此外，由于问题模式与主题间可能具有一对多的关系，我们将该问题转化成多标签分类问题。具体的文本分类模型可以考虑TextRNN、TextCNN、TextRCNN、TextRNN-Attention等基础模型或其它更为复杂的神经网络模型，在训练阶段将采用交叉熵损失函

数评估模型误差。

[0120] 具体的，多标签分类模型的训练过程，包括如下步骤：

[0121] 获取训练用的第一基础模型，第一基础模型为预设的文本分类模型；基于知识图谱中的三元组生成第一问题模式样本，第一问题模式样本所标注的主题标签为三元组中的谓语关系；将第一问题模式样本中用于本次训练的样本和其所标注的主题标签输入至文本分类模型中，并计算文本分类模型的交叉熵损失函数值；在交叉熵损失函数值不符合预设的第一结束条件的情况下，调整文本分类模型的权重参数，并返回执行将第一问题模式样本中用于本次训练的样本和其所标注的主题标签输入至文本分类模型中；在交叉熵损失函数值符合第一结束条件的情况下，将本次训练后的文本分类模型作为多标签分类模型。

[0122] 本申请实施例中，在生成用于训练多标签分类模型的第一问题模式样本时，可以随机从知识图谱中提取一定数量的三元组，将一个三元组中的主语实体用诸如“head”的通用符号替换即可得到一个问题模式样本，并将该三元组中的谓语关系作为该问题模式样本的主题标签。

[0123] 还需要说明的是，第一结束条件可以为一个交叉熵损失函数阈值。

[0124] S205，将问题模式输入至多标签分类模型中，通过多标签分类模型获得问题模式所属主题的第一概率。

[0125] 本申请实施例中，将问题模式输入到多标签分类模型中，由该多标签分类模型输出问题模式所属的主题、以及属于该主题的概率。

[0126] S206，确定候选实体的主题，并从第一概率中获取问题模式属于候选实体的主题的第二概率。

[0127] 本申请实施例中，对于每一个候选实体，其在知识图谱中可能具体多个谓语关系，即具有多个主题，因此，可以确定多标签分类模型输出的结果中被该候选实体所具有的主题的概率，也就是第二概率。

[0128] S207，计算候选实体与问题提及的编辑距离，并基于编辑距离和第二概率中的最大概率确定候选实体的评分。

[0129] 本申请实施例中，形式化定义如下，给出问题模式p、实体提及m、候选实体e_i、问题模式所属主题R_{ei}、多标签分类模型M。

[0130] 采用如下公式(1)计算候选实体的评分，考虑了字符相关性以及问题模式分类：

[0131] $$S_{EL}(e_i,p,m)=S_{ed}(e_i,m)+\max_{r\in R_{e_i}}(M(p)_r) \quad (1)$$

[0132] 其中，$S_{EL}(e_i,p,m)$为候选实体的评分，$S_{ed}(e_i,m)$为编辑距离，$\max_{r\in R_{e_i}}(M(p)_r)$则表示从问题模式p所属的候选实体的主题$r\in R_{e_i}$的概率(即第二概率)中选出最大值。

[0133] S208，筛选候选实体中评分符合预设排名的实体。

[0134] 本申请实施例中，按照由大到小的顺序对候选实体的评分进行排序，能够得到候选实体的排名，并保留其中top-N个实体作为最终的候选实体。

[0135] 由此，本申请实施例在实体链接的过程中，首先基于字符匹配的实体链接方法初步得到候选实体，进而根据问题模式与主题的映射关系建立多标签分类模型，并将该模型应用于实体链接过程中，从而为初步得到的候选实体额外分配额外的问题模式分类的评

分，最后通过重新排序，使得正确的实体能够出现在排名靠前的候选实体中，从而缓解歧义现象以及主语实体的同名现象。

[0136] S30，获取知识图谱中候选实体的谓语关系，并计算问题模式与候选实体的谓语关系间的语义相似度。

[0137] 本申请实施例中，该步骤为关系检测，具体的，可以构建基于注意力机制的关系检测模型来计算问题模式与候选实体的谓语关系间的余弦相似度，从而将该余弦相似度作为两者的语义相似度。

