(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 118673923 A
(43)申请公布日 2024.09.20

(21)申请号 202411139659.5
(22)申请日 2024.08.20
(71)申请人 安徽思高智能科技有限公司
地址 230088 安徽省合肥市高新区望江西路900号中安创谷科技园A1栋408
(72)发明人 袁水平 余鳌 刘宇 王健 张泽锟
(74)专利代理机构 武汉知产时代知识产权代理有限公司 42238
专利代理师 程泽
(51)Int.Cl.
G06F 40/295(2020.01)
G06F 40/30(2020.01)
G06F 16/332(2019.01)
G06F 16/33(2019.01)
G06N 5/04(2023.01)

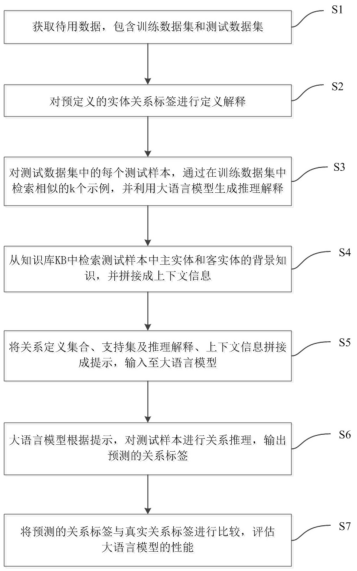
权利要求书2页 说明书4页 附图2页

(54)发明名称

一种基于大语言模型的实体关系抽取方法

(57)摘要

本发明公开了一种基于大语言模型的实体关系抽取方法,旨在提升关系抽取任务的泛化性能。该方法首先准备包含训练数据集和测试数据集的数据,并利用预训练词嵌入模型对文本编码。然后使用大语言模型对预定义的实体关系标签生成定义描述,帮助模型理解各类关系涵义。对于测试数据集中的每个样本,通过计算与训练样本的相似度,检索得到k个最相似的支持集样本。并对支持集样本利用大语言模型生成关系成立的推理解释。同时,从知识库中检索测试样本中主客体实体的背景知识。最后,将关系定义、支持集样本及推理解释、背景知识等信息构建为提示,输入大语言模型进行关系推理,得到预测结果。



CN 118673923 A

1.一种基于大语言模型的实体关系抽取方法,其特征在于:包括:

S1:获取待用数据,包含训练数据集D_{train}和测试数据集D_{test};

S2:对预定义的实体关系标签进行定义解释;

S3:对测试数据集D_{test}中的每个测试样本($content_i$, e_{si}, r_i, e_{oi}),通过在训练数据集D_{train}中检索相似的k个示例,并利用大语言模型生成推理解释;其中,r_i表示测试样本的真实关系标签,e_{si}表示测试样本的主实体,e_{oi}表示测试样本的客实体,$content_i$表示测试样本的原始文本,i=1,2,3,...,m,m为正整数;

S4:从知识库KB中检索测试样本中主实体e_{si}和客实体e_{oi}的背景知识,并拼接成上下文信息c_i;

S5:将关系定义集合D_{rel}、支持集S_i及推理解释E_i、上下文信息c_i拼接成提示p_i,输入至大语言模型;

S6:大语言模型根据提示p_i,对测试样本($content_i$,e_{si},r_i,e_{oi})进行关系推理,输出预测的关系标签y_i';

S7:将预测的关系标签y_i'与真实关系标签r_i进行比较,评估大语言模型的性能,当达到预设精度后,得到最终的大语言模型,利用最终的大语言模型可以实现对任一实体关系的抽取。

2.如权利要求1所述的一种基于大语言模型的实体关系抽取方法,其特征在于:

步骤S1的实现过程为:

S1.1:训练数据集D_{train}包含n个训练样本($content_j$, e_{sj}, r_j, e_{oj}),其中,$content_j$表示训练样本的原始文本,e_{sj}表示训练样本的主实体,e_{oj}表示训练样本的客实体,r_j表示训练样本的真实关系标签,j=1,2,3,...,n,$r_j \in R$,R为预定义的实体关系标签集合;

