(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 117648923 B
(45)授权公告日 2024.05.10

(21)申请号 202410120343.5
(22)申请日 2024.01.29
(65)同一申请的已公布的文献号
申请公布号 CN 117648923 A
(43)申请公布日 2024.03.05
(73)专利权人 安徽省立医院(中国科学技术大学附属第一医院)
地址 230001 安徽省合肥市庐阳区庐江路9号
(72)发明人 高敏 陈恩红 刘昌春 蒋浚哲 张凯 王慕秋 李京秀 宋雪莉 丁蓓蓓 张梦云
(74)专利代理机构 合肥天明专利事务所(普通合伙) 34115
专利代理师 金凯
(51)Int.Cl.
G06F 40/232(2020.01)
G06F 40/166(2020.01)
G06F 40/216(2020.01)
(56)对比文件
CN 111310443 A,2020.06.19
CN 115114919 A,2022.09.27
US 2022121822 A1,2022.04.21
CN 112530597 A,2021.03.19
CN 113657098 A,2021.11.16
CN 113673228 A,2021.11.19
CN 113935317 A,2022.01.14
CN 114881006 A,2022.08.09
CN 115081430 A,2022.09.20
CN 115862040 A,2023.03.28
CN 116522905 A,2023.08.01
审查员 朱江

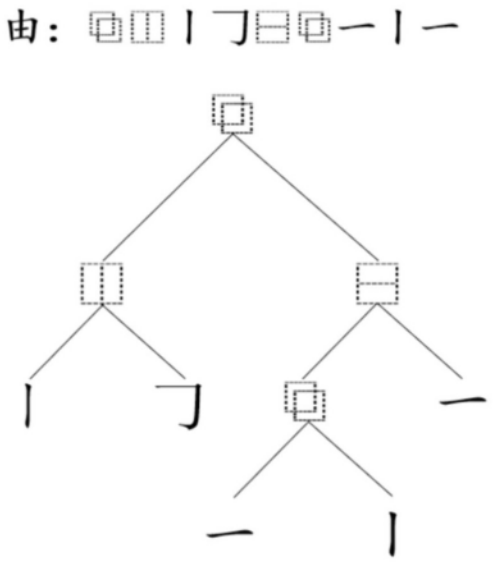
权利要求书3页 说明书8页 附图4页

(54)发明名称

一种适用于医疗语境的中文拼写纠错方法

(57)摘要

本发明涉及人工智能领域,具体涉及一种适用于医疗语境的中文拼写纠错方法,包括将句子转换为汉字标号序列后输入到BERT预训练中文语言模型中,得到语境信息特征,将语境信息特征进行线性变换使其尺寸对齐词表;计算每一个位置的前*J*个候选项的归一化置信度,得到每一个位置的前*J*个候选项的置信度;计算每一个位置的前*J*个候选项对应汉字与输入汉字的视觉相似度和语音相似度,并将二者加权,得到相似度;融合相似度与置信度计算每一个位置的前*J*个候选项的综合权重;以每一个位置综合权重最高的汉字作为纠错后的汉字。本发明通过对汉字的视觉相似度和语音相似度进行建模,解决了相似字错误的问题。



CN 117648923 B

1.一种适用于医疗语境的中文拼写纠错方法,其特征在于,包括以下步骤:

步骤一,将待纠错的句子以汉字为单位划分得到I个汉字,第i个汉字为C_i,$i \in [1,I]$,将I个汉字通过词表进行映射得到序列,在序列之前加上$[CLS]$,在序列之后加上$[SEP]$,得到待纠错的句子的汉字标号序列seq;

步骤二,将汉字标号序列seq输入到BERT预训练中文语言模型中得到语境信息特征$Hide$,将语境信息特征$Hide$的维度转换为$I \times 21127$,得到置信度预测pre;

步骤三,定义置信度预测pre中对应汉字C_i的置信度预测为汉字置信度预测pre_i,将汉字置信度预测pre_i中所有值从大到小排序后选取前J个值作为待纠错的句子中第i个位置处的候选汉字概率集合,将候选汉字概率集合进行归一化处理,其中待纠错的句子中第i个位置处的第j个候选汉字的归一化置信度为$pre_norm_i[j]$;

步骤四,基于编辑距离算法计算待纠错的句子中第i个汉字C_i与待纠错的句子中第i个位置处的第j个候选汉字之间的语音相似度$pho_sim_{i,j}$;

步骤五,基于编辑距离算法计算待纠错的句子中第i个汉字C_i与待纠错的句子中第i个位置处的第j个候选汉字之间的视觉相似度$vis_sim_{i,j}$;

