

(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
2023 年 2 月 2 日 (02.02.2023)



(10) 国际公布号
WO 2023/005293 A1

(51) 国际专利分类号:
G06F 40/00 (2020.01)

(21) 国际申请号: PCT/CN2022/088892

(22) 国际申请日: 2022 年 4 月 25 日 (25.04.2022)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202110873540.0 2021年7月30日 (30.07.2021) 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 518033 (CN)。

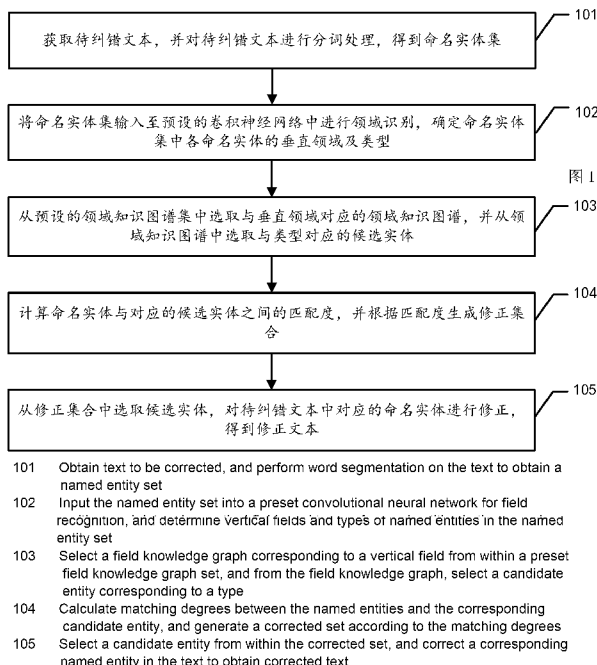
(72) 发明人: 李帅(LI, Shuai); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518033 (CN)。

(74) 代理人: 北京市京大律师事务所(BEIJING JINGDA LAW FIRM); 中国北京市海淀区高粱桥斜街 59 号院 1 号楼 8 层 801-3, Beijing 100044 (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, 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, JM, 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,

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

(54) 发明名称: 文本纠错方法、装置、设备及存储介质



(57) **Abstract:** The present application relates to the field of data analysis, and discloses a text error correction method, apparatus, and device, and a storage medium. The method comprises: performing word segmentation on text to be corrected to obtain a named entity set; inputting the named entity set into a preset convolutional neural network for field recognition, and determining vertical fields and types of named entities in the named entity set; selecting a field knowledge graph and a candidate entity from within a field knowledge graph set; calculating matching degrees between the named entities and the candidate entity, and generating a corrected set according to

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))。

the matching degrees; and selecting a candidate entity from the corrected set, and correcting the text to obtain corrected text. According to the present application, error correction efficiency and accuracy can be increased by means of calling a field knowledge graph, selecting a candidate entity, and performing targeted correction on an error present in text to be corrected. In addition, the present application also relates to blockchain technology, and text to be corrected and corrected text can be stored in blockchains.

(57) 摘要: 本申请涉及数据分析领域, 公开了一种文本纠错方法、装置、设备及存储介质, 该方法包括: 对待纠错文本进行分词处理, 得到命名实体集; 将命名实体集中输入至预设的卷积神经网络中进行领域识别, 确定命名实体集中各命名实体的垂直领域及类型; 从领域知识图谱集中选取领域知识图谱和候选实体; 计算命名实体与候选实体的匹配度, 并根据匹配度生成修正集合; 从修正集合中选取候选实体, 对待纠错文本进行修正, 得到修正文本。本申请通过调用领域知识图谱, 选取候选实体, 对待纠错文本中出现的错误进行针对性的修正, 从而提高了纠错效率和准确度。此外, 本申请还涉及区块链技术, 待纠错文本和修正文本可存储于区块链中。

