

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2024 年 8 月 15 日 (15.08.2024)



(10) 国际公布号
WO 2024/164580 A1

- (51) 国际专利分类号:
G06F 40/30 (2020.01) **G06F 18/22** (2023.01)
G06F 40/211 (2020.01)
- (21) 国际申请号: PCT/CN2023/128479
- (22) 国际申请日: 2023 年 10 月 31 日 (31.10.2023)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
202310085688.7 2023年2月9日 (09.02.2023) CN
- (71) 申请人: 中南大学 (CENTRAL SOUTH UNIVERSITY) [CN/CN]; 中国湖南省长沙市岳麓区麓山南路932号, Hunan 410012 (CN)。
- (72) 发明人: 龙军 (LONG, Jun); 中国湖南省长沙市岳麓区麓山南路932号, Hunan 410012 (CN)。 杨柳 (YANG, Liu); 中国湖南省长沙市岳麓区麓山南路932号, Hunan 410012 (CN)。 齐倩倩 (QI, Qianqian);

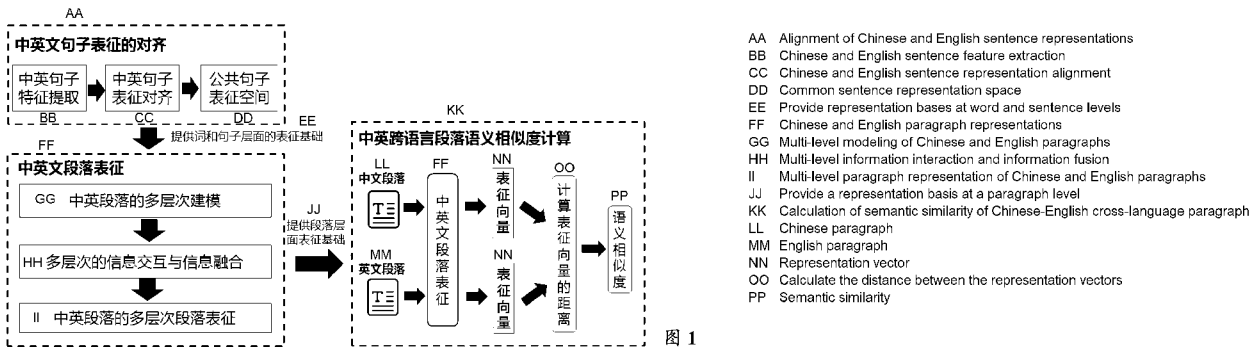
中国湖南省长沙市岳麓区麓山南路932号, Hunan 410012 (CN)。 唐自强 (TANG, Ziqiang); 中国湖南省长沙市岳麓区麓山南路932号, Hunan 410012 (CN)。 黄金彩 (HUANG, Jincai); 中国湖南省长沙市岳麓区麓山南路932号, Hunan 410012 (CN)。

(74) 代理人: 北京高沃律师事务所 (BEIJING GAOWO LAW FIRM); 中国北京市海淀区彩和坊路11号6层602室, Beijing 100080 (CN)。

(81) 指定国(除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD,

(54) Title: METHOD FOR CALCULATING CHINESE-ENGLISH SEMANTIC SIMILARITY OF PARAGRAPH-LEVEL TEXTS

(54) 发明名称: 面向段落级文本的中英文语义相似度计算方法



(57) Abstract: Disclosed in the present invention is a method for calculating the Chinese-English semantic similarity of paragraph-level texts, comprising: respectively performing paragraph representation vector extraction on a Chinese paragraph and an English paragraph; modeling paragraph texts in terms of three levels of topic words, sentences, and paragraphs, and performing information interaction within and among the levels on the basis of a graph attention network; then fusing information of the three levels of topic words, sentences, and paragraphs to obtain paragraph representation vectors; and calculating the distance between the paragraph representation vectors to obtain the semantic similarity of the Chinese paragraph and the English paragraph. According to the method, information of three levels of topic words, sentences, and paragraphs are fused to obtain high-quality paragraph representation vectors, thereby achieving high-accuracy calculation of the semantic similarity of Chinese-English cross-language paragraphs.

(57) 摘要: 本发明公开了一种面向段落级文本的中英文语义相似度计算方法, 其分别对中文段落和英文段落进行段落表征向量提取, 从主题词、句子和段落三个层次对段落文本进行建模, 并基于图注意力网络在各个层次内部和各个层次信息之间进行信息交互, 然后融合主题词、句子和段落三个层次的信息得到段落表征向量, 通过计算段落表征向量的距离得到中文段落和英文段落的语义相似度。该方法通过融合主题词、句子和段落三个层次的信息得到高质量的段落表征向量, 能实现中英跨语言段落语义相似度的高精度计算。



SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ,
UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW。

- (84) 指定国(除另有指明，要求每一种可提供的地区
保护): ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ,
NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚
(AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE,
BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR,
HU, IE, IS, IT, LT, LU, LV, MC, ME, MK, MT, NL, NO,
PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF,
CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN,
TD, TG)。

本国际公布:

- 包括国际检索报告(条约第21条(3))。

面向段落级文本的中英文语义相似度计算方法

本申请要求于 2023 年 02 月 09 日提交中国专利局、申请号为 202310085688.7、发明名称为“面向段落级文本的中英文语义相似度计算方法”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

5 技术领域

本发明涉及信息技术领域，特别是涉及一种面向段落级文本的中英文语义相似度计算方法。

背景技术

现在，随着世界文化交流程度的加强以及各国跨语言之间的资源共享，
10 跨语言场景变得越来越普遍，这导致了跨语言应用需求的越发迫切。为了解决由此带来的技术壁垒，实现跨语言之间的资源共享，学术界和工业界一直都在积极探索跨语言的自然语言处理技术。而其中，跨语言段落语义相似度计算就是一项重要的研究内容，它研究的是如何对不同语言的段落级文本进行语义相似度计算，这在许多跨语言处理应用及其相关领域中扮演
15 着重要角色，如对不同语言申请书的摘要进行语义相似度计算，完成相似申请书的检测和遴选；对不同语言论文的摘要进行语义相似度计算，从而实现相似论文检测，完成跨语言论文推荐和跨语言抄袭检测。

现阶段的中英跨语言段落语义相似度计算方法精度还有待提高，其主要存在以下三点不足：

20 (1) 跨语言段落语义相似度计算需要解决语言之间的障碍。但是，目前的大多数跨语言相似度检测使用翻译的技巧来解决语言之间的障碍，包括：基于字典、基于平行语料库、基于机器翻译的方法等。相当于把跨语言问题转换为单语或者中间语言之后，再来解决相似度度量的问题。然而，在转换的过程不管是字典、平行语料库、机器翻译都存在缺陷，极有
25 可能丢失文本的一部分深层信息。

(2) 中英句子表征对齐里的负样本应该与锚样本语义不相似。中英句子表征对齐使用中英平行句对数据集进行训练，对于中英平行句对数据集，当锚样本是中文句子的时候，那么它的正样本是与中文句子平行的英文句子，负样本则需要自己构造，现有方法把其它平行句子对里的英文句子
30 作为负样本，但是这种方法生成的负样本有可能是语义与锚样本相似的

假负样本。如果拉开锚样本与假负样本的表征距离,会影响句子表征的对齐,损害中英句子表征模型的性能。

(3)段落文本的表征应该同时考虑词、句子和段落三个层次的信息。但是,现在对段落进行表征的方法或强调通过句子层面的信息对段落进行表征,或强调通过词层面的信息对段落进行表征,缺少对段落不同层次信息的信息交互和信息融合,不能有效地从词、句子和段落三个层次同时对段落文本进行完备的表征。

发明内容

基于此,有必要提供一种面向段落级文本的中英文语义相似度计算方法,以解决现有的中英跨语言段落语义相似度计算方法精度不高的问题。

为实现上述目的,本发明提供了如下方案:

一种面向段落级文本的中英文语义相似度计算方法,包括:

分别对中文段落和英文段落进行段落表征向量提取,包括如下过程:

为中文段落/英文段落中的每个主题词构建主题词节点,为中文段落/英文段落的每个句子构建一个句子节点,为中文段落/英文段落构建一个全局段落节点;

提取主题词节点、句子节点及全局段落节点的初始化特征向量;

根据句子是否含有主题词建立句子节点与主题词节点间的边关系,根据句子在对应段落里的上下文关系建立句子节点之间的边关系,根据主题词在句子中的共现关系建立主题词节点之间的边关系,全局段落节点与每一个主题词节点和句子节点都用一条边连接;每个节点建立一条与自身连接的边;对节点间的边关系进行建模;

将所有节点的初始化特征向量及节点间的边关系输入图注意力网络,输出信息交互后各节点的特征向量;

分别将主题词节点、句子节点的特征向量作平均池化,然后对平均池化后的主题词节点表征向量、句子节点表征向量与全局段落节点的特征向量进行拼接及降维,得到中文段落/英文段落的段落表征向量;

计算中文段落的段落表征向量与英文段落的段落表征向量之间的距离得到中文段落与英文段落的语义相似度。

进一步地,在分别对中文段落和英文段落进行段落表征向量提取之前

还包括:

基于中英平行句对数据集训练得到中英文句子表征对齐模型, 该中英文句子表征对齐模型用于提取主题词节点、句子节点及全局段落节点的初始化特征向量。

5 进一步地, 所述基于中英平行句对数据集训练得到中英文句子表征对齐模型, 包括:

从中英文平行句数据集中选定锚样本, 与锚样本平行的句子为正样本, 其他中英文平行句对中与锚样本语言不同的句子为负样本;

10 对于每个负样本, 若其与锚样本的语义相似度大于设定阈值, 则该负样本为假负样本, 将假负样本的权重赋值为 0; 否则, 将负样本的权重赋值为 1;

15 利用去假负样本后得到的训练集对中英文句子表征对齐模型进行训练, 得到训练完成后的中英文句子表征对齐模型; 其中, 中英文句子表征对齐模型包括两个特征提取分支, 其中一个分支包括中文句子编码器和平均池化层, 另一个分支包括英文句子编码器和平均池化层。

进一步地, 对于每个负样本, 其与锚样本的语义相似度通过如下方法计算:

利用单语言句子语义相似度计算模型计算负样本平行的句子与锚样本之间的第一语义相似度作为负样本与锚样本的语义相似度;

20 或, 利用单语言句子语义相似度计算模型计算负样本与锚样本平行的句子之间的第二语义相似度作为负样本与锚样本的语义相似度;

或, 取第一语义相似度和第二语义相似度中的最大值作为负样本与锚样本的语义相似度。

进一步地, 训练中英文句子表征对齐模型时, 目标函数表示如下:

$$25 \quad \min L(x_i, y_i) + L(y_i, x_i)$$

其中, $L(x_i, y_i)$ 和 $L(y_i, x_i)$ 分别表示锚样本为中文句子和英文句子时的对齐损失函数, 表示如下:

$$L(x_i, y_i) = -\log \frac{\exp(d(x_i, y_i)/\tau)}{\exp(d(x_i, y_i)/\tau) + \sum_{j=1, j \neq i}^N \alpha_{i,j} \times \exp(d(x_i, y_j)/\tau)}$$

$$L(y_i, x_i) = -\log \frac{\exp(d(x_i, y_i)/\tau)}{\exp(d(x_i, y_i)/\tau) + \sum_{j=1, j \neq i}^N \alpha_{j,i} \times \exp(d(x_j, y_i)/\tau)}$$

其中, $d(x, y)$ 表示中文句子 x 与英文句子 y 的向量相似度; τ 表示温度超参数; $\alpha_{i,j}$ 表示锚样本 x_i 的负样本 y_j 的权重; 下标 i, j 表示中英平行句对的编号, N 为该训练批次中中英平行句对的总数。

5 进一步地, 所述为中文段落/英文段落中的每个主题词构建主题词节点, 为中文段落/英文段落的每个句子构建一个句子节点, 包括:

利用 TF-IDF 算法提取中文段落/英文段落中的关键词, 提取重要性最高的前 k 个关键词作为中文段落/英文段落的主题词并对应构建主题词节点 $node_{w_1}, node_{w_2} \cdots node_{w_k}$, 分别记录这 k 个主题词出现的句子序

10 号和句子中的词序;

对中文段落/英文段落进行分句, 为中文段落/英文段落的每个句子构建一个句子节点, 表示为 $node_{s_1}, node_{s_2}, \cdots node_{s_c}$, c 为中文段落/英文段落的句子数。

15 进一步地, 所述提取主题词节点、句子节点及全局段落节点的初始化特征向量, 包括:

对于每个主题词节点, 统计其出现的句子序号和在句子中的词序, 利用中英文句子表征对齐模型中的中文句子编码器/英文句子编码器对出现对应主题词的句子进行编码, 得到句子中每个词的特征向量, 根据主题词在句子中的词序提取该主题词的特征向量作为该主题词节点的初始化特征向量; 如果该主题词在中文段落/英文段落中出现多次, 则将每个位置

20 的主题词的特征向量进行平均池化得到该主题词节点的初始化特征向量;

对于每个句子节点, 利用中英文句子表征对齐模型中的中文句子编码器/英文句子编码器对句子进行编码得到每个词的特征向量, 然后进行平均池化得到特征向量作为该句子节点的初始化特征向量;

25 对所有主题词节点和句子节点的初始化特征向量进行平均池化, 得到全局段落节点的初始化特征向量。

进一步地, 各节点间的边关系通过如下方法确定:

如果句子 A 中含有主题词 a , 则句子 A 对应的句子节点与主题词 a 对应的主题词节点之间建立一条边进行连接;

如果句子 A 和句子 B 在中文段落/英文段落里有上下文关系, 则句子 A 对应的句子节点与句子 B 对应的句子节点间建立一条边进行连接;

如果主题词 a 和主题词 b 在中文段落/英文段落的句子里存在共现关系, 则在它们对应的主题词节点之间建立一条边进行连接;

5 全局段落节点与所有的句子节点和主题词节点都建立一条边进行连接;

每个节点建立一条与自身连接的边;

节点间的边关系建模表示如下:

$$e_{pq} = \begin{cases} 1, \text{节点} p \text{和节点} q \text{之间有边连接} \cup p = q \\ 0, \text{节点} p \text{和节点} q \text{之间无边连接} \cap p \neq q \end{cases}$$

10 其中, e_{pq} 表示节点 p 与节点 q 之间的边关系。

进一步地, 所述将所有节点的初始化特征向量及节点间的边关系输入图注意力网络, 输出信息交互后各节点的特征向量, 包括:

对于每一层图注意力网络, 将上一层图注意力网络输出的节点特征向量 $\mathbf{h} = \{\vec{h}_1, \vec{h}_2, \dots, \vec{h}_n\}$ 作为该层图注意力网络的输入, 输出信息交互后各节点的特征向量 $\mathbf{h}' = \{\vec{h}'_1, \vec{h}'_2, \dots, \vec{h}'_n\}$; 其中 n 为节点总数; 第一层图注意力网络的输入为所有节点的初始化特征向量和节点间的边关系; 其中, 每层图注意力网络构建过程如下:

学习一个共享的线性转换 $\mathbf{W}$, 在节点之间进行自注意力以计算节点之间边的权重, 如果节点 p 和节点 q 之间有一条边连接的话, 则边关系 e_{pq} 的权重计算公式如下:

$$A_{pq} = \mathbf{d}(\mathbf{W}\vec{h}_p, \mathbf{W}\vec{h}_q)$$

其中, A_{pq} 表示节点 q 对节点 p 的重要性, $\mathbf{d}(\cdot)$ 表示两个向量之间的相似度函数;

每个节点的邻居节点信息都会传递到该节点身上, 其中邻居节点包括该节点自身, 对该节点的邻居节点的边关系的权重用 softmax 进行归一化:

$$\beta_{pq} = \text{softmax}_q(A_{pq}) = \frac{\exp(A_{pq})}{\sum_{m \in \mathcal{N}_p} \exp(A_{pm})}$$

其中, $\mathcal{N}_p$ 是节点 p 的邻居节点集, m 表示节点 m ;

节点 p 更新后的特征向量 $\vec{h}'_p$ 表示如下:

$$\vec{h}'_p = \sigma \left(\sum_{q \in N_p} \beta_{pq} \mathbf{W} \vec{h}_q \right)$$

其中, σ 表示非线性变换函数;

将每层图注意力网络最终输出的各节点的特征向量输入到下一层图注意力网络, 经过最后一层图注意力网络输出后, 得到各节点最终的特征向量 $\vec{h}'_p, p = 1, 2 \cdots n$ 。

进一步地, 中文段落/英文段落的段落表征向量通过如下方法得到:

用 $\vec{h}'_p, p = 1, 2 \cdots k$ 表示主题词节点的特征向量, 用 $\vec{h}'_p, p = k + 1, k + 2 \cdots k + c$ 表示句子节点的特征向量, 用 $\vec{h}'_p, p = k + c + 1$ 表示全局段落节点的特征向量; 其中, k 为主题词节点总数, c 为句子节点总数;

分别将主题词节点、句子节点的特征向量作平均池化, 表示如下:

$$h_{word} = \frac{1}{k} \sum_{i=1}^k h'_i$$

$$h_{sentence} = \frac{1}{c} \sum_{i=k+1}^{k+c} h'_i$$

其中, h_{word} 、 $h_{sentence}$ 分别表示平均池化后的主题词节点表征向量、句子节点表征向量;

对平均池化后的主题词节点表征向量、句子节点表征向量与全局段落节点的特征向量进行拼接后, 再经过两层全连接层进行维度压缩, 得到中文段落/英文段落的段落表征向量 $h_{paragraph}$, 表示如下:

$$h_{paragraph} = W_2^T (W_1^T [h_{word} || h_{sentence} || h_{global}] + b_1) + b_2$$

其中, $h_{global} = h'_{k+c+1}$ 表示全局段落节点的特征向量, W_1^T 、 W_2^T 分别表示两层全连接层的权重矩阵, b_1 、 b_2 分别表示两层全连接层的偏置系数。

本发明提出了一种面向段落级文本的中英文语义相似度计算方法,

提出了一种多层次的段落表征方法,能够从主题词、句子和段落三个层次对段落文本进行建模,并基于图注意力网络在各个层次内部和各个层次信息之间进行信息交互,然后融合主题词、句子和段落三个层次的信息得到高质量的段落表征向量,通过计算段落表征向量的距离得到中文段落和英文段落的语义相似度,该方法能实现中英跨语言段落语义相似度的高精度计算,可以应用于对不同语言申请书的摘要进行相似度计算,完成相似申请书的检测和遴选;还可以应用于对不同语言论文的摘要进行相似度计算,从而实现相似论文检测,完成跨语言的论文推荐。