步骤六,基于语音相似度$pho_sim_{i,j}$与视觉相似度$vis_sim_{i,j}$计算汉字C_i与待纠错的句子中第i个位置处的第j个候选汉字之间的相似度$sim_{i,j}$,基于相似度$sim_{i,j}$与归一化置信度$pre_norm_i[j]$计算待纠错的句子中第i个位置处的第j个候选汉字的综合权重$weight_{i,j}$,根据综合权重$weight_{i,j}$计算待纠错的句子中第i个位置处纠错后的汉字O_i。

2.根据权利要求1所述一种适用于医疗语境的中文拼写纠错方法,其特征在于,步骤二中所述将汉字标号序列seq输入到BERT预训练中文语言模型中得到语境信息特征$Hide$,具体指将汉字标号序列seq输入到BERT预训练中文语言模型中得到语境信息特征$Hide$:

$$Hide = BERT(seq);$$

其中,$BERT()$代表通过BERT预训练中文语言模型提取特征操作。

3.根据权利要求1所述一种适用于医疗语境的中文拼写纠错方法,其特征在于,步骤二中所述将语境信息特征$Hide$的维度转换为$I \times 21127$,得到置信度预测pre,具体指将语境信息特征$Hide$进行维度转换,得到置信度预测pre:

$$pre = Linear(Hide);$$

其中,$Linear()$代表线性变换操作,置信度预测pre的维度为$I \times 21127$。

4.根据权利要求1所述一种适用于医疗语境的中文拼写纠错方法,其特征在于,步骤三中归一化置信度$pre_norm_i[j]$的计算方法为:

$$pre\_norm_i[j] = \frac{(pre_i[j]-pre_i[J+1])}{(pre_i[1]-pre_i[J+1])}, j \in [1,J];$$

其中,$pre_i[j]$代表汉字置信度预测pre_i的向量按照数值从大到小排序后第j个值。

5.根据权利要求1所述一种适用于医疗语境的中文拼写纠错方法,其特征在于,步骤四具体包括:用每个汉字的拼音和声调编码组成该汉字的拼音序列,定义待纠错的句子中第i个汉字C_i的拼音序列为$pho(C_i)$,基于编辑距离算法计算待纠错的句子中第i个汉字C_i与待纠错的句子中第i个位置处的第j个候选汉字之间的语音相似度$pho_sim_{i,j}$:

$$pho\_sim_{i,j}=1-\frac{lev(pho(d(index_i[j])),pho(C_i))}{max(|d(index_i[j])|,|C_i|)};$$

其中,$index_i[j]$代表待纠错的句子中第i个位置处的第j个候选汉字的词表索引,$d()$代表将词表索引转换为对应汉字的解码函数,$d(index_i[j])$代表待纠错的句子中第i个位置处的第j个候选汉字,$pho(d(index_i[j]))$代表待纠错的句子中第i个位置处的第j个候选汉字的拼音序列,$lev()$代表编辑距离计算函数,$||$代表绝对值运算,max ()代表求最大值函数。

6.根据权利要求1所述一种适用于医疗语境的中文拼写纠错方法,其特征在于,步骤五具体包括:定义待纠错的句子中第i个汉字C_i的表意文字描述序列为$vis(C_i)$,基于编辑距离算法计算待纠错的句子中第i个汉字C_i与待纠错的句子中第i个位置处的第j个候选汉字之间的视觉相似度$vis_sim_{i,j}$:

$$vis\_sim_{i,j}=1-\frac{lev(vis(d(index_i[j])),vis(C_i))}{max(|d(index_i[j])|,|C_i|)};$$

其中,$index_i[j]$代表待纠错的句子中第i个位置处的第j个候选汉字的词表索引,$d()$代表将词表索引转换为对应汉字的解码函数,$d(index_i[j])$代表待纠错的句子中第i个位置处的第j个候选汉字,$vis(d(index_i[j]))$代表待纠错的句子中第i个位置处的第j个候选汉字的表意文字描述序列,$lev()$代表编辑距离计算函数,$||$代表绝对值运算,max ()代表求最大值函数。

7.根据权利要求6所述一种适用于医疗语境的中文拼写纠错方法,其特征在于,表意文字描述序列,具体指:

将每个汉字以独体字为单位进行拆分得到内部构字部件,对于不能完全拆分为独体字的汉字,将拆分剩余的笔画与最接近的独体字结合作为一个内部构字部件;

按照汉字书写规则的先后顺序对每个内部构字部件继续拆分直到得到单独的笔画;

