(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
2022 年 5 月 27 日 (27.05.2022)

WIPO | PCT

(10) 国际公布号
WO 2022/105083 A1

(51) 国际专利分类号:
G06F 40/232 (2020.01) *G06F 40/242* (2020.01)

(21) 国际申请号: PCT/CN2021/084546

(22) 国际申请日: 2021 年 3 月 31 日 (31.03.2021)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202011302530.3 2020年11月19日 (19.11.2020) CN

(71) 申请人: 平安科技（深圳）有限公司 (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) [CN/CN]; 中国广东省深圳市福田区福田街道福安社区益田路5033号平安金融中心23楼, Guangdong 518000 (CN)。

(72) 发明人: 郑立颖 (ZHENG, Liying); 中国广东省深圳市福田区福田街道福安社区益田路5033号平安金融中心23楼, Guangdong 518000 (CN)。 徐亮 (XU, Liang); 中国广东省深圳市福田区福田街道福安社区益田路5033号平安金融中心23楼, Guangdong 518000 (CN)。 肖京 (XIAO, Jing); 中国广东省深圳市福田区福田街道福安社区益田路5033号平安金融中心23楼, Guangdong 518000 (CN)。

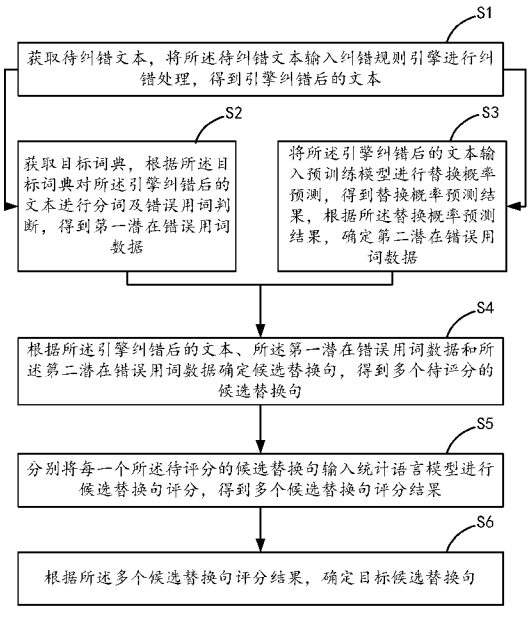
(74) 代理人: 深圳市明日今典知识产权代理事务所（普通合伙） (SHENZHEN MINGRIJINDIAN INTELLECTUAL PROPERTY AGENCY FIRM (GENERAL)); 中国广东省深圳市南山区南头街道安乐社区关口二路15号智恒产业园30栋4层406室, Guangdong 518000 (CN)。

(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,

(54) **Title:** TEXT ERROR CORRECTION METHOD AND APPARATUS, DEVICE, AND MEDIUM

(54) **发明名称:** 文本纠错方法、装置、设备及介质

附图



S1 Obtain text to be corrected, and input said text into an error correction rule engine for error correction to obtain text corrected by the engine

S2 Obtain a target dictionary, and perform word segmentation and error word determination on the text, corrected by the engine, to obtain first potential error word data

S3 Input the text corrected by the engine into a pre-training model for replacement probability prediction to obtain a replacement probability prediction result, and determine second potential error word data according to the replacement probability prediction result

S4 Determine candidate replacement sentences according to the text corrected by the engine, the first potential error word data, and the second potential error word data to obtain a plurality of candidate replacement sentences to be scored

S5 Separately input each candidate replacement sentences into a statistical language model for candidate replacement sentence scoring to obtain a plurality of candidate replacement sentence scoring results

S6 Determine a target candidate replacement sentence according to the plurality of candidate replacement sentence scoring results

图 1

(57) **Abstract:** A text error correction method and apparatus, a device, and a medium, relating to the technical field of artificial intelligence. The method comprises: performing, according to a target dictionary, word segmentation and error word determination on text, corrected by an engine, to obtain first potential error word data; inputting the text corrected by the engine into a pre-training model for replacement probability prediction to obtain a replacement probability prediction result, and determining second potential error word data according to the replacement probability prediction result; determining candidate replacement sentences according to the text

[见续页]

WO 2022/105083 A1

BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, 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, GH, GM, KE, LR, LS, MW, MZ, NA, RW, 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, 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))。

corrected by the engine, the first potential error word data, and the second potential error word data to obtain a plurality of candidate replacement sentences to be scored; separately inputting each candidate replacement sentence into a statistical language model for candidate replacement sentence scoring to obtain a plurality of candidate replacement sentence scoring results; and determining a target candidate replacement sentence according to the plurality of candidate replacement sentence scoring results. Thus, both errors within and outside a rule can be recognized, and the precision of text error correction is improved.

(57) 摘要：一种文本纠错方法、装置、设备及介质，涉及人工智能技术领域，其中方法包括：根据目标词典对引擎纠错后的文本进行分词及错误用词判断得到第一潜在错误用词数据；将引擎纠错后的文本输入预训练模型进行替换概率预测得到替换概率预测结果，根据替换概率预测结果确定第二潜在错误用词数据；根据引擎纠错后的文本、第一潜在错误用词数据和第二潜在错误用词数据确定候选替换句得到多个待评分的候选替换句；分别将每一个待评分的候选替换句输入统计语言模型进行候选替换句评分得到多个候选替换句评分结果；根据多个候选替换句评分结果确定目标候选替换句。从而实现了对规则以内和规则以外的错误情况的识别，提高了文本纠错的准确性。

说明书

发明名称：文本纠错方法、装置、设备及介质

[0001] 本申请要求于2020年11月19日提交中国专利局、申请号为2020113025303，发明名称为“文本纠错方法、装置、设备及介质”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。

技术领域

[0002] 本申请涉及到人工智能技术领域，特别是涉及到一种文本纠错方法、装置、设备及介质。

背景技术

[0003] 文本纠错指的是对自然语言在使用过程中出现的问题自动进行识别和纠正，比如，用字错误（痛点写成通点）、语法错误（的地得混用）、用词搭配错误（辅助决策写成扶助决策）、多字、漏字等。

[0004] 因为特定场景的相关术语、专业术语，比如，机构的缩写：广东分公司写成广分，公司内部的缩写用语：会议纪要写成会纪，导致采用通用语料（如维基百科中文语料、人民日报中文语料）训练出的文本纠错模型的纠错效果不会太好。

[0005] 发明人意识到目前已有的纠错技术针对特定场景大都结合规则引擎，但是仅依赖规则引擎会造成文本纠错模型覆盖率有限，针对规则以外的错误情况无法处理，同时规则引擎也会容易引起误判。

发明概述

技术问题

[0006] 现有技术的纠错技术仅依赖规则引擎会造成文本纠错模型覆盖率有限，针对规则以外的错误情况无法处理，同时规则引擎也会容易引起误判的技术问题。

问题的解决方案

技术解决方案

[0007] 本申请的主要目的为提供一种文本纠错方法、装置、设备及介质，旨在解决现有技术的纠错技术仅依赖规则引擎会造成文本纠错模型覆盖率有限，针对规则

以外的错误情况无法处理，同时规则引擎也会容易引起误判的技术问题。

[0008] 为了实现上述发明目的，本申请提出一种文本纠错方法，所述方法包括：

[0009] 获取待纠错文本，将所述待纠错文本输入纠错规则引擎进行纠错处理，得到引擎纠错后的文本；

[0010] 获取目标词典，根据所述目标词典对所述引擎纠错后的文本进行分词及错误用词判断，得到第一潜在错误用词数据；

[0011] 将所述引擎纠错后的文本输入预训练模型进行替换概率预测，得到替换概率预测结果，根据所述替换概率预测结果，确定第二潜在错误用词数据；

[0012] 根据所述引擎纠错后的文本、所述第一潜在错误用词数据和所述第二潜在错误用词数据确定候选替换句，得到多个待评分的候选替换句；

[0013] 分别将每一个所述待评分的候选替换句输入统计语言模型进行候选替换句评分，得到多个候选替换句评分结果；

[0014] 根据所述多个候选替换句评分结果，确定目标候选替换句。

[0015] 本申请还提出了一种文本纠错装置，所述装置包括：

[0016] 引擎纠错模块，用于获取待纠错文本，将所述待纠错文本输入纠错规则引擎进行纠错处理，得到引擎纠错后的文本；

[0017] 第一潜在错误用词数据确定模块，用于获取目标词典，根据所述目标词典对所述引擎纠错后的文本进行分词及错误用词判断，得到第一潜在错误用词数据；

[0018] 第二潜在错误用词数据确定模块，用于将所述引擎纠错后的文本输入预训练模型进行替换概率预测，得到替换概率预测结果，根据所述替换概率预测结果，确定第二潜在错误用词数据；