S1.2:测试数据集D_{test}包含m个测试样本($content_i$, e_{si}, r_i, e_{oi}),格式与训练数据集相同。

3.如权利要求1所述的一种基于大语言模型的实体关系抽取方法,其特征在于:步骤S2的实现过程为:

S2.1:对每个测试样本的真实关系标签r_i,构造提示"请解释关系<r_i>的定义和判断条件",输入大语言模型生成响应d_i;

S2.2:将d_i与真实关系标签r_i组成元组(r_i,d_i),得到关系定义集合D_{rel}。

4.如权利要求1所述的一种基于大语言模型的实体关系抽取方法,其特征在于:步骤S3中的实现过程为:

S3.1:对训练数据集和测试数据集中的样本进行预处理和词嵌入处理,样本包括训练样本和测试样本;

S3.2:对每个测试样本($content_i$, e_{si}, r_i, e_{oi}),计算其与所有训练样本($content_j$, e_{sj}, r_j, e_{oj})的三元组相似度sim_tri(($content_i$, e_{si}, r_i, e_{oi}), ($content_j$, e_{sj}, r_j, e_{oj})),选取相似度最高的k个样本作为支持集S_i;三元组相似度的计算公式为:

$$sim\_tri((content_i, e_{si}, r_i, e_{oi}), (content_j, e_{sj}, r_j, e_{oj})) = \cos(v_{si}', v_{sj}') + \cos(v_{oi}', v_{oj}')$$

其中,v_{si}'为测试样本($content_i$, e_{si}, r_i, e_{oi})的主实体和上下文的拼接嵌入,v_{oi}'为测试样本($content_i$, e_{si}, r_i, e_{oi})的客实体和上下文的拼接嵌入,v_{sj}'为训练样本

($content_j$, e_{sj}, r_j, e_{oj})的主实体和上下文拼接嵌入,v_{oj}'为训练样本($content_j$, e_{sj}, r_j, e_{oj})的客实体和上下文的拼接嵌入,cos(•)为余弦相似度函数;

S3.3:对S_i中每个样本($content_{ik}$,s_{ik},r_{ik},o_{ik}),构造提示并输入至大语言模型,生成关系(s_{ik},r_{ik},o_{ik})成立的推理解释E_i,其中,$content_{ik}$表示支持集中样本的原始文本,s_{ik}表示支持集中样本的主客体,r_{ik}表示支持集中样本的真实关系标签,o_{ik}表示支持集中样本的客实体;

S3.4:将S_i与对应的推理解释E_i组成测试样本的注释数据D_i。

5.如权利要求4所述的一种基于大语言模型的实体关系抽取方法,其特征在于:步骤S3.1的实现过程为:

S3.1.1:使用预训练词嵌入模型对样本的原始文本进行编码,得到上下文嵌入向量v_c;对主实体和客实体进行类似的编码,得到实体嵌入v_s和v_o;

S3.1.2:将实体嵌入和上下文嵌入进行拼接,得到增强的实体表示$v_s'=[v_s;v_c]$,$v_o'=[v_o;v_c]$。

6.如权利要求1所述的一种基于大语言模型的实体关系抽取方法,其特征在于:

步骤S4的实现过程为:

S4.1:知识库KB由SPO三元组(实体,属性,属性值)构成,对每个待检索实体e,在KB中找到其对应的所有属性值,记为KB(e);

S4.2:对主实体e_{si}和客实体e_{oi},检索背景知识KB(e_{si})和KB(e_{oi}),将KB(e_{si})和KB(e_{oi})拼接为上下文信息c_i。

7.如权利要求1所述的一种基于大语言模型的实体关系抽取方法,其特征在于:步骤S6的实现过程为:

S6.1:将提示p_i输入至大语言模型,生成回复a_i;

S6.2:通过设计一个映射函数$f:a_i \rightarrow y_i'$,从大语言模型生成的回复a_i中进一步提取预测的关系标签,记为y_i'。

8.一种存储设备,其特征在于:所述存储设备存储指令及数据用于实现权利要求1~7任一项所述的基于大语言模型的实体关系抽取方法。

9.一种基于大语言模型的实体关系抽取设备,其特征在于:包括:处理器及存储设备;所述处理器加载并执行所述存储设备中的指令及数据用于实现权利要求1~7任一项所述的基于大语言模型的实体关系抽取方法。

一种基于大语言模型的实体关系抽取方法

技术领域

[0001] 本发明属于自然语言处理(NLP)和人工智能(AI)技术领域,特别是一种基于大型语言模型的实体关系抽取方法。

背景技术

[0002] 关系抽取旨在从文本中识别预定义的实体对之间的语义关系,是知识获取和自然语言理解的核心任务。早期的关系抽取系统主要采用基于模式匹配、启发式规则等方法,但难以应对语言表达的多样性和复杂性。近年来,随着深度学习技术的发展,一系列基于神经网络的关系抽取模型被相继提出,如CNN、RNN、Transformer等,大幅提升了关系抽取的性能。然而,这些模型通常需要大量高质量的标注数据进行训练,在实际应用中受到标注成本和领域适应性的限制。

[0003] 近年来,大型预训练语言模型(如BERT、GPT等)在多项NLP任务上取得了显著成就。尤其是GPT-4等大规模生成式语言模型,展现出了惊人的上下文学习能力(in-context learning),即只需提供少量演示样本作为上下文,即能在新样本上进行关系抽取。然而,尽管GPT-4在许多NLP任务上取得了显著成就,但在关系抽取方面的表现仍显著落后于使用大量标注数据微调的模型(如BERT)。这主要是当前基于大语言模型关系抽取任务提示构造缺乏相应背景信息以及推理路径指导。现有方法在利用大语言模型进行关系抽取时仍存在检索质量不高、推理能力弱等局限。

发明内容

[0004] 本发明提出了一种基于大语言模型的实体关系抽取方法,主要通过标签关系定义、支持集样本及推理解释、背景知识等信息来构建更好的提示,从而提升大语言模型在关系抽取任务中的表现。

[0005] 一种基于大语言模型的实体关系抽取方法,包括:

S1:获取待用数据,包含训练数据集D_{train}和测试数据集D_{test};

S2:对预定义的实体关系标签进行定义解释;

S3:对测试数据集D_{test}中的每个测试样本($content_i$, e_{si}, r_i, e_{oi}),通过在训练数据集D_{train}中检索相似的k个示例,并利用大语言模型生成推理解释;其中,r_i表示测试样本的真实关系标签,e_{si}表示测试样本的主实体,e_{oi}表示测试样本的客实体,$content_i$表示测试样本的原始文本,i=1,2,3,...,m,m为正整数;

S4:从知识库KB中检索测试样本中主实体e_{si}和客实体e_{oi}的背景知识,并拼接成上下文信息c_i;

S5:将关系定义集合D_{rel}、支持集S_i及推理解释E_i、上下文信息c_i拼接成提示p_i,输入至大语言模型;

S6:大语言模型根据提示p_i,对测试样本($content_i$,e_{si},r_i,e_{oi})进行关系推理,输出预测的关系标签y_i';

S7:将预测的关系标签y_i'与真实关系标签r_i进行比较,评估大语言模型的性能,当达到预设精度后,得到最终的大语言模型,利用最终的大语言模型可以实现对任一实体关系的抽取。

[0006] 一种存储设备,所述存储设备存储指令及数据用于实现一种基于大语言模型的实体关系抽取方法。

[0007] 一种基于大语言模型的实体关系抽取设备,包括:处理器及所述存储设备;所述处理器加载并执行所述存储设备中的指令及数据用于实现一种基于大语言模型的实体关系抽取方法。