按照拆分的顺序,构建树状结构的汉字的表意文字描述树,表意文字描述树的根节点为描述第一次拆分得到的内部构字部件的相对位置的结构信息编码,叶子节点为单个笔画的笔画编码,中间节点为描述内部构字部件间或笔画间相对位置的结构信息编码;

汉字的表意文字描述序列即遍历表意文字描述树得到的序列。

8.根据权利要求7所述一种适用于医疗语境的中文拼写纠错方法,其特征在于,遍历表意文字描述树具体指:按照前序遍历顺序遍历表意文字描述树。

9.根据权利要求1所述一种适用于医疗语境的中文拼写纠错方法,其特征在于,步骤六具体包括,计算待纠错的句子中第i个汉字C_i与待纠错的句子中第i个位置处的第j个候选汉字之间的相似度$sim_{i,j}$:

$sim_{i,j} = pho\_sim_{i,j} \times W + vis\_sim_{i,j} \times (1 - W), W \in [0,1]$;

其中,W为调节语音相似度与视觉相似度的调节因子;

综合相似度与归一化置信度得到待纠错的句子中第i个位置处的第j个候选汉字的综合权重$weight_{i,j}$:

$weight_{i,j} = sim_{i,j} \times pre\_norm_i[j]$;

则待纠错的句子中第i个位置处纠错后的汉字O_i为:

$O_i = d(index(argmax([weight_{i,1}, \dots, weight_{i,j}, \dots, weight_{i,J}])))$;

其中,$argmax()$表示选出括号中最大值的函数,$index()$表示将综合权重转换为词表索引的函数,$d()$代表将词表索引转换为对应汉字的解码函数。

一种适用于医疗语境的中文拼写纠错方法

技术领域

[0001] 本发明涉及人工智能领域,具体涉及一种适用于医疗语境的中文拼写纠错方法。

背景技术

[0002] 随着我国人口增长和老龄化,就医人数大幅增加,导致医生更多时间用于接诊,无法专注于其他工作,如写病历、开处方等。繁重的工作压力使医生在工作中出现拼写错误的几率大大提高,从而引发信息传递偏差甚至事故。比如,药品名称拼写错误可能导致患者获得错误的药物;疾病名称的纰漏可能造成误诊;手术操作的记录失误也可能严重影响治疗效果。自动拼写纠错系统可以发现文字中的错字并提出修改建议,帮助医务人员减少拼写错误,提高医疗记录的准确性,节省医生时间,将更多精力用在治疗上。

[0003] 一些研究人员尝试基于深度学习方法在医疗环境下进行拼写纠错。这种拼写纠错方法主要需要构建神经网络模型,这些神经网络模型通过大量的训练数据学习和理解语言的复杂模式,包括上下文关系、语法结构和语义含义等。通过编码器将输入的错误拼写文本编码为一个固定长度的向量,捕获文本中的重要信息,然后,解码器基于这个向量生成正确的拼写文本。训练过程中,神经网络模型通过比较生成的文本和真实的正确文本,不断调整内部参数,使得生成的文本更接近真实的正确文本。

[0004] 这些神经网络模型依赖局部上下文信息来进行预测,因为这类神经网络模型的设计使其在处理长距离依赖问题时存在困难,模型可能无法充分理解前后文中带有丰富含义的短语或者句子,导致无法对语境相关的错误进行纠正。例如,病情描述中的甲状腺功能亢进症被错误写成了甲状腺功能减退症,虽然在病情描述中根据具体症状能推测出正确的疾病名称,但是由于模型无法充分理解上下文信息中的关联性和语义含义,导致其无法正确纠正拼写错误。

[0005] 另一方面,对于相似汉字的拼写错误,例如,将窦性心律错误写成窦性心率,律与率具有相同发音,并且心率和心律都有各自实际的意义,现有的深度学习的神经网络模型在处理这种复杂的非线性关系时存在困难,导致在面对形态相似或读音相似的拼写错误时,模型可能无法做出正确的预测。

发明内容

[0006] 为解决上述问题,本发明提供一种适用于医疗语境的中文拼写纠错方法。

[0007] 该方法包括:

[0008] 步骤一,将待纠错的句子以汉字为单位划分得到I个汉字,第i个汉字为C_i,$i \in [1,I]$,将I个汉字通过词表进行映射得到序列,在序列之前加上$[CLS]$,在序列之后加上$[SEP]$,得到待纠错的句子的汉字标号序列seq;

[0009] 步骤二,将汉字标号序列seq输入到BERT预训练中文语言模型中得到语境信息特征$Hide$,将语境信息特征$Hide$的维度转换为$I \times 21127$,得到置信度预测pre;