[0019] 待评分的候选替换句确定模块，用于根据所述引擎纠错后的文本、所述第一潜在错误用词数据和所述第二潜在错误用词数据确定候选替换句，得到多个待评分的候选替换句；

[0020] 候选替换句评分结果确定模块，用于分别将每一个所述待评分的候选替换句输入统计语言模型进行候选替换句评分，得到多个候选替换句评分结果；

[0021] 目标候选替换句确定模块，用于根据所述多个候选替换句评分结果，确定目标候选替换句。

[0022] 本申请还提出了一种计算机设备，包括存储器和处理器，所述存储器存储有计算机程序，所述处理器执行所述计算机程序时实现如下方法步骤：

[0023] 获取待纠错文本，将所述待纠错文本输入纠错规则引擎进行纠错处理，得到引擎纠错后的文本；

[0024] 获取目标词典，根据所述目标词典对所述引擎纠错后的文本进行分词及错误用词判断，得到第一潜在错误用词数据；

[0025] 将所述引擎纠错后的文本输入预训练模型进行替换概率预测，得到替换概率预测结果，根据所述替换概率预测结果，确定第二潜在错误用词数据；

[0026] 根据所述引擎纠错后的文本、所述第一潜在错误用词数据和所述第二潜在错误用词数据确定候选替换句，得到多个待评分的候选替换句；

[0027] 分别将每一个所述待评分的候选替换句输入统计语言模型进行候选替换句评分，得到多个候选替换句评分结果；

[0028] 根据所述多个候选替换句评分结果，确定目标候选替换句。

[0029] 本申请还提出了一种计算机可读存储介质，其上存储有计算机程序，所述计算机程序被处理器执行时实现如下方法步骤：

[0030] 获取待纠错文本，将所述待纠错文本输入纠错规则引擎进行纠错处理，得到引擎纠错后的文本；

[0031] 获取目标词典，根据所述目标词典对所述引擎纠错后的文本进行分词及错误用词判断，得到第一潜在错误用词数据；

[0032] 将所述引擎纠错后的文本输入预训练模型进行替换概率预测，得到替换概率预测结果，根据所述替换概率预测结果，确定第二潜在错误用词数据；

[0033] 根据所述引擎纠错后的文本、所述第一潜在错误用词数据和所述第二潜在错误用词数据确定候选替换句，得到多个待评分的候选替换句；

[0034] 分别将每一个所述待评分的候选替换句输入统计语言模型进行候选替换句评分，得到多个候选替换句评分结果；

[0035] 根据所述多个候选替换句评分结果，确定目标候选替换句。

发明的有益效果

有益效果

[0036] 本申请的一种文本纠错方法、装置、设备及介质，通过在错误检测阶段使用规则引擎、目标词典及预训练模型提高错误位置识别的可能性，实现了对规则以内和规则以外的错误情况的识别，从而提高了覆盖率；在错误纠正阶段，根据引擎纠错后的文本、第一潜在错误用词数据和第二潜在错误用词数据确定候选替换句得到多个待评分的候选替换句，然后再结合统计语言模型判断替换词在候选替换句中存在的合理程度，减少了错误检测阶段带来的误判，从而提高了文本纠错的准确性。

对附图的简要说明

附图说明

[0037] 图1为本申请一实施例的文本纠错方法的流程示意图；

[0038] 图2 为本申请一实施例的文本纠错装置的结构示意框图；

[0039] 图3 为本申请一实施例的计算机设备的结构示意框图。

[0040] 本申请目的的实现、功能特点及优点将结合实施例，参照附图做进一步说明。

发明实施例

本发明的实施方式

[0041] 为了使本申请的目的、技术方案及优点更加清楚明白，以下结合附图及实施例，对本申请进行进一步详细说明。应当理解，此处描述的具体实施例仅仅用以解释本申请，并不用于限定本申请。

[0042] 为了解决现有技术的纠错技术仅依赖规则引擎会造成文本纠错模型覆盖率有限，针对规则以外的错误情况无法处理，同时规则引擎也会容易引起误判的技术问题，本申请提出了文本纠错方法，所述方法应用于人工智能技术领域，所述方法进一步应用于人工智能的自然语言处理技术领域。所述文本纠错方法通过先进采用纠错规则引擎进行纠错，再用词典找出第一潜在错误用词数据和用预训练模型确定第二潜在错误用词数据，然后根据第一潜在错误用词数据和第二潜在错误用词数据确定候选替换句，对候选替换句进行评分，根据评分确定文本纠错结果，实现了对规则以内和规则以外的错误情况的识别，从而提高了覆盖率，提高了文本纠错的准确性。

[0043] 参照图1，本申请实施例中提供一种文本纠错方法，所述方法包括：

[0044] S1：获取待纠错文本，将所述待纠错文本输入纠错规则引擎进行纠错处理，得到引擎纠错后的文本；

[0045] S2：获取目标词典，根据所述目标词典对所述引擎纠错后的文本进行分词及错误用词判断，得到第一潜在错误用词数据；

[0046] S3：将所述引擎纠错后的文本输入预训练模型进行替换概率预测，得到替换概率预测结果，根据所述替换概率预测结果，确定第二潜在错误用词数据；

[0047] S4：根据所述引擎纠错后的文本、所述第一潜在错误用词数据和所述第二潜在错误用词数据确定候选替换句，得到多个待评分的候选替换句；

[0048] S5：分别将每一个所述待评分的候选替换句输入统计语言模型进行候选替换句评分，得到多个候选替换句评分结果；

[0049] S6：根据所述多个候选替换句评分结果，确定目标候选替换句。

[0050] 本实施例通过在错误检测阶段使用规则引擎、目标词典及预训练模型提高错误位置识别的可能性，实现了对规则以内和规则以外的错误情况的识别，从而提高了覆盖率；在错误纠正阶段，根据引擎纠错后的文本、第一潜在错误用词数据和第二潜在错误用词数据确定候选替换句得到多个待评分的候选替换句，然后再结合统计语言模型判断替换词在候选替换句中存在的合理程度，减少了错误检测阶段带来的误判，从而提高了文本纠错的准确性。

[0051] 对于S1，可以从数据库中获取待纠错文本，也可以是用户输入的待纠错文本，还可以是其他应用系统发送的待纠错文本。

[0052] 待纠错文本，是需要进行文本纠错的文本。

[0053] 其中，将所述待纠错文本输入纠错规则引擎进行错误词识别和错误词替换，得到引擎纠错后的文本。

[0054] 纠错规则引擎是采用通用语料对神经网络训练得到的模型，其中，通用语料包括但不限于：维基百科中文语料、人民日报中文语料。

[0055] 可以理解的是，所述待纠错文本和训练纠错规则引擎的语料的语言类别相同。比如，采用中文预料对神经网络训练得到的纠错规则引擎，用于对中文的待纠错文本进行错误词识别和错误词替换，在此举例不做具体限定。

[0056] 对于S2，可以从数据库中获取目标词典，也可以是用户输入的目标词典，还可

以是其他应用系统发送的目标词典。

[0057] 目标词典包括至少一个词语。

[0058] 其中，对所述引擎纠错后的文本进行分词，采用所述目标词典对分词结果进行错误用词判断，将判断为错误用词的词语放在集合中，得到第一潜在错误用词数据。也就是说，第一潜在错误用词数据是一个集合。

[0059] 对于S3，将所述引擎纠错后的文本输入预训练模型进行每个字是否被替换的替换概率预测，得到替换概率预测结果。也就是说，替换概率预测结果中包含至少一个替换概率预测值。