[0138] 具体实现过程中，步骤S30“计算问题模式与候选实体的谓语关系间的语义相似度”可以采用如下步骤：

[0139] 调用已训练的关系检测模型，关系检测模型是预先通过第二问题模式样本、以及为第二问题模式样本所标注的关系标签训练得到的；

[0140] 将问题模式与候选实体的谓语关系输入至关系检测模型中，通过关系检测模型获得问题模式与候选实体的谓语关系的语义相似度。

[0141] 其中，关系检测模型的训练过程，包括如下步骤：

[0142] 获取训练用的第二基础模型，第二基础模型包括第一编码层、第二编码层、分类模型和输出层；基于知识图谱中的三元组生成第二问题模式样本，第二问题模式样本包括正样本和负样本，正样本所标注的关系标签为三元组中的谓语关系，负样本所标注的关系标签非三元组中的谓语关系；按照预设比例分别对正样本和负样本进行样本采集，得到用于本次训练的样本；针对用于本次训练的样本，通过第一编码层生成该样本所标注的关系标签的嵌入向量，并将该样本所标注的关系标签的嵌入向量作为该样本所标注的关系标签的第一低维向量；通过第二编码层生成该样本中词组的嵌入向量；通过分类模型采用注意力机制处理词组的嵌入向量得到该样本的第二低维向量；通过输出层计算第一低维向量与第二低维向量的关联程度，并基于关联程度确定折页损失函数值；在折页损失函数值不符合预设的第二结束条件的情况下，基于折页损失函数值分别调整第一编码层、第二编码层和分类模型的权重参数，并返回执行按照预设比例分别对正样本和负样本进行样本采集，得到用于本次训练的样本；在折页损失函数值符合第二结束条件的情况下，将本次训练后的第二基础模型作为关系检测模型。

[0143] 参见图4所示的关系检测模型的训练示意图。通过第一编码层生成关系标签“时区”的嵌入向量，该嵌入向量作为其的低维向量表示，即第一低维向量。通过第二编码层生成样本“<head>的时区是多少?”的嵌入向量，并将该嵌入向量输入至分类模型“BiGRU模型”中，由BiGRU模型先基于该嵌入向量获得样本的隐含状态，进而采用注意力机制将样本的隐含状态转换为样本的低维向量表示，即第二低维向量。使用注意力机制能够为问题模式中每个词组分配不同的权重，通过可视化的注意力权重，能够提高模型的可解释性。

[0144] 具体的，采用以下公式(2)为注意力机制计算第二低维向量：

[0145] $$\begin{cases} v_p = \sum_{i=1}^{L} \alpha_i h_i \\ \alpha_i = \dfrac{\exp(w_i)}{\sum_{j=1}^{L} \exp(w_j)} \\ w_i = v^T \tanh(W v_r + U h_i) \end{cases} \quad (2)$$

[0146] 其中，h_i为样本的第i个词组的隐含状态$[h_1, h_2, \ldots, h_L]$中的一个，L为样本的词组长度，α_i为第i个词组的注意力权重，v、W、以及U为分类模型的权重参数。

[0147] 进一步，采用以下公式(3)计算第一低维向量v_r与第二低维向量v_p的余弦相似度，将该余弦相似度作为关联程度：

[0148] $S_{RD}(v_r, v_p) = \text{cosine}(v_r, v_p)$ (3)

[0149] 其中，$S_{RD}(v_r, v_p)$为余弦相似度。

[0150] 此外，可以按照如下公式(4)计算第二基础模型的折页损失函数值：

[0151] $$L(\theta) = \sum_{i=1}^{N} \sum_{j=1}^{M} \max[0, \gamma - S_{RD}(v_r^{(i)}, v_p) + S_{RD}(v_r^{(j)}, v_p)] \quad (4)$$