[0010] 步骤三,定义置信度预测pre中对应汉字C_i的置信度预测为汉字置信度预测pre_i,将汉字置信度预测pre_i中所有值从大到小排序后选取前J个值作为待纠错的句子中第i个位置处的候选汉字概率集合,将候选汉字概率集合进行归一化处理,其中待纠错的句子中第i个位置处的第j个候选汉字的归一化置信度为$pre_norm_i[j]$;

[0011] 步骤四,基于编辑距离算法计算待纠错的句子中第i个汉字C_i与待纠错的句子中第i个位置处的第j个候选汉字之间的语音相似度$pho_sim_{i,j}$;

[0012] 步骤五,基于编辑距离算法计算待纠错的句子中第i个汉字C_i与待纠错的句子中第i个位置处的第j个候选汉字之间的视觉相似度$vis_sim_{i,j}$;

[0013] 步骤六,基于语音相似度$pho_sim_{i,j}$与视觉相似度$vis_sim_{i,j}$计算汉字C_i与待纠错的句子中第i个位置处的第j个候选汉字之间的相似度$sim_{i,j}$,基于相似度$sim_{i,j}$与归一化置信度$pre_norm_i[j]$计算待纠错的句子中第i个位置处的第j个候选汉字的综合权重$weight_{i,j}$,根据综合权重$weight_{i,j}$计算待纠错的句子中第i个位置处纠错后的汉字O_i。

[0014] 进一步的,步骤二中所述将汉字标号序列seq输入到BERT预训练中文语言模型中得到语境信息特征$Hide$,具体指将汉字标号序列seq输入到BERT预训练中文语言模型中得到语境信息特征$Hide$:

[0015] $$Hide = BERT(seq);$$

[0016] 其中,$BERT()$代表通过BERT预训练中文语言模型提取特征操作。

[0017] 进一步的,步骤二中所述将语境信息特征$Hide$的维度转换为$I \times 21127$,得到置信度预测pre,具体指将语境信息特征$Hide$进行维度转换,得到置信度预测pre:

[0018] $$pre = Linear(Hide);$$

[0019] 其中,$Linear()$代表线性变换操作,置信度预测pre的维度为$I \times 21127$。

[0020] 进一步的,步骤三中归一化置信度$pre_norm_i[j]$的计算方法为:

[0021] $$pre\_norm_i[j] = \frac{(pre_i[j]-pre_i[J+1])}{(pre_i[1]-pre_i[J+1])}, j \in [1,J];$$

[0022] 其中,$pre_i[j]$代表汉字置信度预测pre_i的向量按照数值从大到小排序后第j个值。

[0023] 进一步的,步骤四具体包括:用每个汉字的拼音和声调编码组成该汉字的拼音序列,定义待纠错的句子中第i个汉字C_i的拼音序列为$pho(C_i)$,基于编辑距离算法计算待纠错的句子中第i个汉字C_i与待纠错的句子中第i个位置处的第j个候选汉字之间的语音相似度$pho_sim_{i,j}$:

[0024] $$pho\_sim_{i,j} = 1 - \frac{lev(pho(d(index_i[j])),pho(C_i))}{max(|d(index_i[j])|,|C_i|)};$$

[0025] 其中,$index_i[j]$代表待纠错的句子中第i个位置处的第j个候选汉字的词表索引,$d()$代表将词表索引转换为对应汉字的解码函数,$d(index_i[j])$代表待纠错的句子中第i个位

置处的第j个候选汉字，$pho(d(index_i[j]))$代表待纠错的句子中第i个位置处的第j个候选汉字的拼音序列，$lev()$代表编辑距离计算函数，$||$代表绝对值运算，max ()代表求最大值函数。

[0026] 进一步的，步骤五具体包括：定义待纠错的句子中第i个汉字C_i的表意文字描述序列为$vis(C_i)$，基于编辑距离算法计算待纠错的句子中第i个汉字C_i与待纠错的句子中第i个位置处的第j个候选汉字之间的视觉相似度$vis_sim_{i,j}$：

[0027] $$vis\_sim_{i,j} = 1 - \frac{lev(vis(d(index_i[j])),vis(C_i))}{max(|d(index_i[j])|,|C_i|)};$$

[0028] 其中，$index_i[j]$代表待纠错的句子中第i个位置处的第j个候选汉字的词表索引，$d()$代表将词表索引转换为对应汉字的解码函数，$d(index_i[j])$代表待纠错的句子中第i个位置处的第j个候选汉字，$vis(d(index_i[j]))$代表待纠错的句子中第i个位置处的第j个候选汉字的表意文字描述序列，$lev()$代表编辑距离计算函数，$||$代表绝对值运算，max ()代表求最大值函数。