文本纠错方法、装置、设备及存储介质

本申请要求于 2021 年 7 月 30 日提交中国专利局、申请号为 202110873540.0、发明名称为“文本纠错方法、装置、设备及存储介质”的中国专利申请的优先权，其全部内容通过引用结合在申请中。

技术领域

本申请涉及数据分析领域，尤其涉及一种文本纠错方法、装置、设备及存储介质。

背景技术

人工智能正在以一种前所未有的力量影响着各行各业，而智能客服作为此次技术革新的排头兵已经在垂直领域生根发芽，即将开花。但是用户在文字交互中经常出现输入错误，包括错字、缺字、多字、字序错误、拼音汉字混合等各种各样的错误。其中有些错误无关紧要，不影响系统后续处理，有些错误会对系统的后续自动处理有非常大的影响，差之毫厘谬以千里。

目前，对文本进行纠错主要是基于语言模型的混淆度和字形字音的相似度去选取最大概率的替换字。但是发明人意识到现有技术只能处理错字，不能处理多字、漏字等其他的情况，需要其他多种技术去配合，不能从整体上去系统解决，从而导致纠错效率低、准确率低。

发明内容

本申请的主要目的在于解决现有技术中文本纠错效率低、准确率低的技术问题。

本申请第一方面提供了一种文本纠错方法，所述文本纠错方法包括：获取待纠错文本，并对所述待纠错文本进行分词处理，得到命名实体集；将所述命名实体集输入至预设的卷积神经网络中进行领域识别，确定所述命名实体集中各命名实体的垂直领域及类型；从预设的领域知识图谱集中选取与所述垂直领域对应的领域知识图谱，并从所述领域知识图谱中选取与所述类型对应的候选实体；计算所述命名实体与对应的候选实体之间的匹配度，并根据所述匹配度生成修正集合；从所述修正集合中选取候选实体，对所述待纠错文本中对应的命名实体进行修正，得到修正文本。

本申请第二方面提出一种文本纠错设备，包括存储器、处理器及存储在所述存储器上并可在所述处理器上运行的计算机可读指令，所述处理器执行所述计算机可读指令时实现如下步骤：获取待纠错文本，并对所述待纠错文本进行分词处理，得到命名实体集；将所述命名实体集输入至预设的卷积神经网络中进行领域识别，确定所述命名实体集中各命名实体的垂直领域及类型；从预设的领域知识图谱集中选取与所述垂直领域对应的领域知识图谱，并从所述领域知识图谱中选取与所述类型对应的候选实体；计算所述命名实体与对应的候选实体之间的匹配度，并根据所述匹配度生成修正集合；从所述修正集合中选取候选实体，对所述待纠错文本中对应的命名实体进行修正，得到修正文本。

本申请第三方面提出一种计算机可读存储介质，所述计算机可读存储介质中存储计算机指令，当所述计算机指令在计算机上运行时，使得计算机执行如下步骤：获取待纠错文本，并对所述待纠错文本进行分词处理，得到命名实体集；将所述命名实体集输入至预设的卷积神经网络中进行领域识别，确定所述命名实体集中各命名实体的垂直领域及类型；从预设的领域知识图谱集中选取与所述垂直领域对应的领域知识图谱，并从所述领域知识图谱中选取与所述类型对应的候选实体；计算所述命名实体与对应的候选实体之间的匹配度，并根据所述匹配度生成修正集合；从所述修正集合中选取候选实体，对所述待纠错文本中对应的命名实体进行修正，得到修正文本。

本申请第四方面提出一种文本纠错装置，所述文本纠错装置包括：分词模块，用于获取待纠错文本，并对所述待纠错文本进行分词处理，得到命名实体；识别模块，用于将所

述命名实体集输入至预设的卷积神经网络中进行领域识别，确定所述命名实体集中各命名实体的垂直领域及类型；选取模块，用于从预设的领域知识图谱集中选取与所述垂直领域对应的领域知识图谱，并从所述领域知识图谱中选取与所述类型对应的候选实体；计算模块，用于计算所述命名实体与对应的候选实体之间的匹配度，并根据所述匹配度生成修正集合；修正模块，用于从所述修正集合中选取候选实体，对所述待纠错文本中对应的命名实体进行修正，得到修正文本。