[0008] 本发明提供的技术方案带来的有益效果是:本发明通过引入关系定义解释、支持集检索与推理解释、实体背景知识等丰富信息,构建了一种新颖的提示学习范式。该方法利用大语言模型强大的语言理解和生成能力,在只给定少量示例的情况下,即可实现对新样本的关系推理。与传统方法相比,本发明不依赖大规模人工标注数据,大大降低了数据标注成本;通过从多角度引入先验知识,增强了模型对实体关系的语义理解能力;采用更加自然的提示形式,使得模型能够根据具体语境进行推理决策,提高了方法的泛化性和鲁棒性。此外,本发明生成的自然语言推理解释也大大增强了模型预测的可解释性,有助于人们理解模型的判断依据。总之,本发明提供了一种数据高效、泛化性强、可解释性好的实体关系抽取新方法,有望在知识图谱构建、智能问答、语义搜索等领域得到广泛应用。该方法巧妙地将关系定义解释、支持集示例推理、实体背景知识等多源信息融入提示学习范式,充分利用了大语言模型的语言理解和生成能力,实现了高效关系抽取。

附图说明

[0009] 下面将结合附图及实施例对本发明作进一步说明,附图中:

图1是本发明基于大语言模型的实体关系抽取方法的流程图;

图2是本发明的硬件设备工作的示意图。

具体实施方式

[0010] 为了对本发明的技术特征、目的和效果有更加清楚的理解,现对照附图详细说明本发明的具体实施方式。

[0011] 如图1所示,本发明提供了一种基于大语言模型的实体关系抽取方法,主要包括以下步骤:

S1:数据准备。准备包含训练数据集D_{train}和测试数据集D_{test}的数据。

[0012] 步骤S1的实现过程为:

S1.1:训练数据集D_{train}包含n个训练样本($content_j$, e_{sj}, r_j, e_{oj}),其中,$content_i$为训练样本的原始文本,e_{sj}表示训练样本的主实体,e_{oj}表示训练样本的客实体,r_j表示训练样本的真实关系标签,j=1,2,3,...,n,n为正整数;

S1.2:测试数据集D_{test}包含m个测试样本($content_i$, e_{si}, r_i, e_{oi}),格式与训练数据集相同。其中,$content_i$表示测试样本的原始文本,r_i表示测试样本的真实关系标签,e_{si}表示测试样本的主实体,e_{oi}表示测试样本的客实体,i=1,2,3,...,m,m为正整数,r_i,r_j∈R,R为预定义的关系标签集合。关系标签r_i运用输出性能评测模型性能。

[0013] S2:对预定义的实体关系标签进行定义解释。利用大语言模型生成每个关系标签$r_i \in R$的定义描述d_i,使模型理解各类关系的涵义。

[0014] 步骤S2的实现过程为:

S2.1:对每个测试样本的真实关系标签r_i,构造提示"请解释关系<r_i>的定义和判断条件",输入大语言模型生成响应d_i。

[0015] S2.2:将生成的响应d_i与真实关系标签r_i组成元组(r_i, d_i),得到关系定义集合D_{rel}。

[0016] S3:对测试数据集D_{test}中的每个测试样本(content, e_s, r, e_o),通过在训练数据集D_{train}中检索相似的k个示例,并利用大语言模型生成推理解释。其中,r表示预定义的某一实体关系标签,e_s表示主实体,e_o表示客实体,content表示测试样本的原始文本。

[0017] 步骤S3的实现过程为:

S3.1:对训练数据集和测试数据集中的原始文本content、主实体e_s、客实体e_o进行预处理和词嵌入处理。

[0018] S3.1.1:使用预训练词嵌入模型(如Word2Vec、GloVe等)对原始文本content进行编码,得到上下文嵌入向量v_c。对主实体e_s和客实体e_o进行类似的编码,得到实体嵌入v_s和v_o。