[0060] 根据所述替换概率预测结果中所有的替换概率预测值，确定第二潜在错误用词数据。

[0061] 预训练模型可以从现有技术中选择，也可以是基于神经网络训练得到的模型。

[0062] 可以理解的是，所述引擎纠错后的文本和训练预训练模型的文本的语言类别相同。比如，采用中文文本对神经网络训练得到的预训练模型，用于对中文的所述引擎纠错后的文本进行替换概率预测，在此举例不做具体限定。

[0063] 可以理解的是，步骤S2和步骤S3还可以同步执行，还可以按步骤S3和步骤S2的顺序异步执行，在此不做具体限定。

[0064] 对于S4，根据所述第一潜在错误用词数据和所述第二潜在错误用词数据确定候选词，然后根据所述候选词和所述引擎纠错后的文本确定候选替换句，得到多个待评分的候选替换句。多个待评分的候选替换句中包括了所有可能的替换句组合。

[0065] 对于S5，分别将每一个所述待评分的候选替换句输入统计语言模型进行候选替换句评分，也就是说，每个所述待评分的候选替换句对应一个候选替换句评分结果。

[0066] 统计语言模型可以从现有技术中选择，也可以是基于神经网络训练得到的模型。

[0067] 可以理解的是，所述待评分的候选替换句和统计语言模型的文本的语言类别相同。比如，采用中文文本对神经网络训练得到的统计语言模型，用于对中文的所述待评分的候选替换句进行候选替换句评分，在此举例不做具体限定。

[0068] 对于S6，从所述多个候选替换句评分结果中提取评分值最大的候选替换句评分结果对应的候选替换句作为目标候选替换句。比如，所述多个候选替换句评分结果为：候选替换句B的评分为80、候选替换句A的评分为70、候选替换句D的评分为60，候选替换句B的评分80分为最大值，则80分对应的候选替换句B作为目标候选替换句，在此举例不做具体限定。

[0069] 在一个实施例中，上述获取目标词典的步骤之前，包括：

[0070] S021：获取多个业务场景文本样本；

[0071] S022：对所述多个业务场景文本样本进行分词，得到待统计的词语集合；

[0072] S023：对所述待统计的词语集合中每个词语进行词频统计，得到多个待分析词语词频；

[0073] S024：获取词频阈值；

[0074] S025：判断所述多个待分析词语词频是否大于所述词频阈值；

[0075] S026：当所述待分析词语词频大于所述词频阈值时，将所述待分析词语词频对应的词语作为业务场景常用词数据；

[0076] S027：采用点间互信息和左右熵的新词发现算法对所述多个业务场景文本样本进行新词挖掘，得到业务场景新词数据；

[0077] S028：获取业务场景特定词数据和通用场景常用词数据；

[0078] S029：根据所述业务场景常用词数据、所述业务场景新词数据、所述业务场景特定词数据和所述通用场景常用词数据，确定所述目标词典。

[0079] 本实施例实现了根据所述业务场景常用词数据、所述业务场景新词数据、所述业务场景特定词数据和所述通用场景常用词数据确定所述目标词典，使目标词典覆盖了通用场景、业务场景的各种相关术语及专业术语，从而在错误检测阶段通过目标词典提高了错误位置识别的可能性，从而提高了覆盖率。

[0080] 对于S021，可以从数据库中获取多个业务场景文本样本，也可以是用户输入的多个业务场景文本样本，还可以是其他应用系统发送的多个业务场景文本样本。

[0081] 业务场景文本样本，是业务场景使用的文本数据。

[0082] 对于S022，对所述多个业务场景文本样本进行分词，将分词得到的词语放在一

个集合中，得到待统计的词语集合。

[0083] 对于S023，分别计算所述待统计的词语集合中每个词语出现的次数，得到多个目标出现次数；获取所述待统计的词语集合中词语的总数，得到目标词语总数；分别将每一个所述目标出现次数除以所述目标词语总数，得到多个待分析词语词频。也就是说，所述待统计的词语集合中每个词语对应一个待分析词语词频。

[0084] 待分析词语词频，是需要进行分析的词语词频。

[0085] 对于S024，可以从数据库中获取词频阈值，也可以是用户输入的词频阈值，还可以是其他应用系统发送的词频阈值。

[0086] 词频阈值，是一个0到1的具体数值。

[0087] 对于S025，依次判断所述多个待分析词语词频中每个所述待分析词语词频是否大于所述词频阈值。

[0088] 对于S026，当所述待分析词语词频大于所述词频阈值时，意味着所述待分析词语词频对应的词语是业务场景的常用词的概率比较大，此时将所述待分析词语词频对应的词语作为业务场景常用词数据，有利于提高业务场景常用词数据的准确性。

[0089] 对于S027，所述采用点间互信息和左右熵的新词发现算法对所述多个业务场景文本样本进行新词挖掘，得到业务场景新词数据的步骤，包括：

[0090] S0271：根据所述多个业务场景文本样本，生成n_ gram（基于概率的判别模型）词典；

[0091] 其中，根据所述多个业务场景文本样本生成n_ gram词典的方法可以从现有技术中选择，在此不做赘述。

[0092] S0272：采用点间互信息方法从所述n_ gram词典中筛选出备选的新词，得到待选择的新词数据；

[0093] 其中，采用点间互信息方法从所述n_ gram词典中筛选出备选的新词方法可以从现有技术中选择，在此不做赘述。

[0094] S0273：采用左右熵方法从所述待选择的新词数据中进行新词选择，得到业务场景新词数据。

[0095] 其中，采用左右熵方法从所述待选择的新词数据中进行新词选择方法可以从现有技术中选择，在此不做赘述。

[0096] 对于S028，可以从数据库中获取业务场景特定词数据，也可以是用户输入的业务场景特定词数据，还可以是其他应用系统发送的业务场景特定词数据。

[0097] 可以从数据库中获取通用场景常用词数据，也可以是用户输入的通用场景常用词数据，还可以是其他应用系统发送的和通用场景常用词数据。

[0098] 业务场景特定词数据是业务场景的特性形成的词语。比如，组织内部的愿景“先知、先觉、先行”，可以将“先知、先觉、先行”作为业务场景特定词，在此举例不做具体限定。

[0099] 通用场景常用词数据，是大部分场景经常用到的词语。

[0100] 对于S029，将所述业务场景常用词数据、所述业务场景新词数据、所述业务场景特定词数据和所述通用场景常用词数据放到一个集合，将得到的集合作为所述目标词典。也就是说，目标词典覆盖了通用场景、业务场景的各种相关术语及专业术语。

[0101] 在一个实施例中，上述根据所述目标词典对所述引擎纠错后的文本进行分词及错误用词判断，得到第一潜在错误用词数据的步骤，包括：

[0102] S21：对所述引擎纠错后的文本进行分词，得到多个待判定词语；

[0103] S22：判断所述多个待判定词语在所述目标词典中是否存在；

[0104] S23：当所述待判定词语在所述目标词典中不存在时，将多个所述待判定词作为所述第一潜在错误用词数据。

[0105] 本实施例实现了根据目标词典进行错误用词判断，因目标词典覆盖了通用场景、业务场景的各种相关术语及专业术语，从而提高了错误位置识别的可能性，从而提高了第一潜在错误用词数据的覆盖率。

[0106] 对于S21，对所述引擎纠错后的文本进行分词，将分词得到的词语作为待判定词语。

[0107] 待判定词语，是指需要判定是否错误用词的词语。

[0108] 对于S22，分别将所述多个待判定词语中每一个待判定词语在所述目标词典中进行查找，当在所述目标词典中查找到相同词语时意味着该待判定词语是正确

用词，当在所述目标词典中查找不到相同词语时意味着该待判定词语是错误用词，此时确定所述待判定词语在所述目标词典中不存在。

[0109] 对于S23，当所述待判定词语在所述目标词典中不存在时，意味着该待判定词语是错误用词，将所有在所述目标词典中不存在的所述待判定词语作为所述第一潜在错误用词数据。