在本申请提供的技术方案中，通过对待纠错文本进行分词处理，得到命名实体集；将命名实体集输入至预设的卷积神经网络中进行领域识别，确定命名实体集中各命名实体的垂直领域及类型；从预设的领域知识图谱集中选取与垂直领域对应的领域知识图谱，并从领域知识图谱中选取与命名实体的类型对应的候选实体；计算命名实体与候选实体的匹配度，并根据匹配度生成修正集合；从修正集合中选取候选实体，对待纠错文本中对应的命名实体进行修正，得到修正文本。本申请提供的技术方案通过调用领域知识图谱，选取候选实体，对待纠错文本中出现的错误进行针对性的修正，从而提高了纠错效率和准确度。

附图说明

图1为本申请实施例中文本纠错方法的第一个实施例示意图；
图2为本申请实施例中文本纠错方法的第二个实施例示意图；
图3为本申请实施例中文本纠错方法的第三个实施例示意图；
图4为本申请实施例中文本纠错方法的第四个实施例示意图；
图5为本申请实施例中文本纠错装置的一个实施例示意图；
图6为本申请实施例中文本纠错装置的另一个实施例示意图；
图7为本申请实施例中文本纠错设备的一个实施例示意图。

具体实施方式

本申请实施例提供了一种文本纠错方法、装置、设备及存储介质，通过对待纠错文本进行分词处理，得到命名实体集；将命名实体集输入至预设的卷积神经网络中进行领域识别，确定命名实体集中各命名实体的垂直领域及类型；从预设的领域知识图谱集中选取与垂直领域对应的领域知识图谱，并从领域知识图谱中选取与命名实体的类型对应的候选实体；计算命名实体与候选实体的匹配度，并根据匹配度生成修正集合；从修正集合中选取候选实体，对待纠错文本中对应的命名实体进行修正，得到修正文本。本申请实施例通过调用领域知识图谱，选取候选实体，对待纠错文本中出现的错误进行针对性的修正，从而提高了纠错效率和准确度。

本申请的说明书和权利要求书及上述附图中的术语“第一”、“第二”、“第三”、“第四”等（如果存在）是用于区别类似的对象，而不必用于描述特定的顺序或先后次序。应该理解这样使用的数据在适当情况下可以互换，以便这里描述的实施例能够以除了在这里图示或描述的内容以外的顺序实施。此外，术语“包括”或“具有”及其任何变形，意图在于覆盖不排他的包含，例如，包含了一系列步骤或单元的过程、方法、系统、产品或设备不必限于清楚地列出的那些步骤或单元，而是可包括没有清楚地列出的或对于这些过程、方法、产品或设备固有的其它步骤或单元。

为便于理解，下面对本申请实施例的具体内容进行描述，请参阅图1，本申请实施例中文本纠错方法的第一个实施例包括：

101，获取待纠错文本，并对待纠错文本进行分词处理，得到命名实体集；

服务器获取待纠错文本，并对该待纠错文本进行分词处理，其中，分词处理需要结合预设的分词词典。分词词典是指包括有常用的或固定的词语的数据库，其是分词的基准，通过比照分词词典以使输入的待纠错文本中的语句转化为具有最大字符长度的独立词语，即最大字符长度的独立词语为命名实体，并汇集各命名实体形成命名实体集。在本实施例中，命名实体就是人名、机构名、地名以及其他所有以名称为标识的实体。其中，更广泛