[0029] 进一步的，表意文字描述序列，具体指：

[0030] 将每个汉字以独体字为单位进行拆分得到内部构字部件，对于不能完全拆分为独体字的汉字，将拆分剩余的笔画与最接近的独体字结合作为一个内部构字部件；

[0031] 按照汉字书写规则的先后顺序对每个内部构字部件继续拆分直到得到单独的笔画；

[0032] 按照拆分的顺序，构建树状结构的汉字的表意文字描述树，表意文字描述树的根节点为描述第一次拆分得到的内部构字部件的相对位置的结构信息编码，叶子节点为单个笔画的笔画编码，中间节点为描述内部构字部件间或笔画间相对位置的结构信息编码；

[0033] 汉字的表意文字描述序列即遍历表意文字描述树得到的序列。

[0034] 进一步的，遍历表意文字描述树具体指：按照前序遍历顺序遍历表意文字描述树。

[0035] 进一步的，步骤六具体包括，计算待纠错的句子中第i个汉字C_i与待纠错的句子中第i个位置处的第j个候选汉字之间的相似度$sim_{i,j}$：

[0036] $$sim_{i,j} = pho\_sim_{i,j} \times W + vis\_sim_{i,j} \times (1 - W), W \in [0,1];$$

[0037] 其中，W为调节语音相似度与视觉相似度的调节因子；

[0038] 综合相似度与归一化置信度得到待纠错的句子中第i个位置处的第j个候选汉字的综合权重$weight_{i,j}$：

[0039] $$weight_{i,j} = sim_{i,j} \times pre\_norm_i[j];$$

[0040] 则待纠错的句子中第i个位置处纠错后的汉字O_i为：

[0041] $$O_i = d(index(argmax([weight_{i,1}, \dots, weight_{i,j}, \dots, weight_{i,J}])));$$

[0042] 其中，$argmax()$表示选出括号中最大值的函数，$index()$表示将综合权重转换为词表索引的函数，$d()$代表将词表索引转换为对应汉字的解码函数。

[0043] 本申请实施例中提供的一个或多个技术方案,至少具有如下技术效果或优点:

[0044] 本发明提供的拼写纠错方法基于语境置信度和汉字相似度,通过引入BERT预训练中文语言模型,从而引入预训练过程中背景知识,并让输入的句子在此基础上进行编码,从而融入当前的语境特征,解决了部分正确但不合适的词语难以识别的问题。同时,通过对文本结构,即汉字的视觉相似度和语音相似度进行建模,从而帮助模型识别相似的错别字,解决了相似字错误的问题。

附图说明

[0045] 图1为本发明实施例提供的两个内部构字部件为由左至右关系的示意图;

[0046] 图2为本发明实施例提供的两个内部构字部件为由上至下关系的示意图;

[0047] 图3为本发明实施例提供的三个内部构字部件为由左至右关系的示意图;

[0048] 图4为本发明实施例提供的三个内部构字部件为由上至下关系的示意图;

[0049] 图5为本发明实施例提供的两个内部构字部件为由外而内关系的示意图;

[0050] 图6为本发明实施例提供的两个内部构字部件为三面包围且下方开口关系的示意图;

[0051] 图7为本发明实施例提供的两个内部构字部件为三面包围且上方开口关系的示意图;

[0052] 图8为本发明实施例提供的两个内部构字部件为三面包围且右方开口关系的示意图;

[0053] 图9为本发明实施例提供的两个内部构字部件为由左上至右下的两面包围关系的示意图;

[0054] 图10为本发明实施例提供的两个内部构字部件为由右上至左下的两面包围关系的示意图;

[0055] 图11为本发明实施例提供的两个内部构字部件为由左下至右上的两面包围关系的示意图;

[0056] 图12为本发明实施例提供的两个内部构字部件部分重叠关系的示意图;

[0057] 图13为本发明实施例提供的汉字“由”的表意文字描述树。

具体实施方式

[0058] 以下结合附图和具体实施例,对本发明进行详细说明,在详细说明本发明各实施例的技术方案前,对所涉及的名词和术语进行解释说明,在本说明书中,名称相同或标号相同的部件代表相似或相同的结构,且仅限于示意的目的。

[0059] 本发明对用户输入的句子进行纠正,并输出最可能的纠正结果。具体来说,本发明首先对输入的句子进行分词,得到单个汉字的序列,然后加入特殊标识符,得到待纠错的句子的汉字标号序列;将待纠错的句子的汉字标号序列输入到BERT预训练中文语言模型中,得到BERT预训练中文语言模型给出的语境信息特征,将语境信息特征进行线性变换使其尺寸对齐词表;计算每一个位置的前J个候选项的归一化置信度,得到每一个位置的前J个候选项的置信度;计算每一个位置的前J个候选项对应汉字与输入汉字的视觉相似度和语音