说明书附图

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性的前提下,还可以根据这些附图获得其他的附图。

图1是本发明实施例提供的面向段落级文本的中英文语义相似度计算方法整体框架图;

图2是本发明实施例提供的去假负样本的中英文句子表征对齐流程图;

图3是本发明实施例提供的正负样本的构建过程示意图;

图4是本发明实施例提供的中英文句子表征对齐模型框架示意图,其中,(a)和(b)分别为锚样本为中文句子和英文句子的中英文句子表征对齐模型框架示意图;

图5是本发明实施例提供的中英文句子表征对齐后的结果展示图。

具体实施方式

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。

本发明实施例提供了一种面向段落级文本的中英文语义相似度计算方法,首先通过去假负样本的中英文句子表征对齐方法,将不同语言的句

子表征对齐到公共的表征空间下,在公共表征空间下句子的相似度可以通过计算句子向量的距离得到,为实现中英文段落表征提供词和句子层面的表征基础。其次提出一种多层次的段落表征方法,能够从主题词、句子和段落三个层次对段落文本进行建模,并基于图注意力网络在各个层次内部和各个层次信息之间进行信息交互,然后融合主题词、句子和段落三个层次的信息得到高质量的段落表征向量,通过计算段落表征向量的距离得到中文段落和英文段落的语义相似度。

如图 1 所示,本实施例提供的面向段落级文本的中英文语义相似度计算方法包括三个阶段:去假负样本的中英文句子表征对齐,基于图注意力网络的多层次中英文段落表征,中英跨语言段落语义相似度计算。

S1: 第一阶段,去假负样本的中英文句子表征对齐。

去假负样本的中英文句子表征对齐方法选取中英平行句对数据集作为训练数据,首先以中英平行句对数据集为基础,选取锚样本并生成对应的负样本与正样本;然后对负样本进行筛选,筛掉假负样本,保留真负样本;在构建中英文句子表征对齐模型后,就可以用锚样本、正样本和真负样本来训练中英文句子表征对齐模型;训练完成后就得了中英文句子表征模型,它能够对中文句子和英文句子进行编码,将它们表示在一个公共的语义表征空间下,通过计算中文句子和英文句子的表征向量间的距离就可以得到句子间的语义相似度。

本实施例中基于深度语义的特征提取方法,因为中文和英文之间大量的中英平行句对语料作为监督信息,能够通过有监督的句子表征对齐方法将不同语言的句子语义表征对齐到一个公共的表征空间下,在公共表征空间下句子的相似度可以通过计算句子向量的距离得到,而不需要翻译成单语言再进行计算,避免了机器翻译过程的翻译偏差可能带来的语义丢失。

如图 2 所示,去假负样本的中英文句子表征对齐包括如下步骤。

S11: 以中英平行句对数据集为基础,选取锚样本并生成对应的负样本与正样本。

对于一个中文句子的锚样本,它的正样本是与它语义相似的英文句子,而负样本是与它语义不相似的英文句子,正样本和负样本的质量会对中英文句子表征对齐模型训练有很大的影响,会使得模型的表征能力下降。目

前可用的训练数据集只有中英平行句对数据集,每个中英平行句对由一对平行的中文句子和英文句子组成,假设一个训练批次里有 N 对中英平行句对 $(x_i, y_i), i = 1, 2 \dots N$, 则对于中文句子 x_i 来说, 它的一个正样本就是与之平行的英文句子 y_i , 负样本是同一训练批次中其它平行句对里的英文句子 $y_j, j = 1, 2 \dots N \cup j \neq i$ 。

同理, 当英文句子 y_i 作为锚样本的时候, 正样本是与之平行的中文句子 x_i , 负样本是同一训练批次中其它平行句对里的中文句子 $x_j, j = 1, 2 \dots N \cup j \neq i$ 。

S12: 对负样本进行筛选, 筛掉假负样本, 保留真负样本。

对于负样本 y_j , 需要满足与锚样本 x_i 语义不相似这一条件, 但是通过同一训练批次下其它中英平行句对构造的负样本并不一定满足这个条件, 需要对负样本进行筛选。对于负样本 y_j , 如果它与锚样本 x_i 的语义相似度 $\text{sim}(x_i, y_j)$ 大于设定阈值 $\emptyset$, 则说明负样本与锚样本是语义相似的, 不符合负样本的要求, 即这个负样本是假负样本。如果拉开锚样本与假负样本的表征距离, 会影响句子表征的对齐, 损害中英文句子表征对齐模型的性能。

但是因为锚样本 x_i 与负样本 y_j 是不同语言的句子, 而现有的中英文句子语义相似度计算技术并不成熟(精度不高), 不能直接计算它们的相似度, 这里提出一种间接计算相似度的方法。当锚样本是中文句子 x_i 时, 如果与英文句子 y_j 计算相似度, 可以用中文句子 x_i 的平行英文句子 y_i 来与英文句子 y_j 计算相似度得到 $\text{sim}_{en}(y_i, y_j)$, 同理也可以利用英文句子 y_j 的平行中文句子 x_j 来和中文句子 x_i 计算相似度得到 $\text{sim}_{zh}(x_i, x_j)$, 这里 $\text{sim}_{zh}(\cdot)$ 和 $\text{sim}_{en}(\cdot)$ 分别表示中文句子相似度计算和英文句子相似度计算。具体实施时, 利用成熟的单语言句子语义相似度计算模型来计算锚样本与负样本的语义相似度, 在计算中文句子相似度时, 可以利用中文句子编码器 RoBERTa-wwm-ext(Chinese version)来对中文句子进行编码, 然后将编码后得到的词的表征向量进行平均池化得到中文句子的特征向量, 然后通过计算两个中文句子的特征向量的余弦相似度得到中文句子相似度 $\text{sim}_{zh}(\cdot)$; 在计算英文句子相似度时, 可以利用英文句子编码器 bert-base-uncased (English version)来对英文句子进行编码, 然后将编码后

得到的词的表征向量进行平均池化得到英文句子的特征向量,然后通过计算两个英文句子的特征向量的余弦相似度得到英文句子相似度 $sim_{en}(\cdot)$ 。

因为负样本与锚样本的相似度越大则说明这个负样本是假负样本的可能性越大,为了提高判断假负样本的鲁棒性,在本发明一优选实施例中,
5 中文句子 x_i 和英文句子 y_j 的语义相似度 $sim(x_i, y_j)$ 取两者的最大值,其表示如下:

$$sim(x_i, y_j) = \max(sim_{zh}(x_i, x_j), sim_{en}(y_i, y_j))。$$

得到相似度后,需要设置一个阈值 $\emptyset$,如果负样本与锚样本的相似度 $sim(x_i, y_j) > \emptyset$,则说明这个负样本是假负样本,不能用来训练模型,
10 需要通过给假负样本赋权重为0来去除它对模型训练的影响。负样本的权重赋值可表示如下:

$$\alpha_{i,j} = \begin{cases} 0, & sim(x_i, y_j) \geq \emptyset \\ 1, & sim(x_i, y_j) < \emptyset \end{cases}$$

其中, $\alpha_{i,j}$ 表示锚样本 x_i 的负样本 y_j 的权重。当然,需要说明的是,上述是以锚样本为中文句子 x_i 为例的说明,图3所示为一个正负样本的
15 构建过程示意图。同理,当以英文句子 y_j 为锚样本时,上述筛选掉假负样本的过程原理相同,只需对处理对象作对应调整即可。

S13: 构建中英文句子表征对齐模型。

中英文句子表征对齐模型如图4所示,其中图4(a)为锚样本为中文句子的示意图,图4(b)为锚样本为英文句子的示意图。中文句子进入中文句子编码器得到句子中词的表征向量,经过平均池化层得到中文句子的表征向量,英文句子进入英文句子编码器得到句子中词的表征向量,
20 经过平均池化层得到英文句子的表征向量。这里的英文句子编码器采用bert-base-uncased (English version),它是已经训练好的单语言英语句子编码器,可以对英文句子进行语义表征;中文句子的编码器采用
25 RoBERTa-wwm-ext(Chinese version),它是已经训练好的单语言中文句子编码器,可以对中文句子进行语义表征。

S14: 训练中英文句子表征对齐模型。

用锚样本、正样本、负样本来训练中英文句子表征对齐模型,训练过程中,损失函数采用 InfoNCE loss 的变体,当锚样本是中文句子 时,正

样本是英文句子，负样本是 $N-1$ 个英文句子 $y_j, j = 1, 2 \dots N \cup j \neq i$ 。
对齐损失函数 $L(x_i, y_i)$ 为：

$$L(x_i, y_i) = -\log \frac{\exp(d(x_i, y_i)/\tau)}{\exp(d(x_i, y_i)/\tau) + \sum_{j=1, j \neq i}^N \alpha_{i,j} \times \exp(d(x_i, y_j)/\tau)}$$