的实体还包括数字、日期、货币、地址等等。

在本实施例中，分词是指将待纠错文本中的字符串划分为词串的过程。其中，分词方法可以为正向最大匹配法、逆向最大匹配法、条件随机场模型或隐马尔可夫模型。正向最大匹配法的特点是分词效率高，具有线性时间复杂度，容易实现，不需要指定词语的最大长度；逆向最大匹配法的特点是具有线性时间复杂度，需要指定词语的最大长度（maxLen）；隐马尔可夫模型的特点是对未登录词的识别效果优于最大匹配法，但整体效果依赖于训练语料；条件随机场模型的特点是不仅考虑了词语出现的频率，还考虑上下文，具备较好的学习能力，因此其对歧义词和未登录词的识别都具有良好的效果。

进一步的，在本实施例中，在调用正向最大匹配法对待纠错文本中的语句进行顺向扫描，当存在交集型歧义时很可能产生分词错误。因此，本实施例通过增加回溯机制来校正正向最大匹配法的分词结果。其中，回溯是指在分词过程中，采用后退的策略以修正当前分词结果的试探方法。通过增加回溯机制能够提高分词准确率，有效改善交集型歧义问题。

102，将命名实体集输入至预设的卷积神经网络中进行领域识别，确定命名实体集中各命名实体的垂直领域及类型；

将命名实体集输入至预设的卷积神经网络（CNN 网络）中进行领域识别，在本实施例中，CNN 网络的网络结构包含输入层、网络层和输出层；其中，输入层是将待纠错文本的命名实体集中的各命名实体输入到网络层，输出层是网络层的输出，其通过逻辑回归函数（softmax 函数）计算命名实体所涉及的各个专业领域的概率，根据该概率确定命名实体的垂直领域（专业领域）；网络层又包含卷积层、池化层、特征连接层、全连接层四个部分；卷积层设计了两个通道，第一个通道的卷积窗大小为 1，第二个通道的卷积窗大小为 2，使 CNN 网络提取出待纠错文本中单个词汇和相邻词汇的特征；池化层采用最大池化获取卷积层输出的每个通道特征最明显的特征；特征连接层是将池化层输出的两个通道的特征拼接到一起得到特征矩阵；全连接层最后将特征连接层输出的特征矩阵进行分类，得到命名实体的类型，根据命名实体的类型将命名实体存储至 {k, v} 集合中，其中，k 表示命名实体，v 表示命名实体的类型。

103，从预设的领域知识图谱集中选取与垂直领域对应的领域知识图谱，并从领域知识图谱中选取与类型对应的候选实体；

从预设的领域知识图谱集中选取与命名实体的垂直领域对应的领域知识图谱，并从领域知识图谱中选取与命名实体的类型一致的候选实体。其中，领域知识图谱集包含多个领域知识图谱，候选实体是领域知识图谱中的命名实体。

在本实施例中，对于通用领域知识图谱，是显示知识发展进程与结构关系的一系列各种不同的图形，用可视化技术描述知识资源及其载体，挖掘、分析、构建、绘制和显示知识及它们之间的相互联系。知识图谱，是通过将应用数学、图形学、信息可视化技术、信息科学等学科的理论与方法与计量学引文分析、共现分析等方法结合，并利用可视化的图谱形象地展示学科的核心结构、发展历史、前沿领域以及整体知识架构达到多学科融合目的的现代理论。其中，特定领域的知识图谱，对知识的精确性要求较高，包括定义数据的概念、类别、关联、属性约束等。

104，计算命名实体与对应的候选实体之间的匹配度，并根据匹配度生成修正集合；

服务器计算命名实体与对应的候选实体之间的匹配度，根据匹配度生成修正集合。在本实施例中，根据与垂直领域对应的领域知识图谱（G），对集合 {k, v}，依次和 G 中类型 v 的候选实体（g）做比较，如果命名实体 k 和候选实体 g 完全匹配则说明待纠错文本并不需要做纠正，即命名实体和候选实体的匹配度越高，该命名实体的修正率越低。如果候选实体 g 与命名实体 k 并不完全匹配，则提取并汇集与命名实体 k 匹配度最大的候选实体 g，形成修正集合 C_k。其中，修正集合只包括候选实体 g。