[0110] 在一个实施例中，上述将所述引擎纠错后的文本输入预训练模型进行替换概率预测，得到替换概率预测结果的步骤之前，包括：

[0111] S031：获取多个训练样本，所述训练样本包括：训练文本样本数据、训练文本样本标定数据；

[0112] S032：将所述训练文本样本数据输入待训练的生成器进行词语替换，得到替换样本句；

[0113] S033：将所述替换样本句输入待训练的判别器进行替换概率预测，得到替换概率样本预测值，其中，所述待训练的判别器采用Electra的Discriminator；

[0114] S034：根据所述替换概率样本预测值和所述训练文本样本标定数据对所述待训练的生成器和所述待训练的判别器进行训练，并将训练后的所述待训练的判别器作为所述预训练模型。

[0115] 本实施例待训练的判别器采用Electra（Efficiently Learning an Encoder that Classifies Token Replacement
Accurately）的Discriminator，Electra相对于Bert（Bidirectional Encoder Representations
from Transformers，预训练语言表示模型）的去预测Mask的正确值，Electra则是去预测Token是不是被替换了，从而提升了训练效率；通过将训练后的所述待训练的判别器作为所述预训练模型，使预训练模型可以预测每个被替换的概率。

[0116] S031，可以从数据库中获取多个训练样本，也可以是用户输入的多个训练样本，还可以是其他应用系统发送的多个训练样本。

[0117] 每个所述训练样本包括一个训练文本样本数据和一个训练文本样本标定数据，训练文本样本标定数据是对训练文本样本数据中每个字被替换的标定值。

[0118] 训练文本样本数据，是文本数据。

[0119] 训练文本样本标定数据，是一个一维向量，每个向量元素代表训练文本样本数据中一个字被替换的标定值。

[0120] S032，将所述训练文本样本数据输入待训练的生成器进行词语替换，得到替换样本句，也就是说，每个所述训练文本样本数据对应一个替换样本句。

[0121] 优选的，待训练的生成器采用Generator模型。

[0122] 其中，将所述训练文本样本数据经过随机选择设置[MASK]，然后输入给Generator模型，Generator模型负责把[MASK]变成替换过的词。Generator模型并不像对抗神经网络那样需要等待训练的判别器中传回来的梯度，而是像Bert一样去尝试预测正确的词语。

[0123] S033，将所述替换样本句输入待训练的判别器进行替换概率预测，得到替换概率样本预测值，也就是说，每个所述替换样本句对应一个替换概率样本预测值。

[0124] Discriminator预测所述替换样本句中每个位置上的词语是不是被替换过。

[0125] S034，所述根据所述替换概率样本预测值和所述训练文本样本标定数据对所述待训练的生成器和所述待训练的判别器进行训练，并将训练后的所述待训练的判别器作为所述预训练模型的步骤，包括：

[0126] S0341：将所述替换概率样本预测值和所述训练文本样本标定数据输入损失函数进行计算，得到目标损失值，根据所述目标损失值更新所述待训练的生成器的参数和所述待训练的判别器的参数，更新后的所述待训练的生成器和所述待训练的判别器被用于下一次计算所述替换概率样本预测值；

[0127] S0342：重复执行上述方法步骤直至所述损失值达到第一收敛条件或迭代次数达到第二收敛条件，将所述目标损失值达到第一收敛条件或迭代次数达到第二收敛条件的所述待训练的判别器，确定为所述预训练模型。

[0128] 所述第一收敛条件是指相邻两次计算的目标损失值的大小满足lipschitz条件（利普希茨连续条件）。

[0129] 所述迭代次数是指所述待训练的生成器和所述待训练的判别器被用于计算所述替换概率样本预测值的次数，也就是说，计算一次，迭代次数增加1。第二收敛

条件，是预设次数值。

[0130] 所述损失函数可以从现有技术中选择，在此不做赘述。

[0131] 在一个实施例中，上述根据所述替换概率预测结果，确定第二潜在错误用词数据的步骤，包括：

[0132] S31：获取替换概率阈值；

[0133] S32：从所述替换概率预测结果中提取大于所述替换概率阈值的值，得到目标替换概率预测数据；

[0134] S33：将所述目标替换概率预测数据对应的词语作为所述第二潜在错误用词数据。

[0135] 本实施例通过将所述替换概率预测结果中大于所述替换概率阈值的值对应的词语作为所述第二潜在错误用词数据，从而减少了误判，提高了第二潜在错误用词数据的准确性。

[0136] 对于S31，可以从数据库中获取替换概率阈值，也可以是用户输入的替换概率阈值，还可以是其他应用系统发送的替换概率阈值。

[0137] 替换概率阈值，是一个0到1的具体数值。

[0138] 对于S32，从所述替换概率预测结果的所有替换概率预测值中提取大于所述替换概率阈值的替换概率预测值，将找到的替换概率预测值作为目标替换概率预测数据。也就是说目标替换概率预测数据可以有一个值，也可以有多个值，还可以有零个值。通过将大于所述替换概率阈值的替换概率预测值作为目标替换概率预测数据，删除了小于或等于所述替换概率阈值的替换概率预测值，有利于减少噪音数据，从而减少了误判，提高了第二潜在错误用词数据的准确性。

[0139] 对于S33，将所述目标替换概率预测数据对应的词语放在一个集合中，得到所述第二潜在错误用词数据。

[0140] 在一个实施例中，上述根据所述引擎纠错后的文本、所述第一潜在错误用词数据和所述第二潜在错误用词数据确定候选替换句，得到多个待评分的候选替换句的步骤，包括：

[0141] S41：获取同音字同形字字典；

[0142] S42：在所述同音字同形字字典中选取与所述第一潜在错误用词数据和所述第

二潜在错误用词数据匹配的词作为候选词，得到候选词集合；

[0143] S43：对所述候选词集合中的候选词进行随机选择，得到多个候选词分组；

[0144] S44：分别将每一个所述候选词分组对所述引擎纠错后的文本进行替换，得到所述多个待评分的候选替换句。

[0145] 本实施例通过所述引擎纠错后的文本、所述第一潜在错误用词数据和所述第二潜在错误用词数据确定候选替换句，从而为候选替换句评分提供了数据基础。

[0146] 对于S41，可以从数据库中获取同音字同形字字典，也可以是用户输入的同音字同形字字典，还可以是其他应用系统发送的同音字同形字字典。

[0147] 同音字同形字字典包括：同音字子字典、同形字子字典。同音字子字典包括：第一替换前的词语、第一替换后的词语，第一替换前的词语和第一替换后的词语的读音相同。同形字子字典包括：第二替换前的词语、第二替换后的词语，第二替换前的词语和第二替换后的词语的字形相似。

[0148] 对于S42，将所述第一潜在错误用词数据中每个词语在所述同音字同形字字典中进行匹配，将在所述同音字同形字字典中匹配到的词语作为第一候选词；将所述第二潜在错误用词数据中每个词语在所述同音字同形字字典中进行匹配，将在所述同音字同形字字典中匹配到的词语作为第二候选词；将所有第一候选词和所有第二候选词放入集合，得到候选词集合。

[0149] 对于S43，对所述候选词集合中的候选词进行随机组合，将每一个组合作为一个候选词分组。可以理解的是，多个候选词分组涵盖了对所述候选词集合中的候选词所有可能的分组。

[0150] 对于S44，分别将每一个所述候选词分组对所述引擎纠错后的文本进行替换，得到待评分的候选替换句。也就是说，每个所述候选词分组对应一个待评分的候选替换句。

[0151] 待评分的候选替换句，是指需要进行评分的候选替换句。

[0152] 在另一个实施例中，可以采用同音字同形字字典和同音字同形字字典以外的字典进行候选词匹配确定候选词集合，在此不做具体限定。

[0153] 在一个实施例中，上述根据所述多个候选替换句评分结果，确定目标候选替换句的步骤，包括：

[0154] S61：从所述多个候选替换句评分结果中提取评分值最大的候选替换句评分结果作为目标候选替换句评分结果；

[0155] S62：将所述目标候选替换句评分结果对应的候选替换句作为所述目标候选替换句。

[0156] 本实施例实现了将所述多个候选替换句评分结果中最大值对应的所述候选替换句评分结果作为目标候选替换句评分结果，将所述目标候选替换句评分结果对应的候选替换句作为所述目标候选替换句，从而进一步提高了确定的目标候选替换句的准确性。