其中， $d(x, y)$ 表示中文句子 x 与英文句子 y 的向量相似度； τ 表示温度超参数； $\alpha_{i,j}$ 表示锚样本 x_i 的负样本 y_j 的权重；下标 i, j 表示中英平行句对的编号， N 为该训练批次中中英平行句对的总数。对齐损失函数相当于一个 N 路的 softmax 分类问题，其中正样本是正确的，而 $N-1$ 个负样本是错误的。模型的训练目标是拉近锚样本与正样本的向量距离，拉开与负样本的向量距离，从而能够在拉近语义相似的句子在公共语义空间中的表示，同时拉开语义不相似的句子在公共语义空间中的表示，实现对齐到同一表征空间的目的。

同理，当锚样本是英文句子 y_i 时，对齐损失函数 $L(y_i, x_i)$ 为：

$$L(y_i, x_i) = -\log \frac{\exp(d(x_i, y_i)/\tau)}{\exp(d(x_i, y_i)/\tau) + \sum_{j=1, j \neq i}^N \alpha_{j,i} \times \exp(d(x_j, y_i)/\tau)}$$

其中， $\alpha_{j,i}$ 表示锚样本 y_i 的负样本 x_j 的权重。

则中英文句子表征对齐模型的目标函数表示如下：

$$\min L(x_i, y_i) + L(y_i, x_i)。$$

总的来说训练的目的在于使得语义相似的句子它们向量间的距离最小化，语义不相似的句子它们向量间的距离最大化，以此来实现不同语言句子表征空间的对齐。

目标函数包括两个对齐损失，相当于从两个方向进行对齐（中->英，英->中）。举个简单的例子，有下面四对中英平行句子对来训练模型：

① 英语是使用最广泛的语言。 English is the most widely spoken language.

② 汉语是使用人数最多的语言。 Chinese is the most spoken language.

③ 世界上现存的语言大约有 6909 种。 There are about 6909 languages in existence in the world.

④ 英语的使用最广泛。 English is the most widely used.

当以平行句[①英语是使用最广泛的语言。 English is the most widely

spoken language.]来计算 loss 训练模型时有:

1、当把“英语是使用最广泛的语言。”作为锚样本时，正样本是“English is the most widely spoken language.”，负样本是“Chinese is the most spoken language.”、“There are about 6909 languages in existence in the world.”和“English is the most widely used.”。对负样本进行筛选可以判断出{“Chinese is the most spoken language.”，“There are about 6909 languages in existence in the world.”}是真负样本，而“English is the most widely used.”语义与“英语是使用最广泛的语言。”语义相似，所以是假负样本。要在公共表征空间把“英语是使用最广泛的语言。”与“English is the most widely spoken language.”拉近，同时把“英语是使用最广泛的语言。”和真负样本{“Chinese is the most spoken language.”，“There are about 6909 languages in existence in the world.”}这两个句子拉远，假负样本不进行拉远。

2、当把“English is the most widely spoken language.”作为锚样本时，正样本是“英语是使用最广泛的语言。”，负样本是“汉语是使用人数最多的语言。”、“世界上现存的语言大约有 6909 种。”和“英语的使用最广泛。”。对负样本进行筛选可以判断出{“汉语是使用人数最多的语言。”，“世界上现存的语言大约有 6909 种。”}是真负样本，而“英语的使用最广泛。”语义与“English is the most widely spoken language.”语义相似，所以是假负样本。要在公共表征空间把“English is the most widely spoken language.”与“英语是使用最广泛的语言。”拉近，同时把“English is the most widely spoken language.”和{“汉语是使用人数最多的语言。”，“世界上现存的语言大约有 6909 种。”}这两个句子拉远，假负样本不进行拉远。

这样双向训练后，对于句子“英语是使用最广泛的语言。”，句子“English is the most widely spoken language.”在表征空间中离它最近，对于句子“English is the most widely spoken language.”，句子“英语是使用最广泛的语言。”在表征空间中也离它最近。如果只训练一个方向，就可能导致这样一个情况：对于句子“英语是使用最广泛的语言。”，句子“English is the most widely spoken language.”在表征空间中离它最近，而

对于句子“English is the most widely spoken language.”，句子“英语是使用最广泛的语言。”在表征空间中并不是离它最近的，也就是可能有其它句子离句子“English is the most widely spoken language.”更近，这样达不到句子表征对齐的目的，所以需要双向对齐。

图5所示为中英文句子表征对齐后的结果展示图(中英句子表征对齐以锚样本为“英语是使用最广泛的语言。”作为例子进行展示)，在中英句子表征对齐之前，英文句子语义表征空间和中文句子语义表征空间是两个独立的表征空间，这时中文句子和英文句子它们表征向量的距离是不能反映它们语义相似度的。在通过中英句子表征对齐后，通过拉近锚样本和正样本的表征距离，同时拉开锚样本和真负样本的表征距离，将英文句子语义表征空间和中文句子语义表征空间对齐到一个公共的句子语义表征空间下，这样在公共句子语义表征空间下，中文句子和英文句子它们表征向量的距离就能反映出它们的语义相似度。

S15: 得到中英文句子表征模型。

中英文句子表征对齐模型训练完成后，中英文句子表征对齐模型中的中文句子编码器和英文句子编码器就可以作为中英句子表征模型，分别对中文句子和英文句子进行编码得到对应的向量，得到的中文句子向量和英文句子向量是在一个公共的句子语义表征空间下的，通过计算向量间的距离可以得到句子间的语义相似度。

S2: 第二阶段，基于图注意力网络的多层次中英文段落表征。

基于图注意力网络的多层次中英文段落表征具体包括如下过程：

S21: 从主题词、句子和段落三个层次对中文段落/英文段落进行建模(段落的多层次信息建模)。

首先根据段落的主题词和句子构建主题词节点和句子节点，并添加一个全局段落节点，其次根据句子是否含有主题词建立句子节点与主题词节点间的边关系，同时根据句子在段落里的上下文关系和主题词在句子里的共现关系分别建立句子节点之间的边关系和主题词节点之间的边关系，全局段落节点与每一个主题词节点和句子节点都用一条边连接。通过这样的方式从主题词、句子和段落三个层面把段落建模成一个段落图，以作为下一步中英段落表征的输入。具体步骤如下：

S211: 利用 TF-IDF 算法提取中文段落/英文段落中的关键词, 关键词是能够表达段落中心内容的词语, 提取重要性最高的前 k 个关键词作为段落的主题词并对应构建主题词节点 $node_{w_1}, node_{w_2} \cdots node_{w_k}$, 分别记录这 k 个主题词出现的句子序号和在句子中的词序。

5 S212: 对中文段落/英文段落的进行分句, 为段落里的每个句子构建一个句子节点, 表示为 $node_{s_1}, node_{s_2}, \cdots node_{s_c}$, c 为中文段落/英文段落里的句子数。

S213: 为中文段落/英文段落构建一个全局段落节点 $node_{global}$ 。

S214: 利用中英文句子表征模型对节点的特征进行初始化。

10 假设所构建的段落图中有 k 个主题词节点, c 个句子节点, 1 个全局节点, 一共有 $n=k+c+1$ 个节点, 按照主题词节点、句子节点和全局节点的顺序对节点进行排序得到 $node_i, i = 1, 2 \cdots n$, 对这些节点进行特征初始化, 方法如下:

15 对于每个主题词节点 h_i , 需要统计主题词在段落里出现的位置 (在哪些句子中出现过以及在句子里的位置, 即句子序号和在句子中的词序), 在对主题词进行编码时, 首先将主题词出现的句子输入进中英文句子表征模型中, 在经过中文/英文句子编码器对句子编码后, 得到每个词的特征向量, 根据记录的词序提取出该主题词的特征向量作为该主题词节点的初始化向量; 如果主题词在段落里出现了多次, 则先分别通过上面方法得到
20 每个位置主题词的特征向量, 然后对这些向量进行平均池化得到主题词节点的初始化特征。

对于 k 个主题词节点就可以得到特征向量 $h_1, h_2 \cdots h_k$ 。这样得到的主题词特征包含了句子的上下文信息, 在不同句子里的词根据上下文信息的不同可以有着不同的语义, 能够有效解决一词多义问题。

25 句子节点的初始化特征向量就是句子通过中英文句子表征模型进行编码和平均池化后得到的特征向量, 对于 c 个句子节点得到 $h_{k+1}, h_{k+2} \cdots h_{k+c}$ 。

全局节点的初始化特征向量是其它所有主题词节点和句子节点初始化特征向量进行平均池化后得到的 h_{k+c+1} 。

30 S215: 通过如下方式建立节点间的边关系:

(1) 如果句子 A 中含有主题词 a, 则句子 A 对应的句子节点与主题词 a 对应的主题词节点之间建立一条边进行连接;

(2) 如果句子 A 和句子 B 在中文段落/英文段落里有上下文关系, 则句子 A 对应的句子节点与句子 B 对应的句子节点间建立一条边进行连接;

(3) 如果主题词 a 和主题词 b 在中文段落/英文段落的句子里存在共现关系, 则在它们对应的主题词节点之间建立一条边进行连接;

(4) 全局段落节点与所有的句子节点和主题词节点都建立一条边进行连接;

(5) 每个节点建立一条与自身连接的边。

S216: 对节点间的边关系进行建模表示。

对于节点 p 和节点 q , 如果它们之间有一条边连接, 则它们互为邻居节点 (节点自己也算作自己的邻居节点)。如果节点 p 和节点 q 互为邻居节点则 $e_{pq}=1$, 否则 $e_{pq}=0$ 。节点间的边关系建模表示如下:

$$e_{pq} = \begin{cases} 1, & \text{节点 } p \text{ 和节点 } q \text{ 之间有边连接 } \cup p = q \\ 0, & \text{节点 } p \text{ 和节点 } q \text{ 之间无边连接 } \cap p \neq q \end{cases}$$