105，从修正集合中选取候选实体，对待纠错文本中对应的命名实体进行修正，得到

修正文本。

提取修正集合，判断该修正集合中是否包含多个候选实体。当修正集合中只包含一个候选实体，则说明该候选实体就是修正后的命名实体，即根据该候选实体，对命名实体进行修正。当修正集合中包含有多个候选实体时，服务器根据预设的领域语言模型，分别计算修正集合中各候选实体在待纠错文本中的出现概率。

当得到各候选实体对应的出现概率后，对各出现概率进行数值大小的比较，并对各候选实体按照对应的出现概率的比较结果进行排序，生成出现序列。根据排序的结果，从出现序列中选取出现概率最大的候选实体对待纠错文本的命名实体进行修正，从而得到修正文本。

在本实施例中，将需要进行修正的命名实体作为混淆词，并汇集成为混淆词典，调用混淆词典，对分词处理后的待纠错文本中的每个词语进行遍历，提高纠错效率和准确率。

在本申请实施例中，对待纠错文本进行分词处理得到命名实体，并从待纠错文本的垂直领域对应的领域知识图谱中选取与命名实体类型一致的候选实体，计算命名实体与候选实体的匹配度从而生成修正集合，根据该修正集合对待纠错文本进行修正。本实施例通过调用领域知识图谱，选取候选实体，对待纠错文本中出现的错误进行针对性的修正，从而提高了纠错效率和准确度。

请参阅图 2，本申请实施例中文本纠错方法的第二个实施例包括：

201，获取待纠错文本，并根据预设的词典，将待纠错文本生成前缀树；

202，对前缀树进行词图扫描，生成有向无环图；

服务器获取待纠错文本，并调用中文分词工具（jieba）的词典作为对待纠错文本进行分词处理时所用的词典，把一些并不常用的词汇删除，尽可能保留正确且常用的词汇，以减小分词器的容量。调用该词典，将待纠错文本生成前缀树（trie 树）。对 trie 树结构进行词图扫描，即把词典中词语放到一个 trie 树中，一个词语的前面几个字一样，就表示他们具有相同的前缀，就可以使用 trie 树来存储，提高查找速度。在本实施例中，将待纠错文本（由一个或多个语句组成）中的句子根据预设的词典进行词图扫描处理，生成有向无环图。前缀树的生成及词图扫描都采用的是现有技术，故在此不做细述。

203，调用预设的动态规划查找最大概率路径算法，从有向无环图中查找基于词频的最大切分组合；

查找待纠错文本中已经切分好的词语，并计算该词语出现的频率，如果没有该词，就把词典中出现频率最小的那个词语的频率作为该词的频率；然后根据动态规划查找最大概率路径的算法，对待纠错文本中的句子从右往左反向计算最大概率，因为通常情况下形容词太多，后面的才是主干。因此，从右往左计算，正确率要高于从左往右计算，这里类似于逆向最大匹配， $P(\text{NodeN}) = 1.0$ ， $P(\text{NodeN}-1) = P(\text{NodeN}) * \text{Max}(P(\text{倒数第一个词})) \cdots$ 依次类推，最后结合有向无环图得到最大概率路径，即得到最大概率的切分组合。

204，将待纠错文本按照最大切分组合进行分词处理，得到词序列；

205，将词序列输入预设的基于词序列的命名实体识别模型，输出得到命名实体集；

将待纠错文本按照所得到的最大切分组合进行分词处理，即按照字符组合进行分词，得到词序列。将词序列输入至服务器预设的基于词序列的命名实体识别模型，输出识别结果，即识别出待纠错文本中的命名实体，并汇集成为命名实体集。本实施例中采用的是基于词序列的命名实体识别模型，模型输入为词序列，而不是字序列，因此可以提升识别效率，同时也可以减小占用内存。

在本实施例中，命名实体识别模型（Named Entity Recognition，简称 NER）是信息提取、问答系统、句法分析、机器翻译等应用领域的重要基础工具，在自然语言处理技术走向实用化的过程中占有重要地位。一般来说，命名实体识别模型的任务就是识别出待处