[0152] 其中，$L(\theta)$为折页损失函数值，N为正样本数量，M为每个正样本对应的负样本的数量，γ为可调节的超参数，$v_r^{(i)}$为正样本对应的第一低维向量，$v_r^{(j)}$为负样本对应的第一低维向量。

[0153] S40，将知识图谱中，语义相似度最大的候选实体和谓语关系所对应的宾语实体作为问题语句的答案。

[0154] 本申请实施例中，该步骤为答案生成，将语义相似度最大的一组候选实体和谓语关系分别作为最优主语实体和最优谓语关系，从而在知识图谱中检索出最优主语实体和最优谓语关系所在的三元组中的宾语实体作为答案。

[0155] 参见图5所示的场景实施例。接收到问题语句“撒哈拉以南非洲的时区是多少?”之后，通过实体检测识别其中的实体提及“撒哈拉以南非洲”、以及问题模式“<head>的时区是多少?”；通过实体链接确定知识图谱中与实体提及“撒哈拉以南非洲”相关的候选实体“m.04whzt2”和“m.06qtn”；进一步通过关系检测获取候选实体“m.04whzt2”和“m.06qtn”的谓语关系“类型、主题、名称、时区……”，并计算问题模式“<head>的时区是多少?”与谓语关系“类型、主题、名称、时区……”中各个关系的语义相似度，从而在知识图谱中确定语义相似度最大的一组候选实体“m.06qtn”和“时区”所在的三元组<m.06qtn，时区，欧洲西部夏令时间>，从而将“欧洲西部夏令时间”作为答案输出。

[0156] 本申请实施例提供的知识图谱问答方法，能够对问题语句的问题模式和知识图谱的谓语关系进行语义的联合分析，从而识别出知识图谱中语义最相关的宾语实体作为答案，从而提高问答结果的准确率。

[0157] 基于上述实施例提供的知识图谱问答方法，本申请实施例还提供一种执行上述知识图谱问答方法的装置，该装置的结构示意图如图6所示，包括：

[0158] 实体检测模块10，用于接收问题语句，并识别问题语句的实体提及和问题模式；

[0159] 实体链接模块20，用于在预设的知识图谱中检索与实体提及相关的主语实体作为

候选实体；

[0160] 关系检测模块30，用于获取知识图谱中候选实体的谓语关系，并计算问题模式与候选实体的谓语关系间的语义相似度；

[0161] 答案生成模块40，用于将知识图谱中，语义相似度最大的候选实体和谓语关系所对应的宾语实体作为问题语句的答案。

[0162] 可选的，实体链接模块20，具体用于：

[0163] 建立知识图谱中主语实体与主语实体的n-gram集合的反向映射索引，主语实体的n-gram集合中包含主语实体的所有组合方式；生成实体提及的n-gram集合，实体提及的n-gram集合包含实体提及的所有组合方式；采用启发式算法匹配实体提及的n-gram集合与主语实体的n-gram集合，基于反向映射索引将匹配到的主语实体作为候选实体。

[0164] 可选的，实体链接模块20，还用于：

[0165] 调用已训练的多标签分类模型，多标签分类模型是预先通过第一问题模式样本、以及为第一问题模式样本所标注的主题标签训练得到的；将问题模式输入至多标签分类模型中，通过多标签分类模型获得问题模式所属主题的第一概率；确定候选实体的主题，并从第一概率中获取问题模式属于候选实体的主题的第二概率；计算候选实体与问题提及的编辑距离，并基于编辑距离和第二概率中的最大概率确定候选实体的评分；筛选候选实体中评分符合预设排名的实体。

[0166] 可选的，实体链接模块20训练多标签分类模型的过程，包括：

[0167] 获取训练用的第一基础模型，第一基础模型为预设的文本分类模型；

[0168] 基于知识图谱中的三元组生成第一问题模式样本，第一问题模式样本所标注的主题标签为三元组中的谓语关系；将第一问题模式样本中用于本次训练的样本和其所标注的主题标签输入至文本分类模型中，并计算文本分类模型的交叉熵损失函数值；在交叉熵损失函数值不符合预设的第一结束条件的情况下，调整文本分类模型的权重参数，并返回执行将第一问题模式样本中用于本次训练的样本和其所标注的主题标签输入至文本分类模型中；在交叉熵损失函数值符合第一结束条件的情况下，将本次训练后的文本分类模型作为多标签分类模型。