其中, e_{pq} 表示节点 p 与节点 q 之间的边关系。

S22: 基于图注意力网络的信息交互和信息融合。

通过图注意力网络对段落图进行处理, 在各个节点间进行信息传递, 最后通过融合主题词、句子和全局段落信息三个层次的信息生成中文段落/英文段落的段落表征向量。

对于每一层图注意力网络, 将上一层图注意力网络输出的节点特征向量 $\mathbf{h} = \{\vec{h}_1, \vec{h}_2, \dots, \vec{h}_n\}$ 作为该层图注意力网络的输入, 输出信息交互后各节点的特征向量 $\mathbf{h}' = \{\vec{h}'_1, \vec{h}'_2, \dots, \vec{h}'_n\}$; 其中 n 为节点总数; 第一层图注意力网络的输入为所有节点的初始化特征向量和节点间的边关系

$e_{pq}, p, q \in \{1, 2, \dots, n\}$ 。每层图注意力网络构建过程如下:

为了获得足够的表达能力, 学习一个共享的线性转换 $\mathbf{W}$, 将节点的输入特征转换为更深层次的特征。然后在节点之间进行自注意力以计算节点之间边的权重, 如果节点 p 和节点 q 之间有一条边连接的话, 则边关系 e_{pq} 的权重计算公式如下:

$$A_{pq} = d(\vec{W}h_p, \vec{W}h_q)$$

其中, A_{pq} 表示节点 q 对节点 p 的重要性; $d(\cdot)$ 表示两个向量之间的相似度函数, 这里采用余弦相似度; 在信息传递过程中, 每个节点的邻居节点信息都会传递到该节点身上, 其中邻居节点包括该节点自身, 对该节点的邻居节点的边关系的权重用 softmax 进行归一化:

$$\beta_{pq} = \text{softmax}_q(A_{pq}) = \frac{\exp(A_{pq})}{\sum_{m \in \mathcal{N}_p} \exp(A_{pm})}$$

其中, $\mathcal{N}_p$ 是节点 p 的邻居节点集, m 表示节点 m 。得到归一化的注意力权重后就可以用来更新节点 p 的特征, 节点 p 更新后的特征向量 $\vec{h}'_p$ 表示如下:

$$\vec{h}'_p = \sigma \left(\sum_{q \in \mathcal{N}_p} \beta_{pq} \vec{W}h_q \right)$$

其中, σ 表示非线性变换函数。

将每层图注意力网络最终输出的各节点的特征向量输入到下一层图注意力网络, 经过最后一层图注意力网络输出后, 得到各节点最终的特征向量 $h'_p, p = 1, 2 \dots n$ 。

对各节点最终的特征向量进行融合得到中文段落/英文段落的段落表征向量, 首先分别将主题词节点、句子节点的特征向量作平均池化, 表示如下:

$$h_{word} = \frac{1}{k} \sum_{i=1}^k h'_i$$

$$h_{sentence} = \frac{1}{c} \sum_{i=k+1}^{k+c} h'_i$$

其中, h_{word} 、 $h_{sentence}$ 分别表示平均池化后的主题词节点表征向量、句子节点表征向量。然后对平均池化后的主题词节点表征向量、句子节点表征向量与全局段落节点的特征向量进行拼接后, 再经过两层全连接层进行维度压缩, 得到中文段落/英文段落的段落表征向量 $h_{paragraph}$, 表示如下:

$$h_{paragraph} = W_2^T (W_1^T [h_{word} || h_{sentence} || h_{global}] + b_1) + b_2$$

其中, $h_{global} = h'_{k+c+1}$ 表示全局段落节点的特征向量, W_1^T 、 W_2^T 分别表示两层全连接层的权重矩阵, b_1 、 b_2 分别表示两层全连接层的偏置系数。

5 S3: 第三阶段, 中英跨语言段落语义相似度计算。

计算中文段落的段落表征向量 $h_{paragraph_{zh}}$ 与英文段落的段落表征向量 $h_{paragraph_{en}}$ 之间的距离得到中文段落与英文段落的语义相似度 *Similarity*:

$$Similarity = \cos(h_{paragraph_{zh}}, h_{paragraph_{en}})$$

10 其中, $\cos(\cdot)$ 是计算向量间的余弦相似度。

本说明书中各个实施例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部分互相参见即可。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
15 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。

权 利 要 求 书

1.一种面向段落级文本的中英文语义相似度计算方法，其特征在于，包括：

分别对中文段落和英文段落进行段落表征向量提取，包括如下过程：

为中文段落/英文段落中的每个主题词构建主题词节点，为中文段落/英文段落的每个句子构建一个句子节点，为中文段落/英文段落构建一个全局段落节点；

提取主题词节点、句子节点及全局段落节点的初始化特征向量；

根据句子是否含有主题词建立句子节点与主题词节点间的边关系，根据句子在对应段落里的上下文关系建立句子节点之间的边关系，根据主题词在句子里的共现关系建立主题词节点之间的边关系，全局段落节点与每一个主题词节点和句子节点都用一条边连接；每个节点建立一条与自身连接的边；对节点间的边关系进行建模；

将所有节点的初始化特征向量及节点间的边关系输入图注意力网络，输出信息交互后各节点的特征向量；

分别将主题词节点、句子节点的特征向量作平均池化，然后对平均池化后的主题词节点表征向量、句子节点表征向量与全局段落节点的特征向量进行拼接及降维，得到中文段落/英文段落的段落表征向量；

计算中文段落的段落表征向量与英文段落的段落表征向量之间的距离得到中文段落与英文段落的语义相似度。

2.根据权利要求1所述的面向段落级文本的中英文语义相似度计算方法，其特征在于，在分别对中文段落和英文段落进行段落表征向量提取之前还包括：

基于中英平行句对数据集训练得到中英文句子表征对齐模型，该中英文句子表征对齐模型用于提取主题词节点、句子节点及全局段落节点的初始化特征向量。

3.根据权利要求2所述的面向段落级文本的中英文语义相似度计算方法，其特征在于，所述基于中英平行句对数据集训练得到中英文句子表征对齐模型，包括：

从中英文平行句数据集中选定锚样本，与锚样本平行的句子为正样本，其他中英文平行句对中与锚样本语言不同的句子为负样本；

对于每个负样本,若其与锚样本的语义相似度大于设定阈值,则该负样本为假负样本,将假负样本的权重赋值为 0; 否则,将负样本的权重赋值为 1;

利用去假负样本后得到的训练集对中英文句子表征对齐模型进行训练,得到训练完成后的中英文句子表征对齐模型; 其中,中英文句子表征对齐模型包括两个特征提取分支,其中一个分支包括中文句子编码器和平均池化层,另一个分支包括英文句子编码器和平均池化层。

4.根据权利要求3所述的面向段落级文本的中英文语义相似度计算方法,其特征在于,对于每个负样本,其与锚样本的语义相似度通过如下方法计算:

利用单语言句子语义相似度计算模型计算负样本平行的句子与锚样本之间的第一语义相似度作为负样本与锚样本的语义相似度;

或,利用单语言句子语义相似度计算模型计算负样本与锚样本平行的句子之间的第二语义相似度作为负样本与锚样本的语义相似度;

或,取第一语义相似度和第二语义相似度中的最大值作为负样本与锚样本的语义相似度。

5.根据权利要求3所述的面向段落级文本的中英文语义相似度计算方法,其特征在于,训练中英文句子表征对齐模型时,目标函数表示如下:

$$\min L(x_i, y_i) + L(y_i, x_i)$$

其中, $L(x_i, y_i)$ 和 $L(y_i, x_i)$ 分别表示锚样本为中文句子和英文句子时的对齐损失函数,表示如下:

$$L(x_i, y_i) = -\log \frac{\exp(d(x_i, y_i)/\tau)}{\exp(d(x_i, y_i)/\tau) + \sum_{j=1, j \neq i}^N \alpha_{ij} \times \exp(d(x_i, y_j)/\tau)}$$

$$L(y_i, x_i) = -\log \frac{\exp(d(x_i, y_i)/\tau)}{\exp(d(x_i, y_i)/\tau) + \sum_{j=1, j \neq i}^N \alpha_{ji} \times \exp(d(x_j, y_i)/\tau)}$$

其中, $d(x, y)$ 表示中文句子 x 与英文句子 y 的向量相似度; τ 表示温度超参数; α_{ij} 表示锚样本 x_i 的负样本 y_j 的权重; 下标 i, j 表示中英平行句对的编号, N 为该训练批次中中英平行句对的总数。

6.根据权利要求1至5任一项所述的面向段落级文本的中英文语义相似度计算方法,其特征在于,所述为中文段落/英文段落中的每个主题词

构建主题词节点，为中文段落/英文段落的每个句子构建一个句子节点，包括：

利用 TF-IDF 算法提取中文段落/英文段落中的关键词，提取重要性最高的前 k 个关键词作为中文段落/英文段落的主题词并对应构建主题词节点 $node_{w_1}, node_{w_2} \cdots node_{w_k}$ ，分别记录这 k 个主题词出现的句子序号和在句子中的词序；