相似度,并将二者加权,得到每一个位置的前J个候选项的相似度;融合相似度与置信度计算每一个位置的前J个候选项的综合权重;以每一个位置综合权重最高的汉字作为纠错后的汉字。

[0060] 本发明提供的方法具体包括:

[0061] 1.对句子进行分词

[0062] 为了使BERT预训练中文语言模型能够处理以字符组成的句子,需要对句子进行分词。

[0063] 将待纠错的句子$C_1,\dots C_i,\dots,C_I$以汉字为单位进行划分,将划分得到的I个汉字通过词表进行映射得到序列$[c_1,\dots,c_i,\dots,c_I]$,$i\in[1,I]$。$I$代表待纠错的句子中包括的汉字的数量,$c_i$代表待纠错的句子中第$i$个汉字$C_i$映射后得到的汉字数字标号。词表是一个将汉字映射成为数字的表,BERT预训练中文语言模型的词表对应的数字范围为0-21127,则$c_i\in[0,21127]$。

[0064] 基于BERT预训练中文语言模型的格式要求,在序列$[c_1,\dots,c_i,\dots,c_I]$之前加上$[CLS]$标记,之后加上$[SEP]$标记,得到待纠错的句子的汉字标号序列

$seq=[[CLS],c_1,\dots,c_i,\dots,c_I,[SEP]]$。

[0065] 2.获取包含语境信息的特征

[0066] BERT预训练中文语言模型可以建模上下文语义信息,将汉字标号序列seq输入到BERT预训练中文语言模型中得到语境信息特征$Hide$:

[0067] $$Hide=BERT(seq);$$

[0068] 其中,$BERT()$代表通过BERT预训练中文语言模型提取特征操作,语境信息特征$Hide$是一个维度为$I\times768$的向量,768是BERT预训练中文语言模型的输出维度。

[0069] 为了后续步骤中能够计算待纠错的句子每个位置上可能出现的汉字的概率,需要将语境信息特征$Hide$进行维度转换,得到置信度预测pre:

[0070] $$pre=Linear(Hide);$$

[0071] 其中,$Linear()$代表线性变换操作,置信度预测pre的维度为$I\times21127$。

[0072] 3.计算置信度

[0073] 定义置信度预测pre中对应待纠错的句子中第i个汉字C_i的置信度预测为汉字置信度预测pre_i,汉字置信度预测pre_i为长度21127的向量,向量中每个值代表BERT预训练中文语言模型的词表中某个值对应的汉字作为第i个汉字C_i的概率。

[0074] 将汉字置信度预测pre_i的向量中所有值按照数值从大到小排序后选取前J个值组成集合作为待纠错的句子中第i个位置处的候选汉字概率集合,候选汉字概率集合对应的汉字是BERT预训练中文语言模型预测出的最有可能作为待纠错的句子中第i个汉字C_i的J个汉字,候选汉字概率集合中第j个值对应的汉字为待纠错的句子中第i个位置处的第j个候选汉字。将候选汉字概率集合进行归一化处理,待纠错的句子中第i个位置处的第j个候选汉字

的归一化置信度$pre_norm_i[j]$为：

[0075] $$pre\_norm_i[j]=\frac{(pre_i[j]-pre_i[J+1])}{(pre_i[1]-pre_i[J+1])},j\in[1,J];$$

[0076] 其中，$pre_i[j]$代表汉字置信度预测pre_i的向量按照数值从大到小排序后第j个值。

[0077] 本发明计算归一化置信度时仅考虑汉字置信度预测pre_i中数值最大的J个值作为候选值，而非词表中的所有值。该设计的目的是拉开置信度差距，因为汉字置信度预测pre_i的前几个候选值的值较为接近，若利用词表中的所有值归一化，各候选值计算得到的归一化置信度仍将过于接近。

[0078] 4.计算语音相似度

[0079] 汉字的发音可以通过其对应的拼音和声调直接表示。本发明通过将汉字转换为对应的拼音序列，来判断汉字的语音相似度，每个汉字的拼音序列定义为拼音和声调编码组成。本实施例中，声调编码为数字。例如，汉字“医”的拼音序列为“yi1”，其中“1”表示声调第一声的编码。定义待纠错的句子中第i个汉字C_i的拼音序列为$pho(C_i)$。