[0019] S3.1.2:将实体嵌入和上下文嵌入进行拼接,得到增强的实体表示$v_s' = [v_s; v_c]$,$v_o' = [v_o; v_c]$。

[0020] S3.2:对每个测试样本$(content_i, e_{si}, r_i, e_{oi}) \in Dtest$,计算其与所有训练样本$(content_j, e_{sj}, r_j, e_{oj}) \in D_{train}$的三元组相似度$sim_tri((content_i, e_{si}, r_i, e_{oi}), (content_j, e_{sj}, r_j, e_{oj}))$,选取相似度最高的k个样本作为支持集$S_i$。三元组相似度的计算公式为:

$$sim\_tri((content_i, e_{si}, r_i, e_{oi}), (content_j, e_{sj}, r_j, e_{oj})) = \cos(v_{si}', v_{sj}') + \cos(v_{oi}', v_{oj}')$$

其中,v_{si}'为测试样本$(content_i, e_{si}, r_i, e_{oi})$的主实体和上下文的拼接嵌入,$v_{oi}'$为测试样本$(content_i, e_{si}, r_i, e_{oi})$的客实体和上下文的拼接嵌入,$v_{sj}'$为训练样本$(content_j, e_{sj}, r_j, e_{oj})$的主实体和上下文拼接嵌入,$v_{oj}'$为训练样本$(content_j, e_{sj}, r_j, e_{oj})$的客实体和上下文的拼接嵌入,$\cos(\cdot)$为余弦相似度函数。

[0021] S3.3:对S_i中每个样本$(content_{ik}, s_{ik}, r_{ik}, o_{ik})$,构造提示并输入大语言模型,生成关系$(s_{ik}, r_{ik}, o_{ik})$成立的推理解释$E_i$。其中,$content_{ik}$表示支持集中样本的原始文本,$s_{ik}$表示支持集中样本的主客体,$r_{ik}$表示支持集中样本的真实关系标签,$o_{ik}$表示支持集中样本的客实体。

[0022] S3.4:将S_i与对应的推理解释E_i组成测试样本的注释数据D_i。注释数据D_i包含了对测试样本预测关系的解释性信息,支持集S_i给出了与测试样本相似的训练样本,推理解释E_i则说明了这些相似样本所蕴含关系成立的原因。这些解释性信息可以帮助使用者理解和验证模型预测关系的依据。

[0023] S4:从知识库KB中检索测试样本中主实体e_{si}和客实体e_{oi}的背景知识,并拼接成上下文信息c_i。

[0024] 步骤S4的实现过程为:

S4.1:知识库KB由SPO三元组(实体,属性,属性值)构成。对每个待检索实体e,在KB中找到其对应的所有属性值,记为KB(e)。

[0025] S4.2:对主实体e_{si}和客实体e_{oi},检索背景知识KB(e_{si})和KB(e_{oi}),将KB(e_{si})和KB(e_{oi})拼接为上下文信息c_i。

[0026] S5:将关系定义D_{rel}、支持集S_i及推理解释E_i、上下文信息c_i等信息拼接组织成提示p_i,输入给大语言模型。提示p_i设置为:拼接(D_{rel}, S_i, E_i, c_i, $content_i$, e_{si}, e_{oi})。

[0027] S6:大语言模型根据提示p_i,对测试样本($content_i$, e_{si}, r_i, e_{oi})进行关系推理,输出预测结果y_i'。

[0028] 步骤S6的实现过程为:

S6.1:将提示p_i输入大语言模型,生成回复a_i。

[0029] S6.2:从大语言模型生成的回复a_i中进一步提取预测的关系标签,记为y_i'。可以通过设计一个映射函数f:$a_i \rightarrow y_i$'来实现,例如基于关键词匹配、正则表达式等方法从a_i中抽取出预测的关系标签y_i'。