对中文段落/英文段落进行分句，为中文段落/英文段落的每个句子构建一个句子节点，表示为 $node_{s_1}, node_{s_2}, \cdots node_{s_c}$ ， c 为中文段落/英文段落的句子数。

7. 根据权利要求 2 至 5 任一项所述的面向段落级文本的中英文语义相似度计算方法，其特征在于，所述提取主题词节点、句子节点及全局段落节点的初始化特征向量，包括：

对于每个主题词节点，统计其出现的句子序号和在句子中的词序，利用中英文句子表征对齐模型中的中文句子编码器/英文句子编码器对出现对应主题词的句子进行编码，得到句子中每个词的特征向量，根据主题词在句子中的词序提取该主题词的特征向量作为该主题词节点的初始化特征向量；如果该主题词在中文段落/英文段落中出现多次，则将每个位置的主题词的特征向量进行平均池化得到该主题词节点的初始化特征向量；

对于每个句子节点，利用中英文句子表征对齐模型中的中文句子编码器/英文句子编码器对句子进行编码得到每个词的特征向量，然后进行平均池化得到特征向量作为该句子节点的初始化特征向量；

对所有主题词节点和句子节点的初始化特征向量进行平均池化，得到全局段落节点的初始化特征向量。

8. 根据权利要求 1 至 5 任一项所述的面向段落级文本的中英文语义相似度计算方法，其特征在于，各节点间的边关系通过如下方法确定：

如果句子 A 中含有主题词 a，则句子 A 对应的句子节点与主题词 a 对应的主题词节点之间建立一条边进行连接；

如果句子 A 和句子 B 在中文段落/英文段落里有上下文关系，则句子 A 对应的句子节点与句子 B 对应的句子节点间建立一条边进行连接；

如果主题词 a 和主题词 b 在中文段落/英文段落的句子里存在共现关

系，则在它们对应的主题词节点之间建立一条边进行连接；

全局段落节点与所有的句子节点和主题词节点都建立一条边进行连接；

每个节点建立一条与自身连接的边；

5 节点间的边关系建模表示如下：

$$e_{pq} = \begin{cases} 1, \text{节点} p \text{和节点} q \text{之间有边连接} \cup p = q \\ 0, \text{节点} p \text{和节点} q \text{之间无边连接} \cap p \neq q \end{cases}$$

其中， e_{pq} 表示节点 p 与节点 q 之间的边关系。

9.根据权利要求 1 至 5 任一项所述的面向段落级文本的中英文语义相似度计算方法，其特征在于，所述将所有节点的初始化特征向量及节点间的边关系输入图注意力网络，输出信息交互后各节点的特征向量，包括：

10 对于每一层图注意力网络，将上一层图注意力网络输出的节点特征向量 $\mathbf{h} = \{\vec{h}_1, \vec{h}_2, \dots, \vec{h}_n\}$ 作为该层图注意力网络的输入，输出信息交互后各节点的特征向量 $\mathbf{h}' = \{\vec{h}'_1, \vec{h}'_2, \dots, \vec{h}'_n\}$ ；其中 n 为节点总数；第一层图注意力网络的输入为所有节点的初始化特征向量和节点间的边关系；其中，每层图注意力网络构建过程如下：

15 学习一个共享的线性转换 $\mathbf{W}$ ，在节点之间进行自注意力以计算节点之间边的权重，如果节点 p 和节点 q 之间有一条边连接的话，则边关系 e_{pq} 的权重计算公式如下：

$$A_{pq} = d(\mathbf{W}\vec{h}_p, \mathbf{W}\vec{h}_q)$$

其中， A_{pq} 表示节点 q 对节点 p 的重要性， $d(\cdot)$ 表示两个向量之间的相似度函数；

20 每个节点的邻居节点信息都会传递到该节点身上，其中邻居节点包括该节点自身，对该节点的邻居节点的边关系的权重用 softmax 进行归一化：

$$\beta_{pq} = \text{softmax}_q(A_{pq}) = \frac{\exp(A_{pq})}{\sum_{m \in \mathcal{N}_p} \exp(A_{pm})}$$

其中， $\mathcal{N}_p$ 是节点 p 的邻居节点集， m 表示节点 m ；

节点 p 更新后的特征向量 $\vec{h}'_p$ 表示如下：

$$\vec{h}_p' = \sigma \left(\sum_{q \in \mathcal{N}_p} \beta_{pq} \mathbf{W} \vec{h}_q \right)$$

其中， σ 表示非线性变换函数；

将每层图注意力网络最终输出的各节点的特征向量输入到下一层图注意力网络，经过最后一层图注意力网络输出后，得到各节点最终的特征向量 $\vec{h}_p', p = 1, 2 \dots n$ 。

5 10.根据权利要求 1 至 5 任一项所述的面向段落级文本的中英文语义相似度计算方法，其特征在于，中文段落/英文段落的段落表征向量通过如下方法得到：

用 $\vec{h}_p', p = 1, 2 \dots k$ 表示主题词节点的特征向量，用 $\vec{h}_p', p = k + 1, k + 2 \dots k + c$ 表示句子节点的特征向量，用 $\vec{h}_p', p = k + c + 1$ 表示全局段落节点的特征向量；其中， k 为主题词节点总数， c 为句子节点总数；

分别将主题词节点、句子节点的特征向量作平均池化，表示如下：

$$\vec{h}_{word} = \frac{1}{k} \sum_{i=1}^k \vec{h}_i'$$

$$\vec{h}_{sentence} = \frac{1}{c} \sum_{i=k+1}^{k+c} \vec{h}_i'$$

其中， $\vec{h}_{word}$ 、 $\vec{h}_{sentence}$ 分别表示平均池化后的主题词节点表征向量、句子节点表征向量；

15 对平均池化后的主题词节点表征向量、句子节点表征向量与全局段落节点的特征向量进行拼接后，再经过两层全连接层进行维度压缩，得到中文段落/英文段落的段落表征向量 $\vec{h}_{paragraph}$ ，表示如下：

$$\vec{h}_{paragraph} = W_2^T (W_1^T [\vec{h}_{word} || \vec{h}_{sentence} || \vec{h}_{global}] + b_1) + b_2$$

其中， $\vec{h}_{global} = \vec{h}_{k+c+1}'$ 表示全局段落节点的特征向量， W_1^T 、 W_2^T 分别表示两层全连接层的权重矩阵， b_1 、 b_2 分别表示两层全连接层的偏置系数。