[0080] 基于编辑距离算法计算待纠错的句子中第i个汉字C_i与待纠错的句子中第i个位置处的第j个候选汉字之间的语音相似度$pho_sim_{i,j}$：

[0081] $$pho\_sim_{i,j}=1-\frac{lev(pho(d(index_i[j])),pho(C_i))}{max(|d(index_i[j])|,|C_i|)};$$

[0082] 其中，$index_i[j]$代表待纠错的句子中第i个位置处的第j个候选汉字的词表索引，$d()$代表将词表索引转换为对应汉字的解码函数，$d(index_i[j])$代表待纠错的句子中第i个位置处的第j个候选汉字，$pho(d(index_i[j]))$代表待纠错的句子中第i个位置处的第j个候选汉字的拼音序列，$lev()$代表编辑距离计算函数，$||$代表绝对值运算，$\max()$代表求最大值函数。

[0083] 5.计算视觉相似度

[0084] 本发明采用表意文字描述序列来表示汉字的视觉信息，表意文字描述序列包括若干个描述汉字内部构字部件的相对位置的结构信息编码及汉字笔画编码，笔画是按照汉字书写规则进行排序的。基于汉字结构信息与汉字笔画信息，每个表意文字描述序列能与其所描述的汉字一一对应。

[0085] 本发明基于十二种结构信息编码描述汉字内部构字部件的相对位置，从而精确表示汉字的结构信息。图1示出了两个内部构字部件为由左至右关系的示意图，图2示出了两个内部构字部件为由上至下关系的示意图，图3示出了三个内部构字部件为由左至右关系的示意图，图4示出了三个内部构字部件为由上至下关系的示意图，图5示出了两个内部构字部件为由外而内关系的示意图，图6示出了两个内部构字部件为三面包围，下方开口关系的示意图，图7示出了两个内部构字部件为三面包围，上方开口关系的示意图，图8示出了两个内部构字部件为三面包围，右方开口关系的示意图，图9示出了两个内部构字部件为由左上至右下的两面包围关系的示意图，图10示出了两个内部构字部件为由右上至左下的两面包围关系的示意图，图11示出了两个内部构字部件为由左下至右上的两面包围关系的示意图，图12示出了两个内部构字部件部分重叠关系的示意图。

[0086] 将每个汉字以独体字为单位进行拆分得到内部构字部件,对于不能完全拆分为独体字的汉字,将拆分剩余的笔画与最接近的独体字结合作为一个内部构字部件,最接近的独体字指与拆分剩余的笔画在书写位置上最接近的独体字。按照汉字书写规则的先后顺序对每个内部构字部件继续拆分直到得到单独的笔画。按照拆分的顺序,构建树状结构的汉字的表意文字描述树,表意文字描述树的根节点为描述第一次拆分得到的内部构字部件的相对位置的结构信息编码,叶子节点为单个笔画的笔画编码,中间节点为描述内部构字部件间或笔画间相对位置的结构信息编码。汉字的表意文字描述序列即遍历表意文字描述树得到的序列。

[0087] 本实施例选用前序遍历顺序遍历表意文字描述树。图13示出了汉字“由”的表意文字描述树。第一次拆分汉字“由”,得到第一内部构字部件“冂”与第二内部构字部件“土”,第一内部构字部件“冂”与第二内部构字部件“土”为图12对应的部分重叠关系,因此汉字“由”的表意文字描述树的根节点为图12对应的部分重叠关系的结构信息编码;对第一内部构字部件“冂”进行拆分得到笔画“丨”与笔画“冖”,笔画“丨”与笔画“冖”为图1对应的由左至右关系;对第二内部构字部件“土”进行拆分得到第三内部构字部件“十”与笔画“一”,第三内部构字部件“十”与笔画“一”为图2对应的由上至下关系;对第三内部构字部件“十”进行拆分得到笔画“一”与笔画“丨”,笔画“一”与笔画“丨”为图12对应的部分重叠关系。至此,汉字“由”完全拆分为单独的笔画,得到如图13所示的表意文字描述树,图13的上部为对汉字“由”的表意文字描述树进行先序遍历得到的序列。

[0088] 定义待纠错的句子中第i个汉字C_i的表意文字描述序列为$vis(C_i)$,基于编辑距离算法计算待纠错的句子中第i个汉字C_i与待纠错的句子中第i个位置处的第j个候选汉字之间的视觉相似度$vis_sim_{i,j}$:

[0089] $$vis\_sim_{i,j} = 1 - \frac{lev(vis(d(index_i[j])),vis(C_i))}{max(|d(index_i[j])|,|C_i|)};$$