理文本中三大类（实体类、时间类和数字类）、七小类（人名、机构名、地名、时间、日期、货币和百分比）命名实体。

206, 将命名实体集输入至预设的卷积神经网络中, 并调用卷积神经网络的逻辑回归函数, 计算命名实体集中各命名实体所涉及的各个领域的领域属性值;

5 207, 对各个领域属性值进行比较, 将领域属性值最大的领域作为命名实体集中各命名实体的垂直领域;

10 将命名实体集输入至预设的卷积神经网络 (CNN 模型) 中, 调用卷积神经网络的逻辑回归函数 (softmax) 函数, 计算命名实体集中各命名实体所涉及的各个领域的领域属性值。对各个领域属性值进行比较, 并从中选取领域属性值最大的领域作为该命名实体的垂直领域。

在本实施例中, 卷积神经网络中的全连接层有两个隐藏层, 而全连接层的输出层节点数与预设的命名实体的类型个数保持一致; CNN 模型的输出层采用 softmax 函数来进行每个领域的概率的计算, 即计算领域属性值。

208, 基于卷积神经网络中的卷积层, 提取命名实体集中各命名实体的类型特征信息;

15 209, 计算类型特征信息与预设的类型之间的匹配度, 根据匹配度确定命名实体的类型;

根据卷积神经网络中的卷积层, 提取命名实体集中各命名实体的类型特征信息, 并计算类型特征信息与预设的命名实体的类型之间的匹配度, 即计算类型特征信息与预设类型之间的相似度, 根据该相似度确定命名实体的类型。

20 在本实施例中, CNN 模型输入层输入命名实体, 维度为 8×271 ; 卷积层有两个通道, 两个通道的卷积窗维度分别为 1×271 、 2×271 , 每个通道都有 512 个卷积核。卷积层的输出分别为 8×512 、 7×512 的矩阵。池化层对卷积层的输出进行最大池化操作, 输出 1×512 和 1×512 的两个类型特征信息, 且该类型特征信息以特征向量的形式输出, 因此, 待纠错文本可得到 1028 种类型特征信息。特征连接层将池化层的两个输出拼接到一起形成一个 1×1028 的类型特征信息, 将类型特征信息输入到 CNN 模型的全连接层中, 输出命名实体的类型。

210, 从预设的领域知识图谱集中选取与垂直领域对应的领域知识图谱, 并从领域知识图谱中选取与类型对应的候选实体;

211, 计算命名实体与对应的候选实体之间的匹配度, 并根据匹配度生成修正集合;

30 212, 从修正集合中选取候选实体, 对待纠错文本中对应的命名实体进行修正, 得到修正文本。

在本实施例中, 步骤 210-212 与上述的文本纠错方法的第一个实施例中的步骤 103-105 一致, 在此不再赘述。

35 在本申请实施例中, 根据预设的词典将待纠错文本生成有向无环图, 并调用预设的查找最大概率路径算法, 从有向无环图中查找基于词频的最大切分组合, 对该最大切分组合进行分词处理得到词序列并将其输入至基于词序列的命名实体识别模型中确定命名实体, 对待纠错文本进行一系列处理生成词序列, 再输入至命名实体识别模型中基于词序列进行命名实体的识别, 从而可以根据命名实体准确、快速的定位错误, 提高了纠错效率。

40 请参阅图 3, 本申请实施例中文本纠错方法的第三个实施例包括:

301, 获取待纠错文本, 并对待纠错文本进行分词处理, 得到命名实体集;

302, 将命名实体集输入至预设的卷积神经网络中进行领域识别, 确定命名实体集中各命名实体的垂直领域及类型;

45 303, 从预设的领域知识图谱集中选取与垂直领域对应的领域知识图谱, 并从领域知识图谱中选取与类型对应的候选实体;

304, 计算命名实体与对应的候选实体之间的字形相似度, 若字形相似度大于预设的字形相似阈值, 则汇集候选实体生成修正集合;