[0169] 可选的，关系检测模块30，具体用于：

[0170] 调用已训练的关系检测模型，关系检测模型是预先通过第二问题模式样本、以及为第二问题模式样本所标注的关系标签训练得到的；将问题模式与候选实体的谓语关系输入至关系检测模型中，通过关系检测模型获得问题模式与候选实体的谓语关系的语义相似度。

[0171] 可选的，关系检测模块30训练关系检测模型的过程，包括：

[0172] 获取训练用的第二基础模型，第二基础模型包括第一编码层、第二编码层、分类模型和输出层；基于知识图谱中的三元组生成第二问题模式样本，第二问题模式样本包括正样本和负样本，正样本所标注的关系标签为三元组中的谓语关系，负样本所标注的关系标签非三元组中的谓语关系；按照预设比例分别对正样本和负样本进行样本采集，得到用于本次训练的样本；针对用于本次训练的样本，通过第一编码层生成该样本所标注的关系标签的嵌入向量，并将该样本所标注的关系标签的嵌入向量作为该样本所标注的关系标签的第一低维向量；通过第二编码层生成该样本中词组的嵌入向量；通过分类模型采用注意力

机制处理词组的嵌入向量得到该样本的第二低维向量;通过输出层计算第一低维向量与第二低维向量的关联程度,并基于关联程度确定折页损失函数值;在折页损失函数值不符合预设的第二结束条件的情况下,基于折页损失函数值分别调整第一编码层、第二编码层和分类模型的权重参数,并返回执行按照预设比例分别对正样本和负样本进行样本采集,得到用于本次训练的样本;在折页损失函数值符合第二结束条件的情况下,将本次训练后的第二基础模型作为关系检测模型。

[0173] 本申请实施例提供的知识图谱问答装置,能够对问题语句的问题模式和知识图谱的谓语关系进行语义的联合分析,从而识别出知识图谱中语义最相关的宾语实体作为答案,从而提高问答结果的准确率。

[0174] 以上对本发明所提供的一种知识图谱问答方法及装置进行了详细介绍,本文中应用了具体个例对本发明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发明的方法及其核心思想;同时,对于本领域的一般技术人员,依据本发明的思想,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发明的限制。

[0175] 需要说明的是,本说明书中的各个实施例均采用递进的方式描述,每个实施例重点说明的都是与其他实施例的不同之处,各个实施例之间相同相似的部分互相参见即可。对于实施例公开的装置而言,由于其与实施例公开的方法相对应,所以描述的比较简单,相关之处参见方法部分说明即可。

[0176] 还需要说明的是,在本文中,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、物品或者设备所固有的要素,或者是还包括为这些过程、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,并不排除在包括所述要素的过程、方法、物品或者设备中还存在另外的相同要素。

[0177] 对所公开的实施例的上述说明,使本领域专业技术人员能够实现或使用本发明。对这些实施例的多种修改对本领域的专业技术人员来说将是显而易见的,本文中所定义的一般原理可以在不脱离本发明的精神或范围的情况下,在其它实施例中实现。因此,本发明将不会被限制于本文所示的这些实施例,而是要符合与本文所公开的原理和新颖特点相一致的最宽的范围。