[0157] 对于S61，从所述多个候选替换句评分结果提取最大的候选替换句评分结果，将提取的最大的候选替换句评分结果作为目标候选替换句评分结果。

[0158] 参照图2，本申请还提出了一种文本纠错装置，所述装置包括：

[0159] 引擎纠错模块100，用于获取待纠错文本，将所述待纠错文本输入纠错规则引擎进行纠错处理，得到引擎纠错后的文本；

[0160] 第一潜在错误用词数据确定模块200，用于获取目标词典，根据所述目标词典对所述引擎纠错后的文本进行分词及错误用词判断，得到第一潜在错误用词数据；

[0161] 第二潜在错误用词数据确定模块300，用于将所述引擎纠错后的文本输入预训练模型进行替换概率预测，得到替换概率预测结果，根据所述替换概率预测结果，确定第二潜在错误用词数据；

[0162] 待评分的候选替换句确定模块400，用于根据所述引擎纠错后的文本、所述第一潜在错误用词数据和所述第二潜在错误用词数据确定候选替换句，得到多个待评分的候选替换句；

[0163] 候选替换句评分结果确定模块500，用于分别将每一个所述待评分的候选替换句输入统计语言模型进行候选替换句评分，得到多个候选替换句评分结果；

[0164] 目标候选替换句确定模块600，用于根据所述多个候选替换句评分结果，确定目标候选替换句。

[0165] 本实施例通过在错误检测阶段使用规则引擎、目标词典及预训练模型提高错误位置识别的可能性，实现了对规则以内和规则以外的错误情况的识别，从而提

高了覆盖率；在错误纠正阶段，根据引擎纠错后的文本、第一潜在错误用词数据和第二潜在错误用词数据确定候选替换句得到多个待评分的候选替换句，然后再结合统计语言模型判断替换词在候选替换句中存在的合理程度，减少了错误检测阶段带来的误判，从而提高了文本纠错的准确性。

[0166] 参照图3，本申请实施例中还提供一种计算机设备，该计算机设备可以是服务器，其内部结构可以如图3所示。该计算机设备包括通过系统总线连接的处理器、存储器、网络接口和数据库。其中，该计算机设计的处理器用于提供计算和控制能力。该计算机设备的存储器包括非易失性存储介质、内存储器。该非易失性存储介质存储有操作系统、计算机程序和数据库。该内存器为非易失性存储介质中的操作系统和计算机程序的运行提供环境。该计算机设备的数据库用于储存文本纠错方法等数据。该计算机设备的网络接口用于与外部的终端通过网络连接通信。该计算机程序被处理器执行时以实现一种文本纠错方法。所述文本纠错方法，包括：获取待纠错文本，将所述待纠错文本输入纠错规则引擎进行纠错处理，得到引擎纠错后的文本；获取目标词典，根据所述目标词典对所述引擎纠错后的文本进行分词及错误用词判断，得到第一潜在错误用词数据；将所述引擎纠错后的文本输入预训练模型进行替换概率预测，得到替换概率预测结果，根据所述替换概率预测结果，确定第二潜在错误用词数据；根据所述引擎纠错后的文本、所述第一潜在错误用词数据和所述第二潜在错误用词数据确定候选替换句，得到多个待评分的候选替换句；分别将每一个所述待评分的候选替换句输入统计语言模型进行候选替换句评分，得到多个候选替换句评分结果；根据所述多个候选替换句评分结果，确定目标候选替换句。

[0167] 本实施例通过在错误检测阶段使用规则引擎、目标词典及预训练模型提高错误位置识别的可能性，实现了对规则以内和规则以外的错误情况的识别，从而提高了覆盖率；在错误纠正阶段，根据引擎纠错后的文本、第一潜在错误用词数据和第二潜在错误用词数据确定候选替换句得到多个待评分的候选替换句，然后再结合统计语言模型判断替换词在候选替换句中存在的合理程度，减少了错误检测阶段带来的误判，从而提高了文本纠错的准确性。

[0168] 本申请一实施例还提供一种计算机可读存储介质，其上存储有计算机程序，计

算机程序被处理器执行时实现一种文本纠错方法，包括步骤：获取待纠错文本，将所述待纠错文本输入纠错规则引擎进行纠错处理，得到引擎纠错后的文本；获取目标词典，根据所述目标词典对所述引擎纠错后的文本进行分词及错误用词判断，得到第一潜在错误用词数据；将所述引擎纠错后的文本输入预训练模型进行替换概率预测，得到替换概率预测结果，根据所述替换概率预测结果，确定第二潜在错误用词数据；根据所述引擎纠错后的文本、所述第一潜在错误用词数据和所述第二潜在错误用词数据确定候选替换句，得到多个待评分的候选替换句；分别将每一个所述待评分的候选替换句输入统计语言模型进行候选替换句评分，得到多个候选替换句评分结果；根据所述多个候选替换句评分结果，确定目标候选替换句。

[0169] 上述执行的文本纠错方法，通过在错误检测阶段使用规则引擎、目标词典及预训练模型提高错误位置识别的可能性，实现了对规则以内和规则以外的错误情况的识别，从而提高了覆盖率；在错误纠正阶段，根据引擎纠错后的文本、第一潜在错误用词数据和第二潜在错误用词数据确定候选替换句得到多个待评分的候选替换句，然后再结合统计语言模型判断替换词在候选替换句中存在的合理程度，减少了错误检测阶段带来的误判，从而提高了文本纠错的准确性。

[0170] 所述计算机可读存储介质可以是非易失性，也可以是易失性。

[0171] 本领域普通技术人员可以理解实现上述实施例方法中的全部或部分流程，是可以通过计算机程序来指令相关的硬件来完成，所述的计算机程序可存储于一非易失性计算机可读取存储介质中，该计算机程序在执行时，可包括如上述各方法的实施例的流程。其中，本申请所提供的和实施例中所使用的对存储器、存储、数据库或其它介质的任何引用，均可包括非易失性和/或易失性存储器。非易失性存储器可以包括只读存储器（ROM）、可编程ROM（PROM）、电可编程ROM（EPROM）、电可擦除可编程ROM（EEPROM）或闪存。易失性存储器可包括随机存取存储器（RAM）或者外部高速缓冲存储器。作为说明而非局限，RAM以多种形式可得，诸如静态RAM（SRAM）、动态RAM（DRAM）、同步DRAM（SDRAM）、双速据率SDRAM（SSRSDRAM）、增强型SDRAM（ESDRAM）、同步链路（Synchlink）DRAM（SLDRAM）、存储器总线（Rambus）直接RAM（RDRAM）、直接存储器总线

动态RAM（DRDRAM）、以及存储器总线动态RAM（RDRAM）等。

[0172] 需要说明的是，在本文中，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、装置、物品或者方法不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、装置、物品或者方法所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括该要素的过程、装置、物品或者方法中还存在另外的相同要素。

[0173] 以上所述仅为本申请的优选实施例，并非因此限制本申请的专利范围，凡是利用本申请说明书及附图内容所作的等效结构或等效流程变换，或直接或间接运用在其他相关的技术领域，均同理包括在本申请的专利保护范围内。

权利要求书

[权利要求 1] 一种文本纠错方法，其中，所述方法包括：

获取待纠错文本，将所述待纠错文本输入纠错规则引擎进行纠错处理，得到引擎纠错后的文本；

获取目标词典，根据所述目标词典对所述引擎纠错后的文本进行分词及错误用词判断，得到第一潜在错误用词数据；

将所述引擎纠错后的文本输入预训练模型进行替换概率预测，得到替换概率预测结果，根据所述替换概率预测结果，确定第二潜在错误用词数据；

根据所述引擎纠错后的文本、所述第一潜在错误用词数据和所述第二潜在错误用词数据确定候选替换句，得到多个待评分的候选替换句；

分别将每一个所述待评分的候选替换句输入统计语言模型进行候选替换句评分，得到多个候选替换句评分结果；

根据所述多个候选替换句评分结果，确定目标候选替换句。

[权利要求 2] 根据权利要求1所述的文本纠错方法，其中，所述获取目标词典的步骤之前，包括：

获取多个业务场景文本样本；

对所述多个业务场景文本样本进行分词，得到待统计的词语集合；

对所述待统计的词语集合中每个词语进行词频统计，得到多个待分析词语词频；

获取词频阈值；

判断所述多个待分析词语词频是否大于所述词频阈值；

当所述待分析词语词频大于所述词频阈值时，将所述待分析词语词频对应的词语作为业务场景常用词数据；

采用点间互信息和左右熵的新词发现算法对所述多个业务场景文本样本进行新词挖掘，得到业务场景新词数据；

获取业务场景特定词数据和通用场景常用词数据；

根据所述业务场景常用词数据、所述业务场景新词数据、所述业务场

景特定词数据和所述通用场景常用词数据，确定所述目标词典。

[权利要求 3] 根据权利要求1所述的文本纠错方法，其中，所述根据所述目标词典对所述引擎纠错后的文本进行分词及错误用词判断，得到第一潜在错误用词数据的步骤，包括：