[0090] 其中,$vis(d(index_i[j]))$代表待纠错的句子中第i个位置处的第j个候选汉字的表意文字描述序列。

[0091] 6.句子纠错

[0092] 计算待纠错的句子中第i个汉字C_i与待纠错的句子中第i个位置处的第j个候选汉字之间的相似度$sim_{i,j}$:

[0093] $$sim_{i,j} = pho\_sim_{i,j} \times W + vis\_sim_{i,j} \times (1 - W), W \in [0,1];$$

[0094] 其中,W为调节语音相似度与视觉相似度的调节因子。

[0095] 综合相似度与归一化置信度得到待纠错的句子中第i个位置处的第j个候选汉字的综合权重$weight_{i,j}$:

[0096] $$weight_{i,j} = sim_{i,j} \times pre\_norm_i[j]。$$

[0097] 则待纠错的句子中第i个位置处纠错后的汉字O_i为:

[0098] $$O_i = d(index(argmax([weight_{i,1}, \dots, weight_{i,j}, \dots, weight_{i,J}])));$$

[0099] 其中,$argmax()$表示选出括号中最大值的函数,$index()$表示将综合权重转换为词

表索引的函数，$index(argmax([weight_{i,1},\dots,weight_{i,j},\dots,weight_{i,J}]))$表示从$[weight_{i,1},\dots,weight_{i,j},\dots,weight_{i,J}]$中选取值最大的综合权重，并计算该综合权重的词表索引。

[0100] 纠错后的汉字O_i就是纠错后的句子中第i个位置所对应汉字，纠错后的汉字O_i可能与汉字C_i相同也可能不同，纠错后的汉字O_i与汉字C_i相同表示纠错后的句子中第i个位置未做修改；纠错后的汉字O_i与汉字C_i不相同表示待纠错的句子中第i个位置做过修改。

[0101] 每个纠错后的汉字按顺序组成$O_1,\dots O_i,\dots,O_I$，即纠错后的句子。

[0102] 以上所述实施方式仅仅是对本发明的优选实施方式进行描述，并非对本发明的范围进行限定，在不脱离本发明设计精神的前提下，本领域普通技术人员对本发明的技术方案做出的各种变形和改进，均应落入本发明的权利要求书确定的保护范围内。

图1

图2

图3

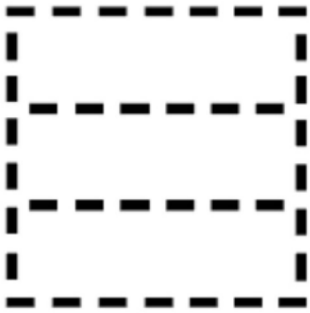


图4

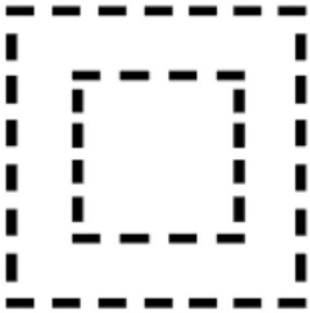


图5

图6

图7

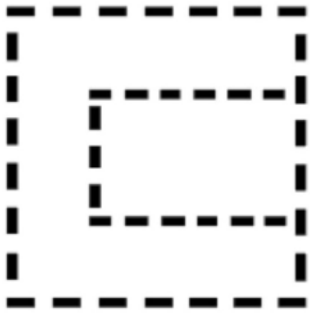


图8

图9

图10

图11

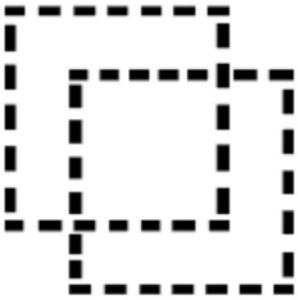


图12

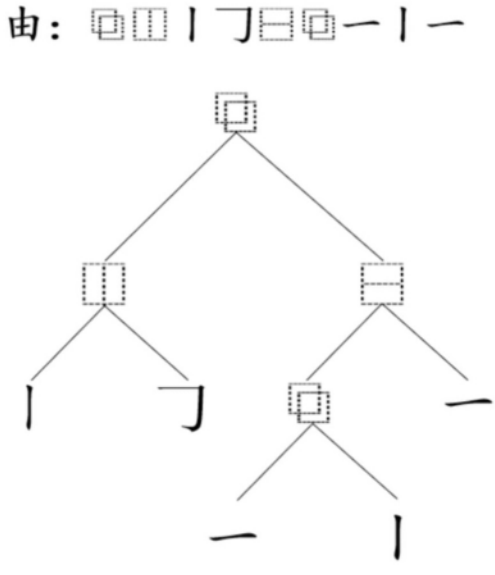


图13