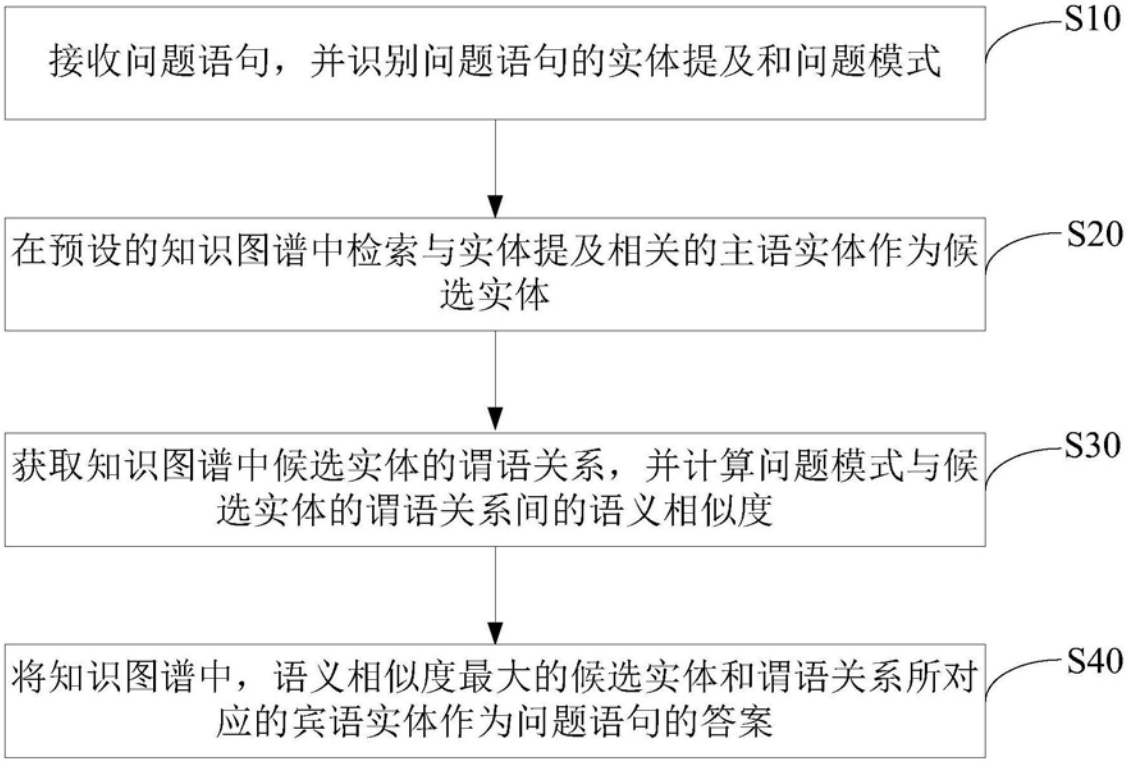

接收问题语句，并识别问题语句的实体提及和问题模式
S10
在预设的知识图谱中检索与实体提及相关的主语实体作为候选实体
S20
获取知识图谱中候选实体的谓语关系，并计算问题模式与候选实体的谓语关系间的语义相似度
S30
将知识图谱中，语义相似度最大的候选实体和谓语关系所对应的宾语实体作为问题语句的答案
S40