对所述引擎纠错后的文本进行分词，得到多个待判定词语；

判断所述多个待判定词语在所述目标词典中是否存在；

当所述待判定词语在所述目标词典中不存在时，将多个所述待判定词作为所述第一潜在错误用词数据。

[权利要求 4] 根据权利要求1所述的文本纠错方法，其中，所述将所述引擎纠错后的文本输入预训练模型进行替换概率预测，得到替换概率预测结果的步骤之前，包括：

获取多个训练样本，所述训练样本包括：训练文本样本数据、训练文本样本标定数据；

将所述训练文本样本数据输入待训练的生成器进行词语替换，得到替换样本句；

将所述替换样本句输入待训练的判别器进行替换概率预测，得到替换概率样本预测值，其中，所述待训练的判别器采用Electra的Discriminator；

根据所述替换概率样本预测值和所述训练文本样本标定数据对所述待训练的生成器和所述待训练的判别器进行训练，并将训练后的所述待训练的判别器作为所述预训练模型。

[权利要求 5] 根据权利要求1所述的文本纠错方法，其中，所述根据所述替换概率预测结果，确定第二潜在错误用词数据的步骤，包括：

获取替换概率阈值；

从所述替换概率预测结果中提取大于所述替换概率阈值的值，得到目标替换概率预测数据；

将所述目标替换概率预测数据对应的词语作为所述第二潜在错误用词数据。

[权利要求 6] 根据权利要求1所述的文本纠错方法，其中，所述根据所述引擎纠错后的文本、所述第一潜在错误用词数据和所述第二潜在错误用词数据确定候选替换句，得到多个待评分的候选替换句的步骤，包括：

获取同音字同形字字典；

在所述同音字同形字字典中选取与所述第一潜在错误用词数据和所述第二潜在错误用词数据匹配的词作为候选词，得到候选词集合；

对所述候选词集合中的候选词进行随机选择，得到多个候选词分组；

分别将每一个所述候选词分组对所述引擎纠错后的文本进行替换，得到所述多个待评分的候选替换句。

[权利要求 7] 根据权利要求1所述的文本纠错方法，其中，所述根据所述多个候选替换句评分结果，确定目标候选替换句的步骤，包括：

从所述多个候选替换句评分结果中提取评分值最大的候选替换句评分结果作为目标候选替换句评分结果；

将所述目标候选替换句评分结果对应的候选替换句作为所述目标候选替换句。

[权利要求 8] 一种文本纠错装置，其中，所述装置包括：

引擎纠错模块，用于获取待纠错文本，将所述待纠错文本输入纠错规则引擎进行纠错处理，得到引擎纠错后的文本；

第一潜在错误用词数据确定模块，用于获取目标词典，根据所述目标词典对所述引擎纠错后的文本进行分词及错误用词判断，得到第一潜在错误用词数据；

第二潜在错误用词数据确定模块，用于将所述引擎纠错后的文本输入预训练模型进行替换概率预测，得到替换概率预测结果，根据所述替换概率预测结果，确定第二潜在错误用词数据；

待评分的候选替换句确定模块，用于根据所述引擎纠错后的文本、所述第一潜在错误用词数据和所述第二潜在错误用词数据确定候选替换句，得到多个待评分的候选替换句；

候选替换句评分结果确定模块，用于分别将每一个所述待评分的候选

替换句输入统计语言模型进行候选替换句评分，得到多个候选替换句评分结果；

目标候选替换句确定模块，用于根据所述多个候选替换句评分结果，确定目标候选替换句。

[权利要求 9] 一种计算机设备，包括存储器和处理器，所述存储器存储有计算机程序，其中，所述处理器执行所述计算机程序时实现如下方法步骤：

获取待纠错文本，将所述待纠错文本输入纠错规则引擎进行纠错处理，得到引擎纠错后的文本；

获取目标词典，根据所述目标词典对所述引擎纠错后的文本进行分词及错误用词判断，得到第一潜在错误用词数据；

将所述引擎纠错后的文本输入预训练模型进行替换概率预测，得到替换概率预测结果，根据所述替换概率预测结果，确定第二潜在错误用词数据；

根据所述引擎纠错后的文本、所述第一潜在错误用词数据和所述第二潜在错误用词数据确定候选替换句，得到多个待评分的候选替换句；

分别将每一个所述待评分的候选替换句输入统计语言模型进行候选替换句评分，得到多个候选替换句评分结果；

根据所述多个候选替换句评分结果，确定目标候选替换句。

[权利要求 10] 根据权利要求9所述的计算机设备，其中，所述获取目标词典的步骤之前，包括：

获取多个业务场景文本样本；

对所述多个业务场景文本样本进行分词，得到待统计的词语集合；

对所述待统计的词语集合中每个词语进行词频统计，得到多个待分析词语词频；

获取词频阈值；

判断所述多个待分析词语词频是否大于所述词频阈值；

当所述待分析词语词频大于所述词频阈值时，将所述待分析词语词频对应的词语作为业务场景常用词数据；

采用点间互信息和左右熵的新词发现算法对所述多个业务场景文本样本进行新词挖掘，得到业务场景新词数据；

获取业务场景特定词数据和通用场景常用词数据；

根据所述业务场景常用词数据、所述业务场景新词数据、所述业务场景特定词数据和所述通用场景常用词数据，确定所述目标词典。

[权利要求 11] 根据权利要求9所述的计算机设备，其中，所述根据所述目标词典对所述引擎纠错后的文本进行分词及错误用词判断，得到第一潜在错误用词数据的步骤，包括：

对所述引擎纠错后的文本进行分词，得到多个待判定词语；

判断所述多个待判定词语在所述目标词典中是否存在；

当所述待判定词语在所述目标词典中不存在时，将多个所述待判定词作为所述第一潜在错误用词数据。

[权利要求 12] 根据权利要求9所述的计算机设备，其中，所述将所述引擎纠错后的文本输入预训练模型进行替换概率预测，得到替换概率预测结果的步骤之前，包括：

获取多个训练样本，所述训练样本包括：训练文本样本数据、训练文本样本标定数据；

将所述训练文本样本数据输入待训练的生成器进行词语替换，得到替换样本句；

将所述替换样本句输入待训练的判别器进行替换概率预测，得到替换概率样本预测值，其中，所述待训练的判别器采用Electra的Discriminator；

根据所述替换概率样本预测值和所述训练文本样本标定数据对所述待训练的生成器和所述待训练的判别器进行训练，并将训练后的所述待训练的判别器作为所述预训练模型。

[权利要求 13] 根据权利要求9所述的计算机设备，其中，所述根据所述替换概率预测结果，确定第二潜在错误用词数据的步骤，包括：

获取替换概率阈值；

从所述替换概率预测结果中提取大于所述替换概率阈值的值，得到目标替换概率预测数据；

将所述目标替换概率预测数据对应的词语作为所述第二潜在错误用词数据。

[权利要求 14] 根据权利要求9所述的计算机设备，其中，所述根据所述引擎纠错后的文本、所述第一潜在错误用词数据和所述第二潜在错误用词数据确定候选替换句，得到多个待评分的候选替换句的步骤，包括：