[0030] S7:评估大语言模型性能。将预测的关系标签y_i'与测试数据集中真实关系标签r_i进行比较,计算准确率、精确率、召回率、F1值等评价指标。若是,当前模型性能不佳,可以采用同样的方法继续进行优化,直到实体关系抽取的精度达到预设精度,得到最终的大语言模型,利用最终的大语言模型可以实现对任一实体关系的抽取。

[0031] 请参见图2,图2是本发明实施例的硬件设备工作示意图,所述硬件设备具体包括:一种基于大语言模型的实体关系抽取设备201、处理器202及存储设备203。

[0032] 一种基于大语言模型的实体关系抽取设备201:所述一种基于大语言模型的实体关系抽取设备201实现所述一种基于大语言模型的实体关系抽取方法。

[0033] 处理器202:所述处理器202加载并执行所述存储设备203中的指令及数据用于实现所述一种基于大语言模型的实体关系抽取方法。

[0034] 存储设备203:所述存储设备203存储指令及数据;所述存储设备203用于实现所述一种基于大语言模型的实体关系抽取方法。

[0035] 以上所述仅为本发明的较佳实施例,并不用以限制本发明,凡在本发明的精神和原则之内,所作的任何修改、等同替换、改进等,均应包含在本发明的保护范围之内。

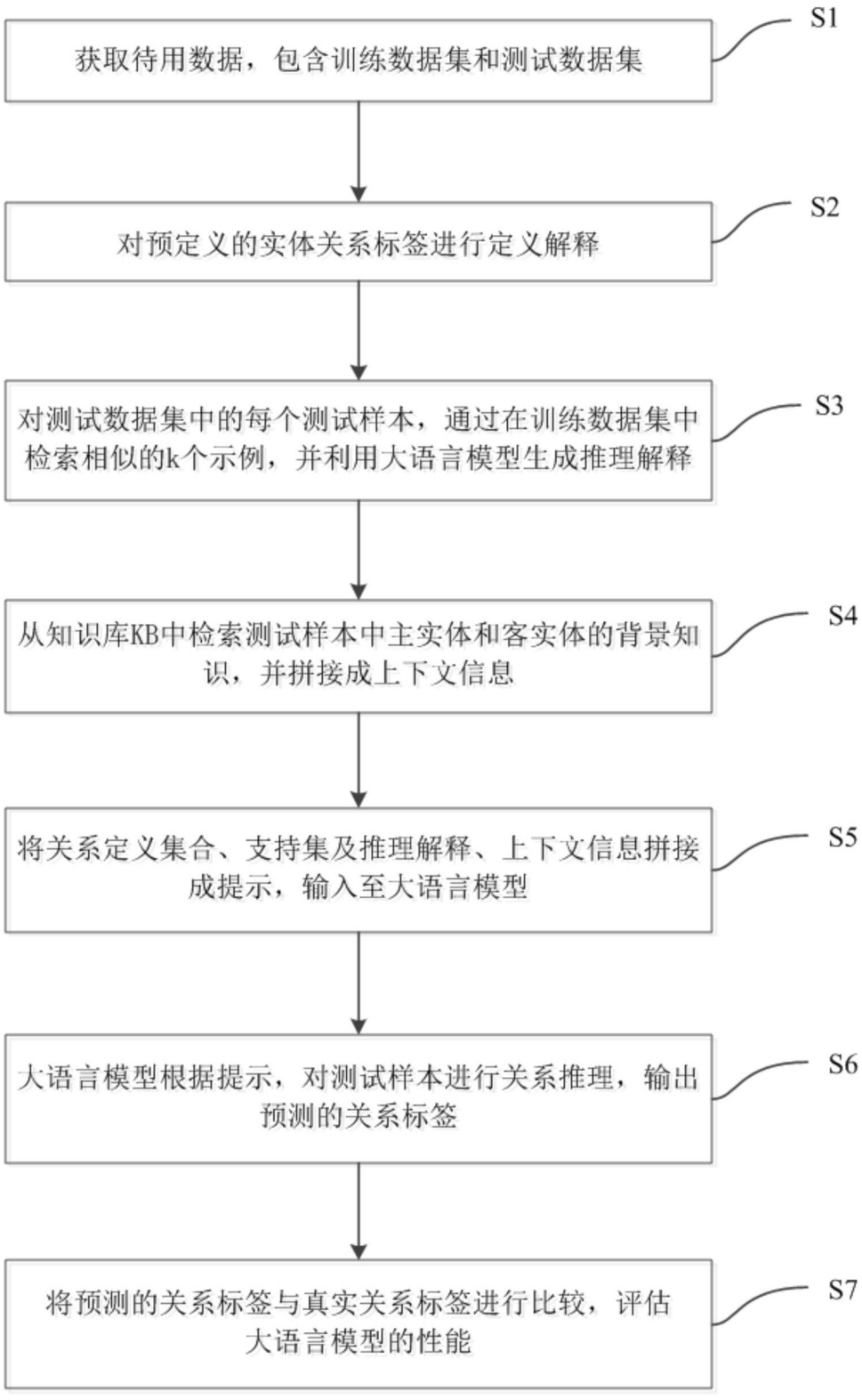

S1
获取待用数据，包含训练数据集和测试数据集
S2
对预定义的实体关系标签进行定义解释
S3
对测试数据集中的每个测试样本，通过在训练数据集中检索相似的k个示例，并利用大语言模型生成推理解释
S4
从知识库KB中检索测试样本中主实体和客实体的背景知识，并拼接成上下文信息
S5
将关系定义集合、支持集及推理解释、上下文信息拼接成提示，输入至大语言模型
S6
大语言模型根据提示，对测试样本进行关系推理，输出预测的关系标签
S7
将预测的关系标签与真实关系标签进行比较，评估大语言模型的性能

图1

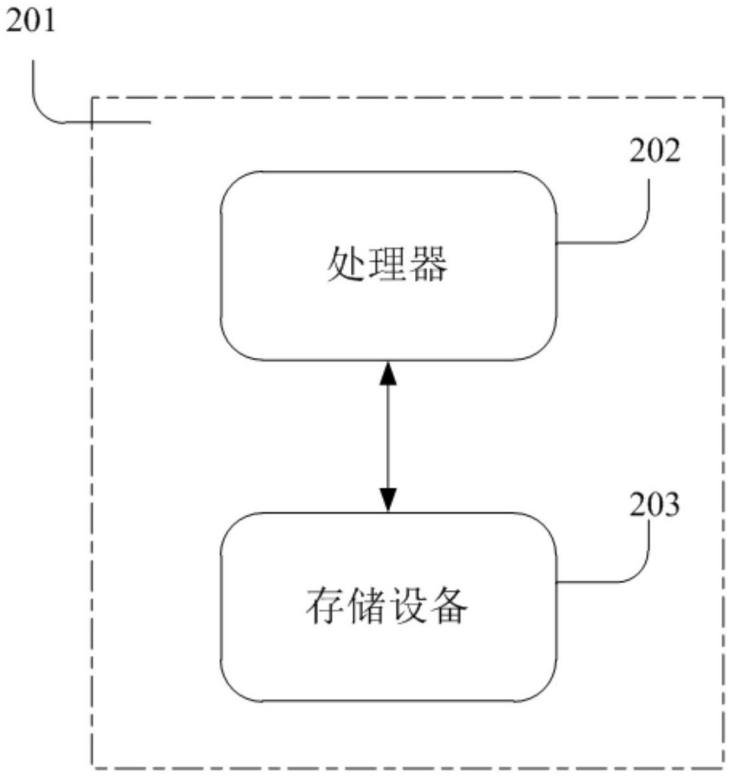


图2