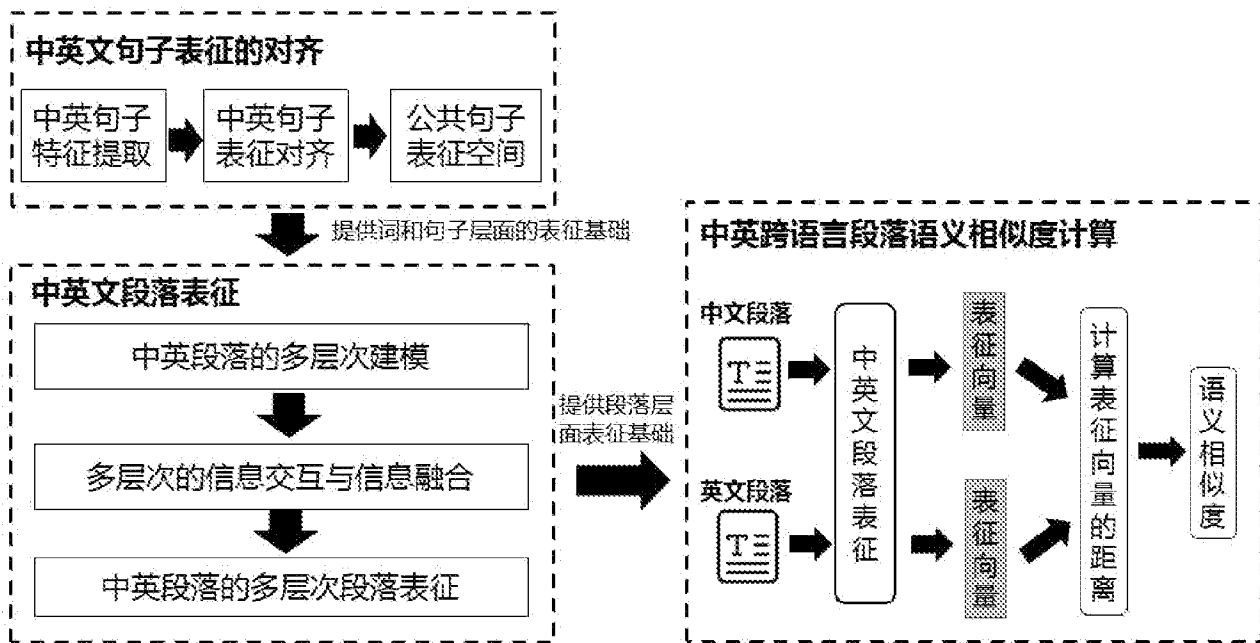


图 1

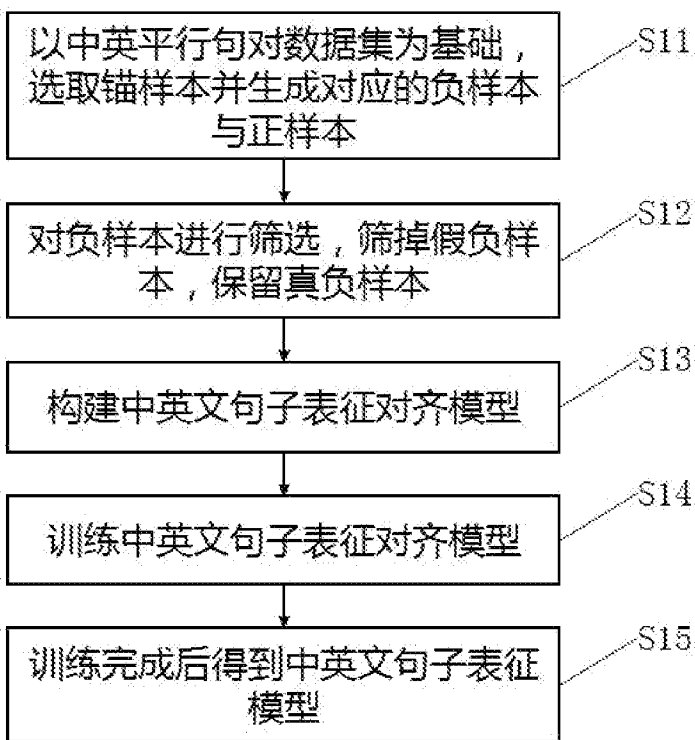


图 2

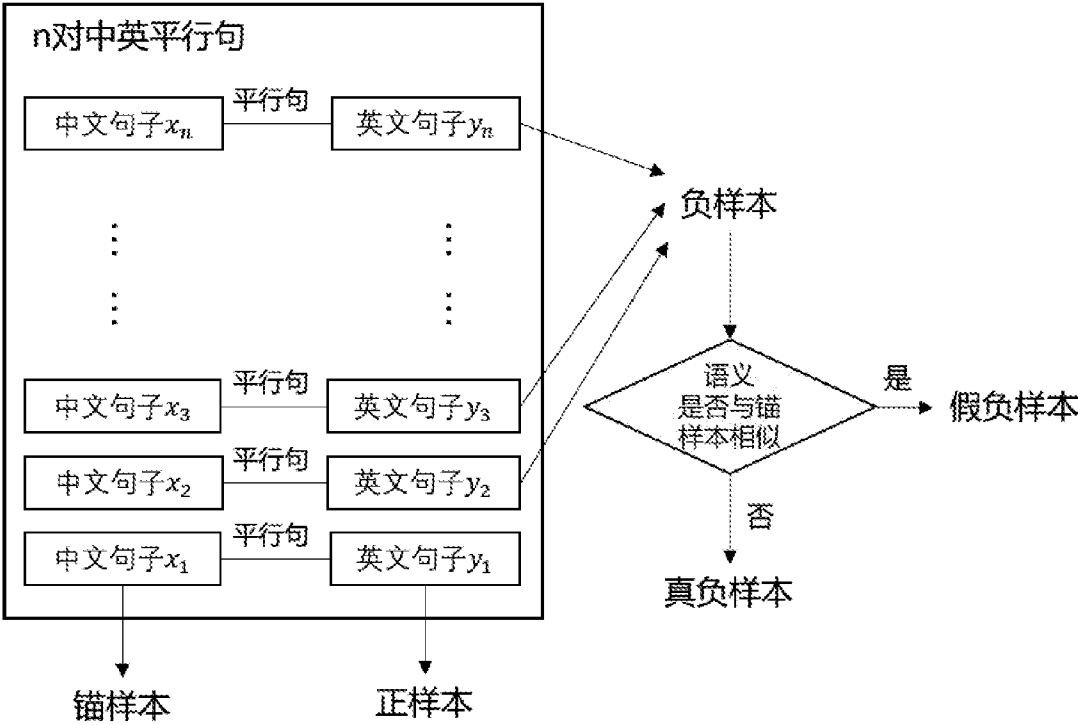


图 3

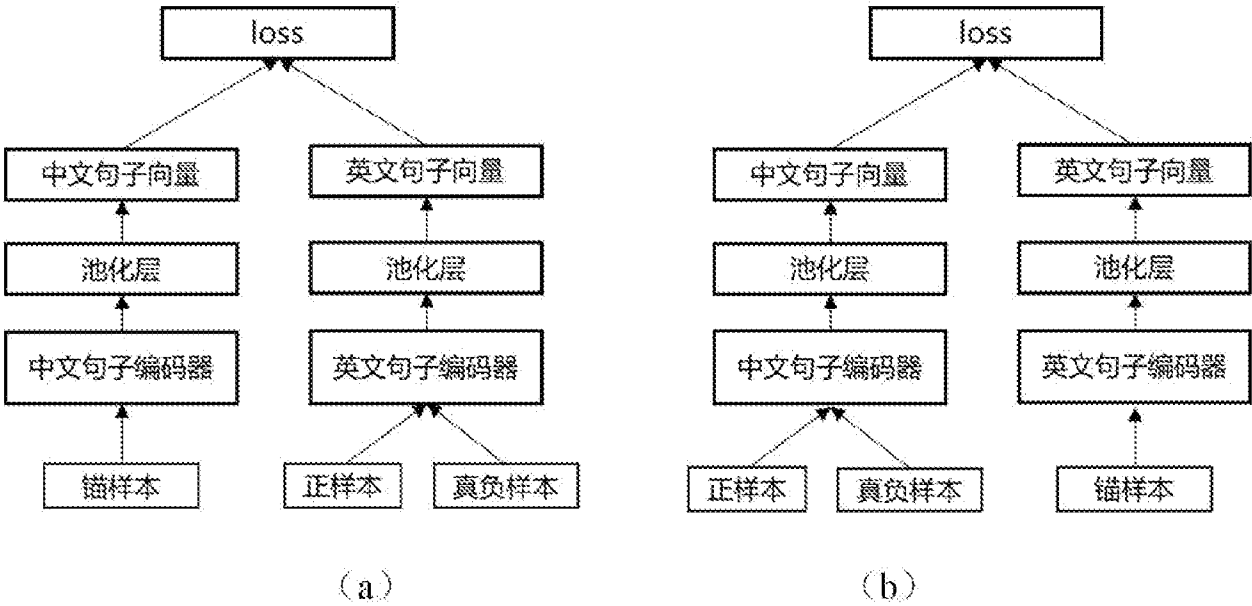


图 4

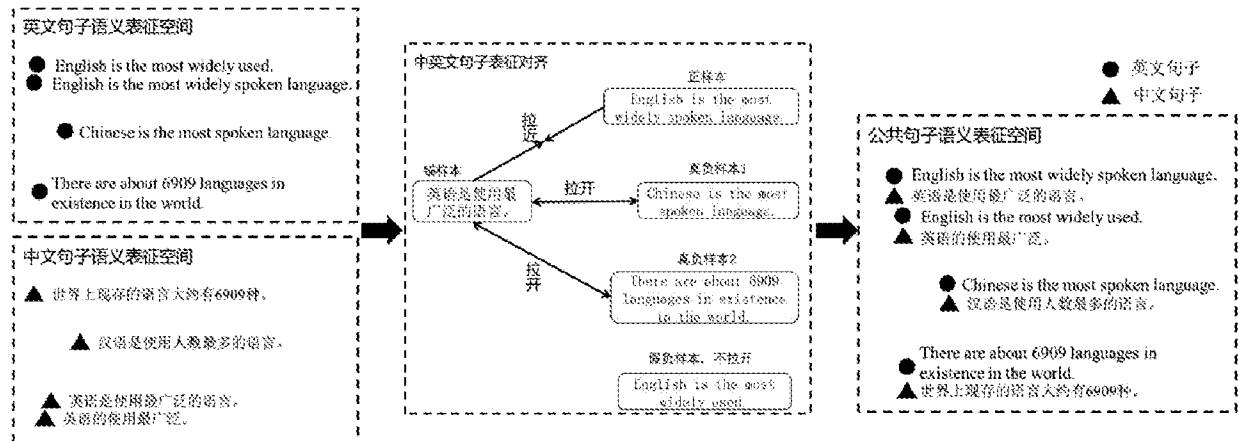


图 5

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2023/128479

A. CLASSIFICATION OF SUBJECT MATTER

G06F 40/30(2020.01)i; G06F40/211(2020.01)i; G06F18/22(2023.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC:G06F40/-; G06F18/-

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNABS, CNTXT, CNKI, BAIDU, VEN, DWPI, USTXT, WOTXT, IEEE: 语义相似度, 段落, 句子, 中英文, 特征向量, 主题, 节点, 边, 池化, 对齐模型, 锚样本, 平行, 假负样本, 编码器, 损失函数, 注意力, semantic similarity, paragraph, sentence, Chinese-English, feature vector, theme, node, edge, pooling, align model, anchor sample, parallel, fake negative sample, encoder, loss function, attention

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 115828931 A (CENTRAL SOUTH UNIVERSITY) 21 March 2023 (2023-03-21) claims 1-10	1-10
Y	CN 112818121 A (RUNLIAN SOFTWARE SYSTEM (SHENZHEN) CO., LTD.) 18 May 2021 (2021-05-18) description, paragraphs 4-132	1-2, 6, 8
Y	CN 109213995 A (HARBIN ENGINEERING UNIVERSITY) 15 January 2019 (2019-01-15) description, paragraphs 7-63	1-2, 6, 8
A	CN 107862045 A (HARBIN ENGINEERING UNIVERSITY) 30 March 2018 (2018-03-30) entire document	1-10
A	CN 113836192 A (SHENYI INFORMATION TECHNOLOGY (HENGQIN) CO., LTD.) 24 December 2021 (2021-12-24) entire document	1-10