获取同音字同形字字典；

在所述同音字同形字字典中选取与所述第一潜在错误用词数据和所述第二潜在错误用词数据匹配的词作为候选词，得到候选词集合；

对所述候选词集合中的候选词进行随机选择，得到多个候选词分组；

分别将每一个所述候选词分组对所述引擎纠错后的文本进行替换，得到所述多个待评分的候选替换句。

[权利要求 15] 一种计算机可读存储介质，其上存储有计算机程序，其中，所述计算机程序被处理器执行时实现如下方法步骤：

获取待纠错文本，将所述待纠错文本输入纠错规则引擎进行纠错处理，得到引擎纠错后的文本；

获取目标词典，根据所述目标词典对所述引擎纠错后的文本进行分词及错误用词判断，得到第一潜在错误用词数据；

将所述引擎纠错后的文本输入预训练模型进行替换概率预测，得到替换概率预测结果，根据所述替换概率预测结果，确定第二潜在错误用词数据；

根据所述引擎纠错后的文本、所述第一潜在错误用词数据和所述第二潜在错误用词数据确定候选替换句，得到多个待评分的候选替换句；

分别将每一个所述待评分的候选替换句输入统计语言模型进行候选替换句评分，得到多个候选替换句评分结果；

根据所述多个候选替换句评分结果，确定目标候选替换句。

[权利要求 16] 根据权利要求15所述的计算机可读存储介质，其中，所述获取目标词

典的步骤之前，包括：

获取多个业务场景文本样本；

对所述多个业务场景文本样本进行分词，得到待统计的词语集合；

对所述待统计的词语集合中每个词语进行词频统计，得到多个待分析词语词频；

获取词频阈值；

判断所述多个待分析词语词频是否大于所述词频阈值；

当所述待分析词语词频大于所述词频阈值时，将所述待分析词语词频对应的词语作为业务场景常用词数据；

采用点间互信息和左右熵的新词发现算法对所述多个业务场景文本样本进行新词挖掘，得到业务场景新词数据；

获取业务场景特定词数据和通用场景常用词数据；

根据所述业务场景常用词数据、所述业务场景新词数据、所述业务场景特定词数据和所述通用场景常用词数据，确定所述目标词典。

[权利要求 17] 根据权利要求15所述的计算机可读存储介质，其中，所述根据所述目标词典对所述引擎纠错后的文本进行分词及错误用词判断，得到第一潜在错误用词数据的步骤，包括：

对所述引擎纠错后的文本进行分词，得到多个待判定词语；

判断所述多个待判定词语在所述目标词典中是否存在；

当所述待判定词语在所述目标词典中不存在时，将多个所述待判定词作为所述第一潜在错误用词数据。

[权利要求 18] 根据权利要求15所述的计算机可读存储介质，其中，所述将所述引擎纠错后的文本输入预训练模型进行替换概率预测，得到替换概率预测结果的步骤之前，包括：

获取多个训练样本，所述训练样本包括：训练文本样本数据、训练文本样本标定数据；

将所述训练文本样本数据输入待训练的生成器进行词语替换，得到替换样本句；

将所述替换样本句输入待训练的判别器进行替换概率预测，得到替换概率样本预测值，其中，所述待训练的判别器采用Electra的Discriminator；

根据所述替换概率样本预测值和所述训练文本样本标定数据对所述待训练的生成器和所述待训练的判别器进行训练，并将训练后的所述待训练的判别器作为所述预训练模型。

[权利要求 19] 根据权利要求15所述的计算机可读存储介质，其中，所述根据所述替换概率预测结果，确定第二潜在错误用词数据的步骤，包括：

获取替换概率阈值；

从所述替换概率预测结果中提取大于所述替换概率阈值的值，得到目标替换概率预测数据；

将所述目标替换概率预测数据对应的词语作为所述第二潜在错误用词数据。

[权利要求 20] 根据权利要求15所述的计算机可读存储介质，其中，所述根据所述引擎纠错后的文本、所述第一潜在错误用词数据和所述第二潜在错误用词数据确定候选替换句，得到多个待评分的候选替换句的步骤，包括：

获取同音字同形字字典；

在所述同音字同形字字典中选取与所述第一潜在错误用词数据和所述第二潜在错误用词数据匹配的词作为候选词，得到候选词集合；

对所述候选词集合中的候选词进行随机选择，得到多个候选词分组；

分别将每一个所述候选词分组对所述引擎纠错后的文本进行替换，得到所述多个待评分的候选替换句。

附图

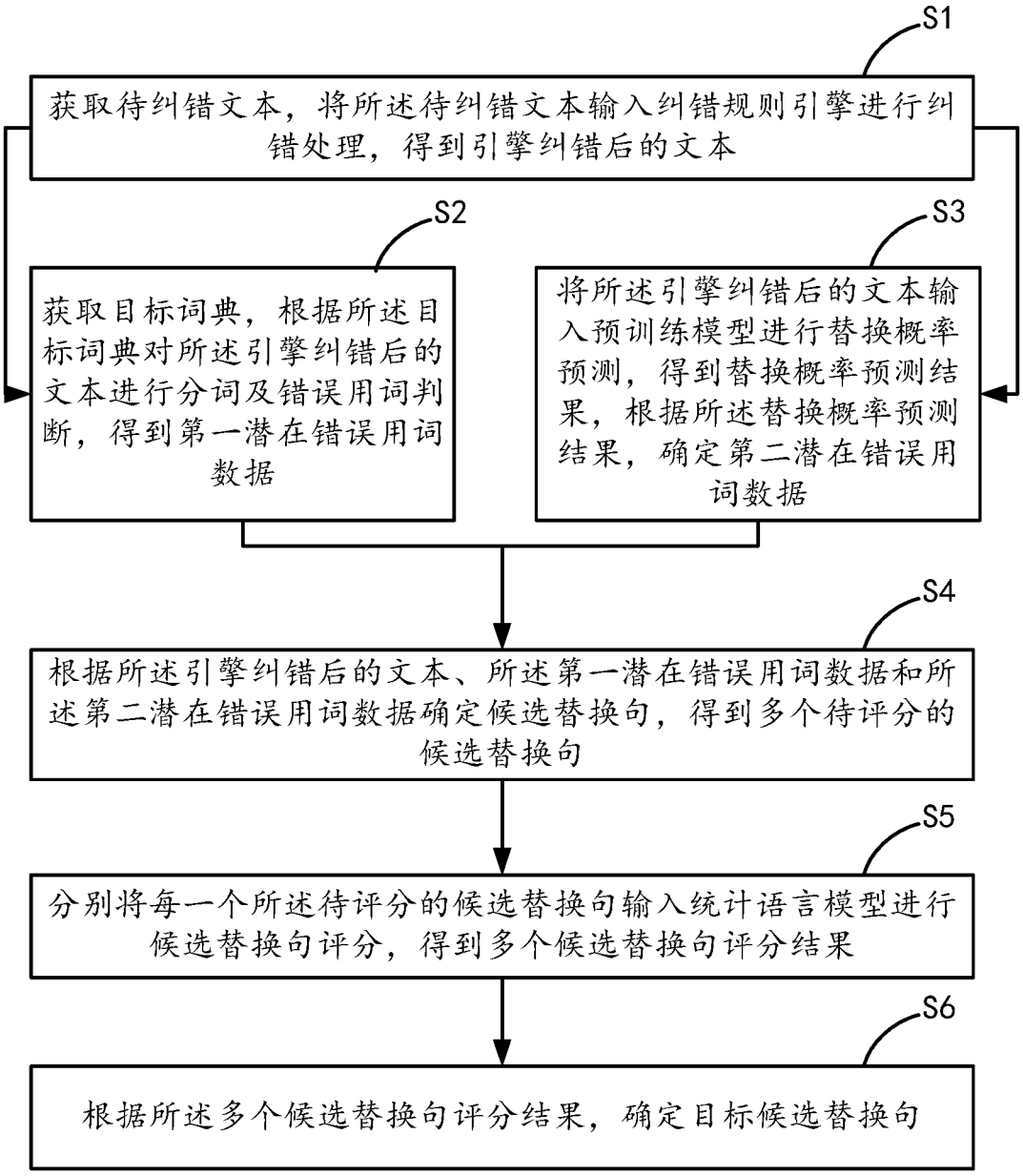

S1
获取待纠错文本，将所述待纠错文本输入纠错规则引擎进行纠错处理，得到引擎纠错后的文本
S2
获取目标词典，根据所述目标词典对所述引擎纠错后的文本进行分词及错误用词判断，得到第一潜在错误用词数据
S3
将所述引擎纠错后的文本输入预训练模型进行替换概率预测，得到替换概率预测结果，根据所述替换概率预测结果，确定第二潜在错误用词数据
S4
根据所述引擎纠错后的文本、所述第一潜在错误用词数据和所述第二潜在错误用词数据确定候选替换句，得到多个待评分的候选替换句
S5
分别将每一个所述待评分的候选替换句输入统计语言模型进行候选替换句评分，得到多个候选替换句评分结果
S6
根据所述多个候选替换句评分结果，确定目标候选替换句