图1

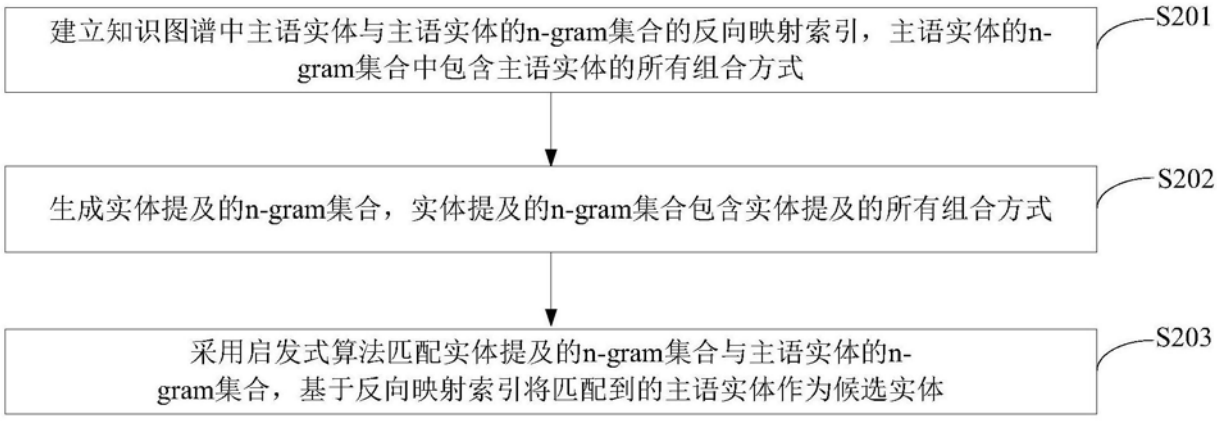

建立知识图谱中主语实体与主语实体的n-gram集合的反向映射索引，主语实体的n-gram集合中包含主语实体的所有组合方式
S201
生成实体提及的n-gram集合，实体提及的n-gram集合包含实体提及的所有组合方式
S202
采用启发式算法匹配实体提及的n-gram集合与主语实体的n-gram集合，基于反向映射索引将匹配到的主语实体作为候选实体
S203