分别对命名实体与对应的候选实体的字形进行分析, 并计算命名实体与候选实体之间的字形相似度, 将字形相似度与预设的字形相似阈值进行比较, 当字形相似度大于字形相似阈值时, 则提取并汇集对应的候选实体, 生成修正集合。

在本实施例中, 每个命名实体均是由一个或多个字组成的, 将该字称为目标字, 并确定目标字的字图像和包含字形特征的字形向量; 其中, 可以基于卷积神经网络确定字图像的字形向量。并将目标字根据该目标字对应的多种字体的写法生成该目标字的字图像。具体的, 确定目标字在不同字体下与目标字相对应的字体图像, 对目标字所有的字体图像进行拼接处理, 生成深度为 D 的字图像, D 为目标字的字体图像的个数。利用多种字体的字体图像来生成包含字形特征的字形向量, 使得目标字的字形向量包含多种字体的字形特征。同时, 按照上述步骤确定候选实体的字形特征, 并根据命名实体和候选实体的字形特征计算字形相似度, 从而将字形相似度与字形相似阈值进行比较。另外, 本实施例中的“字体”还可以包含不同历史时期的字体, 比如金文、草书、魏碑等, 只要是该字体可以包含字形特征即可。

305, 或, 计算命名实体与对应的候选实体之间的字音相似度, 若字音相似度大于预设的字音相似阈值, 则汇集候选实体生成修正集合;

对命名实体进行字音分析, 并判断该命名实体是否为拼音汉字混合拼写。若命名实体为拼音汉字混合拼写, 则对命名实体进行字音转换, 规范化为拼音的字符串。具体的, 服务器根据预设的拼音转化算法, 对命名实体中的汉字进行字音转换, 生成拼音, 将命名实体中的其他拼音进行拼接, 生成拼音串。

进一步的, 分析命名实体和对应的候选实体的字音, 并计算命名实体与候选实体之间的字音相似度, 将字音相似度与预设的字音相似阈值进行比较, 当字音相似度大于字形相似阈值时, 则提取并汇集对应的候选实体, 生成修正集合。

在本实施例中, 制定音形码映射规则, 通过将汉字的字音按照简单的替代规则部分映射到一个字符位, 分为 10 部分。字音主要覆盖韵母, 声母, 补码以及声调的内容, 覆盖 4 个字符位, 第一位韵母位, 从韵母“a”到“ong”24 种韵母, 由数字“1-9”和字母“A-K”代替, 第二位是声母位, 同样的, 也是利用数字“1-9”和字母“A-J”代替, 其中“Z”和“ZH”为相同转化; 第四位是声调位, 分别用“1-4”来替代汉字中的四个声调。根据音形码映射规则分别对命名实体和候选实体进行编码, 并将编码后的结果进行相似度的比较, 其中, 比较两个编码之间的相似度采用距离算法, 得到命名实体和候选实体的字音相似度。

306, 或, 分析命名实体与对应的候选实体的字词结构, 并基于字词结构确定命名实体与候选实体之间的相似度, 若相似度大于预设的字词结构相似阈值, 则汇集候选实体生成修正集合;

分析命名实体和对应的候选实体的字词结构, 并计算命名实体和候选实体之间字词结构的相似度, 其中, 字词结构包括字词组合和字序, 在本实施例中, 当命名实体和候选实体的字词组合的相似度大于预设的字词组合相似阈值时, 则提取并汇集对应的候选实体, 生成修正集合; 或者, 当命名实体和候选实体的字序的相似度大于预设的字序相似阈值时, 则提取并汇集对应的候选实体, 生成修正集合。

分别对命名实体与候选实体的字词组合进行分析, 确定候选实体是否是由命名实体增加一个字或者减少一个字所构成, 即确定该命名实体是否缺字或多字。计算命名实体和候选实体的字词组合的相似度, 并将字词组合的相似度与预设的字词组合相似阈值进行比较, 当字词组合的相似度大于字词组合相似阈值时, 则提取并汇集对应的候选实体, 生成修正集合。