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance
 “D” document cited by the applicant in the international application
 “E” earlier application or patent but published on or after the international filing date
 “L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
 “O” document referring to an oral disclosure, use, exhibition or other means
 “P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

21 November 2023

Date of mailing of the international search report

30 November 2023

Name and mailing address of the ISA/CN

China National Intellectual Property Administration (ISA/
CN)
China No. 6, Xitucheng Road, Jimenqiao, Haidian District,
Beijing 100088

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2023/128479**C. DOCUMENTS CONSIDERED TO BE RELEVANT**

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 113901831 A (KUNMING UNIVERSITY OF SCIENCE AND TECHNOLOGY) 07 January 2022 (2022-01-07) entire document	1-10
A	CN 114821051 A (INSTITUTE OF SOIL FERTILIZER & RESOURCE ENVIRONMENT, JIANGXI ACADEMY OF AGRICULTURAL SCIENCES) 29 July 2022 (2022-07-29) entire document	1-10
A	EP 3249547 A1 (ACCENTURE GLOBAL SOLUTIONS LIMITED) 29 November 2017 (2017-11-29) entire document	1-10

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2023/128479

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	115828931	A	21 March 2023	CN	115828931	B	02 May 2023
CN	112818121	A	18 May 2021	None			
CN	109213995	A	15 January 2019	CN	109213995	B	18 November 2022
CN	107862045	A	30 March 2018	CN	107862045	B	14 January 2022
CN	113836192	A	24 December 2021	CN	113836192	B	03 May 2022
CN	113901831	A	07 January 2022	None			
CN	114821051	A	29 July 2022	None			
EP	3249547	A1	29 November 2017	US	2017344464	A1	30 November 2017
				IN	201641018370	A	01 December 2017
				US	10204032	B2	12 February 2019
				IN	425527	B	24 March 2023

A. 主题的分类 G06F 40/30(2020.01)i; G06F40/211(2020.01)i; G06F18/22(2023.01)i 按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类																							
B. 检索领域 检索的最低限度文献(标明分类系统和分类号) IPC:G06F40/-; G06F18/- 包含在检索领域中的除最低限度文献以外的检索文献 在国际检索时查阅的电子数据库(数据库的名称, 和使用的检索词(如使用)) CNABS,CNXTXT,CNKI,BAIDU,VEN,DWPI,USTXT,WOTXT,IEEE:语义相似度, 段落, 句子, 中英文, 特征向量, 主题, 节点, 边, 池化, 对齐模型, 锚样本, 平行, 假负样本, 编码器, 损失函数, 注意力, semantic similarity, paragraph, sentence, Chinese-English, feature vector, theme, node, edge, pooling, align model, anchor sample, parallel, fake negative sample, encoder, loss function, attention																							
C. 相关文件 <table border="1"> <thead> <tr> <th>类型*</th> <th>引用文件, 必要时, 指明相关段落</th> <th>相关的权利要求</th> </tr> </thead> <tbody> <tr> <td>PX</td> <td>CN 115828931 A (中南大学) 2023年3月21日 (2023 - 03 - 21) 权利要求1-10</td> <td>1-10</td> </tr> <tr> <td>Y</td> <td>CN 112818121 A (润联软件系统(深圳)有限公司) 2021年5月18日 (2021 - 05 - 18) 说明书第4-132段</td> <td>1-2、6、8</td> </tr> <tr> <td>Y</td> <td>CN 109213995 A (哈尔滨工程大学) 2019年1月15日 (2019 - 01 - 15) 说明书第7-63段</td> <td>1-2、6、8</td> </tr> <tr> <td>A</td> <td>CN 107862045 A (哈尔滨工程大学) 2018年3月30日 (2018 - 03 - 30) 全文</td> <td>1-10</td> </tr> <tr> <td>A</td> <td>CN 113836192 A (深译信息科技(横琴)有限公司) 2021年12月24日 (2021 - 12 - 24) 全文</td> <td>1-10</td> </tr> <tr> <td>A</td> <td>CN 113901831 A (昆明理工大学) 2022年1月7日 (2022 - 01 - 07) 全文</td> <td>1-10</td> </tr> </tbody> </table>			类型*	引用文件, 必要时, 指明相关段落	相关的权利要求	PX	CN 115828931 A (中南大学) 2023年3月21日 (2023 - 03 - 21) 权利要求1-10	1-10	Y	CN 112818121 A (润联软件系统(深圳)有限公司) 2021年5月18日 (2021 - 05 - 18) 说明书第4-132段	1-2、6、8	Y	CN 109213995 A (哈尔滨工程大学) 2019年1月15日 (2019 - 01 - 15) 说明书第7-63段	1-2、6、8	A	CN 107862045 A (哈尔滨工程大学) 2018年3月30日 (2018 - 03 - 30) 全文	1-10	A	CN 113836192 A (深译信息科技(横琴)有限公司) 2021年12月24日 (2021 - 12 - 24) 全文	1-10	A	CN 113901831 A (昆明理工大学) 2022年1月7日 (2022 - 01 - 07) 全文	1-10
类型*	引用文件, 必要时, 指明相关段落	相关的权利要求																					
PX	CN 115828931 A (中南大学) 2023年3月21日 (2023 - 03 - 21) 权利要求1-10	1-10																					
Y	CN 112818121 A (润联软件系统(深圳)有限公司) 2021年5月18日 (2021 - 05 - 18) 说明书第4-132段	1-2、6、8																					
Y	CN 109213995 A (哈尔滨工程大学) 2019年1月15日 (2019 - 01 - 15) 说明书第7-63段	1-2、6、8																					
A	CN 107862045 A (哈尔滨工程大学) 2018年3月30日 (2018 - 03 - 30) 全文	1-10																					
A	CN 113836192 A (深译信息科技(横琴)有限公司) 2021年12月24日 (2021 - 12 - 24) 全文	1-10																					
A	CN 113901831 A (昆明理工大学) 2022年1月7日 (2022 - 01 - 07) 全文	1-10																					
☒ 其余文件在C栏的续页中列出。 ☒ 见同族专利附件。																							
<table border="0"> <tr> <td> * 引用文件的具体类型: “A” 认为不特别相关的表示了现有技术一般状态的文件 “D” 申请人在国际申请中引证的文件 “E” 在国际申请日的当天或之后公布的在先申请或专利 “L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件(如具体说明的) “O” 涉及口头公开、使用、展览或其他方式公开的文件 “P” 公布日先于国际申请日但迟于所要求的优先权日的文件 </td> <td> “T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件 “X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性 “Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性 “&” 同族专利的文件 </td> </tr> </table>			* 引用文件的具体类型: “A” 认为不特别相关的表示了现有技术一般状态的文件 “D” 申请人在国际申请中引证的文件 “E” 在国际申请日的当天或之后公布的在先申请或专利 “L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件(如具体说明的) “O” 涉及口头公开、使用、展览或其他方式公开的文件 “P” 公布日先于国际申请日但迟于所要求的优先权日的文件	“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件 “X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性 “Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性 “&” 同族专利的文件																			
* 引用文件的具体类型: “A” 认为不特别相关的表示了现有技术一般状态的文件 “D” 申请人在国际申请中引证的文件 “E” 在国际申请日的当天或之后公布的在先申请或专利 “L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件(如具体说明的) “O” 涉及口头公开、使用、展览或其他方式公开的文件 “P” 公布日先于国际申请日但迟于所要求的优先权日的文件	“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件 “X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性 “Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性 “&” 同族专利的文件																						
国际检索实际完成的日期 2023年11月21日		国际检索报告邮寄日期 2023年11月30日																					
ISA/CN的名称和邮寄地址 中国国家知识产权局 中国北京市海淀区蓟门桥西土城路6号 100088		授权官员 沈晓娟 电话号码 (+86) 028-62967683																					

C. 相关文件		
类 型*	引用文件，必要时，指明相关段落	相关的权利要求
A	CN 114821051 A (江西省农业科学院土壤肥料与资源环境研究所) 2022年7月29日 (2022 - 07 - 29) 全文	1-10
A	EP 3249547 A1 (ACCENTURE GLOBAL SOLUTIONS LIMITED) 2017年11月29日 (2017 - 11 - 29) 全文	1-10

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2023/128479

检索报告引用的专利文件			公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN	115828931	A	2023年3月21日	CN 115828931 B	2023年5月2日
CN	112818121	A	2021年5月18日	无	
CN	109213995	A	2019年1月15日	CN 109213995 B	2022年11月18日
CN	107862045	A	2018年3月30日	CN 107862045 B	2022年1月14日
CN	113836192	A	2021年12月24日	CN 113836192 B	2022年5月3日
CN	113901831	A	2022年1月7日	无	
CN	114821051	A	2022年7月29日	无	
EP	3249547	A1	2017年11月29日	US 2017344464 A1	2017年11月30日
				IN 201641018370 A	2017年12月1日
				US 10204032 B2	2019年2月12日
				IN 425527 B	2023年3月24日