图2

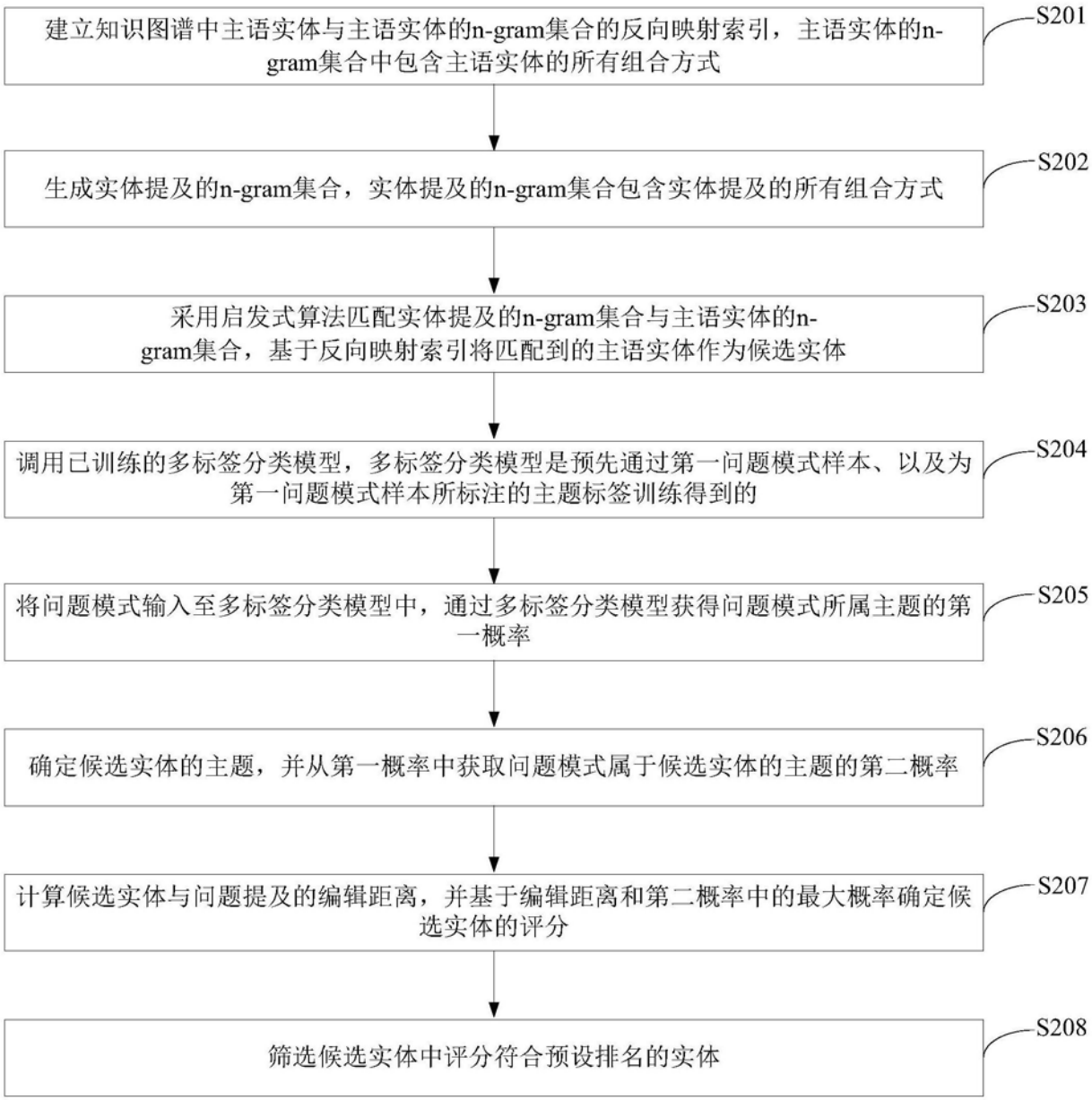

建立知识图谱中主语实体与主语实体的n-gram集合的反向映射索引，主语实体的n-gram集合中包含主语实体的所有组合方式
S201
生成实体提及的n-gram集合，实体提及的n-gram集合包含实体提及的所有组合方式
S202
采用启发式算法匹配实体提及的n-gram集合与主语实体的n-gram集合，基于反向映射索引将匹配到的主语实体作为候选实体
S203
调用已训练的多标签分类模型，多标签分类模型是预先通过第一问题模式样本、以及为第一问题模式样本所标注的主题标签训练得到的
S204
将问题模式输入至多标签分类模型中，通过多标签分类模型获得问题模式所属主题的第一概率
S205
确定候选实体的主题，并从第一概率中获取问题模式属于候选实体的主题的第二概率
S206
计算候选实体与问题提及的编辑距离，并基于编辑距离和第二概率中的最大概率确定候选实体的评分
S207
筛选候选实体中评分符合预设排名的实体
S208