图 1

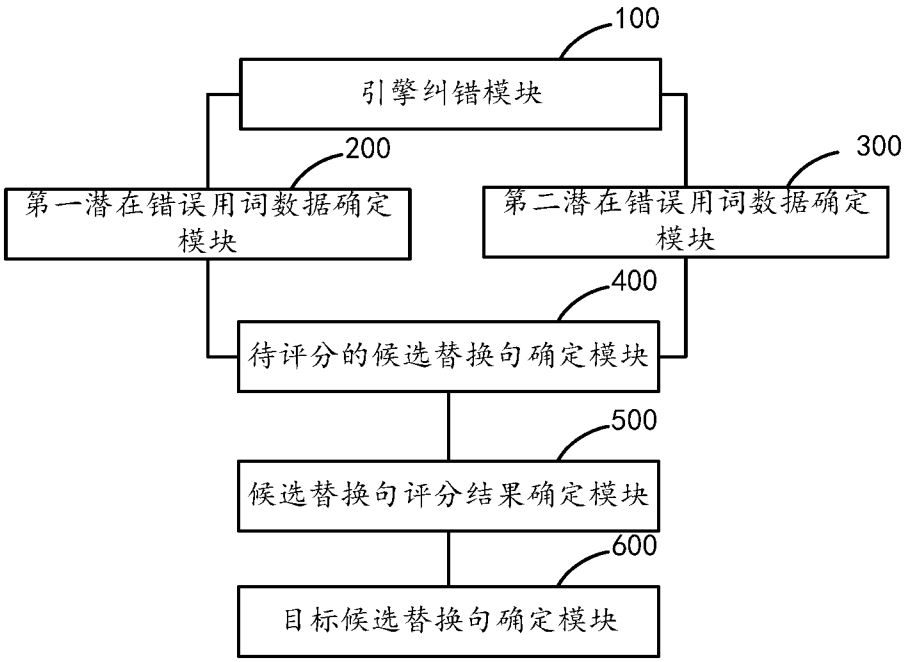

100
引擎纠错模块
200
第一潜在错误用词数据确定模块
300
第二潜在错误用词数据确定模块
400
待评分的候选替换句确定模块
500
候选替换句评分结果确定模块
600
目标候选替换句确定模块

图 2

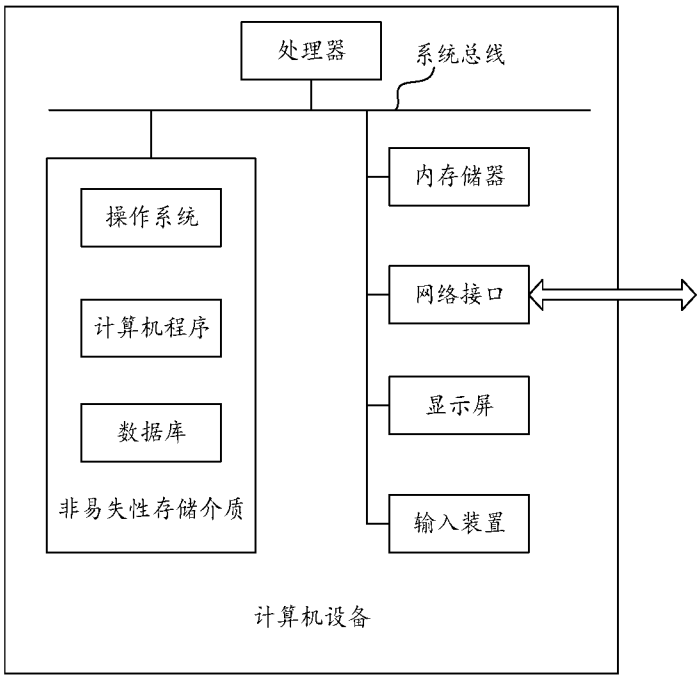

处理器
系统总线
操作系统
计算机程序
数据库
非易失性存储介质
内存储器
网络接口
显示屏
输入装置
计算机设备

图 3

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2021/084546

A. CLASSIFICATION OF SUBJECT MATTER

G06F 40/232(2020.01)i; G06F 40/242(2020.01)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)

G06F

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)

CNPAT, WPI, EPODOC, IEEE, CNKI: 纠错, 检错, 错误, 潜在, 概率, 替换, 候选, 评分, 得分, 排序, 同音字, 业务, 频率, 词频, 阈值, correct, error, mistake, probability, candidate, score, sort, frequency, threshold

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 112380840 A (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) 19 February 2021 (2021-02-19) claims 1-10	1-20
X	CN 111382260 A (TENCENT MUSIC ENTERTAINMENT TECHNOLOGY (SHENZHEN) CO., LTD.) 07 July 2020 (2020-07-07) description paragraphs 0026-0254	1-20
X	CN 111797614 A (ALIBABA GROUP HOLDING LIMITED) 20 October 2020 (2020-10-20) description paragraphs 0119-0315	1-20
A	CN 111753529 A (HANGZHOU YUNJIA CLOUD CALCULATING CO., LTD.) 09 October 2020 (2020-10-09) entire document	1-20
A	CN 111460793 A (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) 28 July 2020 (2020-07-28) entire document	1-20

☐ 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
"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	Date of mailing of the international search report
19 July 2021	**27 July 2021**
Name and mailing address of the ISA/CN **China National Intellectual Property Administration (ISA/CN)** **No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing 100088, China** Facsimile No. **(86-10)62019451**	Authorized officer Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2021/084546

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 112380840 A	19 February 2021	None	
CN 111382260 A	07 July 2020	None	
CN 111797614 A	20 October 2020	None	
CN 111753529 A	09 October 2020	None	
CN 111460793 A	28 July 2020	None	

国际检索报告

国际申请号
PCT/CN2021/084546

A. 主题的分类

G06F 40/232(2020.01)i; G06F 40/242(2020.01)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)

G06F

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））

CNPAT, WPI, EPODOC, IEEE, CNKI：纠错，检错，错误，潜在，概率，替换，候选，评分，得分，排序，同音字，业务，频率，词频，阈值，correct, error, mistake, probability, candidate, score, sort, frequency, threshold

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
PX	CN 112380840 A (平安科技深圳有限公司) 2021年 2月 19日 (2021-02-19) 权利要求1-10	1-20
X	CN 111382260 A (腾讯音乐娱乐科技深圳有限公司) 2020年 7月 7日 (2020-07-07) 说明书第0026-0254段	1-20
X	CN 111797614 A (阿里巴巴集团控股有限公司) 2020年 10月 20日 (2020-10-20) 说明书第0119-0315段	1-20
A	CN 111753529 A (杭州云嘉云计算有限公司) 2020年 10月 9日 (2020-10-09) 全文	1-20
A	CN 111460793 A (平安科技深圳有限公司) 2020年 7月 28日 (2020-07-28) 全文	1-20

☐ 其余文件在C栏的续页中列出。

☑ 见同族专利附件。

* 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

“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	国际检索报告邮寄日期
2021年 7月 19日	2021年 7月 27日
ISA/CN的名称和邮寄地址 中国国家知识产权局(ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10)62019451	受权官员 王芳 电话号码 86-(10)-53961369

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2021/084546

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 112380840 A	2021年 2月 19日	无	
CN 111382260 A	2020年 7月 7日	无	
CN 111797614 A	2020年 10月 20日	无	
CN 111753529 A	2020年 10月 9日	无	
CN 111460793 A	2020年 7月 28日	无	