图3

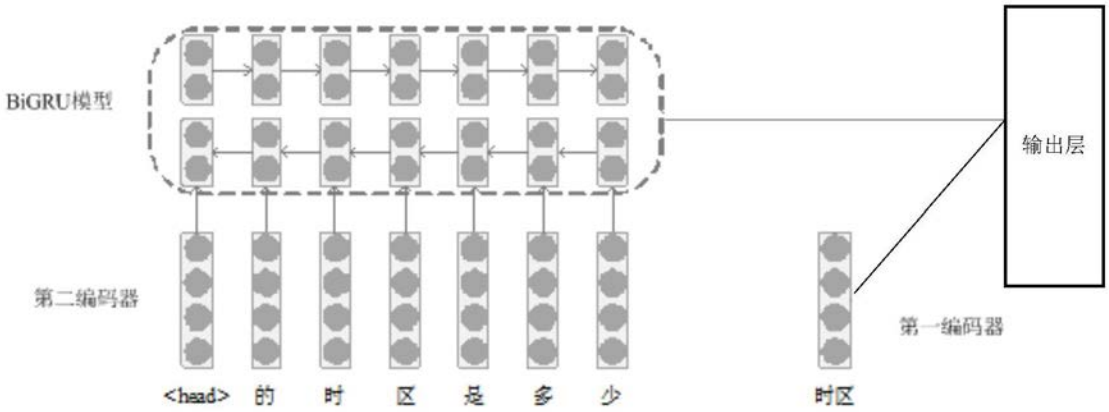

BiGRU模型
输出层
第二编码器
第一编码器
<head>
的
时
区
是
多
少
时区

图4

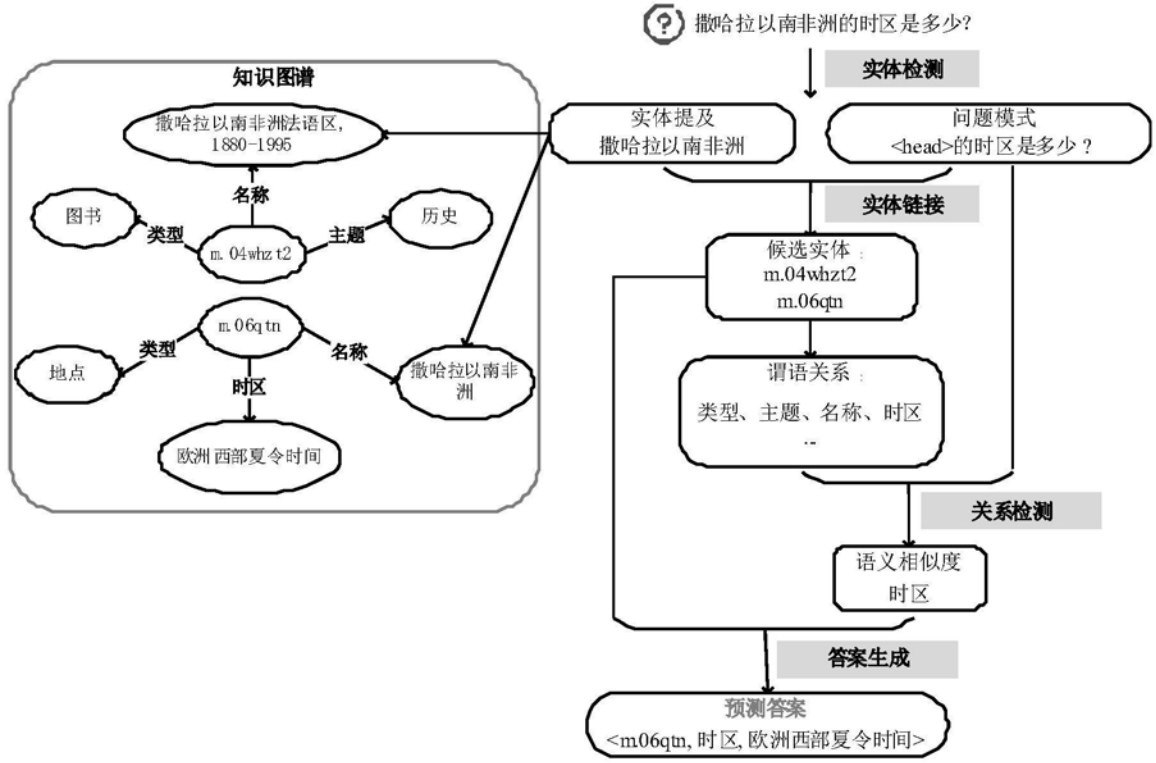

撒哈拉以南非洲的时区是多少?
实体检测
知识图谱
撒哈拉以南非洲法语区,
1880-1995
名称
图书
类型
m.04whzt2
主题
历史
m.06qtn
类型
名称
地点
时区
撒哈拉以南非洲
欧洲西部夏令时间
实体提及
撒哈拉以南非洲
问题模式
<head>的时区是多少?
实体链接
候选实体:
m.04whzt2
m.06qtn
谓语关系:
类型、主题、名称、时区
...
关系检测
语义相似度
时区
答案生成
预测答案
<m.06qtn, 时区, 欧洲西部夏令时间>

图5

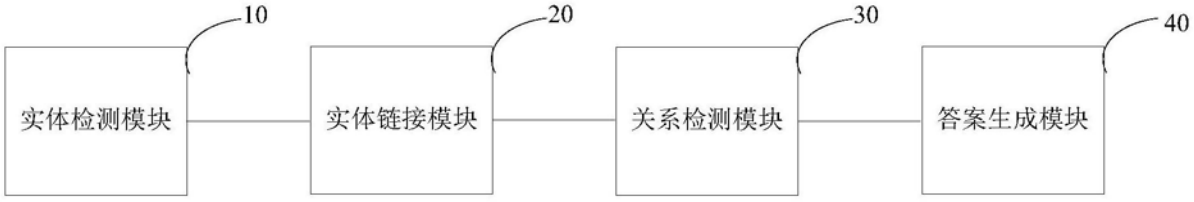

10
实体检测模块
20
实体链接模块
30
关系检测模块
40
答案生成模块

图6