具体的，服务器根据预设的编码规则，分别对命名实体和候选实体中的字或字符进行编码，即预设编码标识符对字或字符进行标识，从而将字或字符转换为对应的编码标识符，生成两个编码标识串，并将编码的结果进行相似度的比较，即比较两个编码标识串之间的编码标识符组成结构，在两个编码标识串编码标识符在排列顺序一致下，判断其个数是否一致，其中，比较两个编码标识串之间的相似度采用距离算法，得到命名实体和候选实体的字词组合的相似度。

另外，分别对命名实体和候选实体的字序进行分析，确定候选实体是否是由命名实体调整字序所构成，计算命名实体和候选实体的字序的相似度，并将字序的相似度与预设的字序相似阈值进行比较，当字序的相似度大于字序相似阈值时，则提取并汇集对应的候选实体，生成修正集合。

具体的，服务器根据预设的编码规则，分别对命名实体和候选实体中的字或字符进行编码，即预设编码标识符对字或字符进行标识，从而将字或字符转换为对应的编码标识符，生成两个编码标识串，并将编码的结果进行相似度的比较，即比较两个编码标识串之间的编码标识符组成结构，在两个编码标识串编码标识符的个数一致且编码标识符一致下，判断其排列顺序是否一致，其中，比较两个编码标识串之间的相似度采用距离算法，得到命名实体和候选实体的字序的相似度。

307，从修正集合中选取候选实体，对待纠错文本中对应的命名实体进行修正，得到修正文本。

提取修正集合，判断该修正集合中是否包含多个候选实体。当修正集合中只包含一个候选实体，则说明该候选实体就是修正后的命名实体，即根据该候选实体，对待纠错文本中对应的命名实体进行修正。当修正集合中包含有多个候选实体时，服务器根据预设的领域语言模型，分别计算修正集合中各候选实体在待纠错文本中的出现概率。

当得到各候选实体对应的出现概率后，对各出现概率进行数值大小的比较，并对各候选实体按照对应的出现概率的比较结果进行排序，生成出现序列。根据排序的结果，从出现序列中选取出现概率最大的候选实体对待纠错文本的命名实体进行修正，从而得到修正文本。

在本实施例中，将需要进行修正的命名实体作为混淆词，并汇集成为混淆词典，调用混淆词典，对分词处理后的待纠错文本中的每个词语进行遍历，提高纠错效率和准确率。

在本申请实施例中，步骤 301-303 与上述的文本纠错方法的第一个实施例中的步骤 101-103 一致，在此不做赘述。

在本申请实施例中，对命名实体和候选实体分别进行字形分析、字音分析和字词结构分析，从而可以识别文本的多种错误，对待纠错文本有针对性地进行修正，提高了文本纠错的准确度。

请参阅图 4，本申请实施例中文本纠错方法的第四个实施例包括：

401，获取待纠错文本，并对待纠错文本进行分词处理，得到命名实体集；

402，将命名实体集输入至预设的卷积神经网络中进行领域识别，确定命名实体集中各命名实体的垂直领域及类型；

403，从预设的领域知识图谱集中选取与垂直领域对应的领域知识图谱，并从领域知识图谱中选取与类型对应的候选实体；

404，计算命名实体与对应的候选实体的匹配度，并根据匹配度生成修正集合；

405，判断修正集合是否包含多个候选实体；

406，若修正集合包含多个候选实体，则根据预设的领域语言模型，计算候选实体在待纠错文本中的出现概率；

提取修正集合，判断该修正集合中是否包含多个候选实体。当修正集合中只包含一个

候选实体，则说明该候选实体就是修正后的命名实体，即根据该候选实体，对命名实体进行修正。当修正集合中包含有多个候选实体时，服务器根据预设的领域语言模型，分别计算修正集合中各候选实体在待纠错文本中的出现概率。

407，对候选实体按照出现概率的大小进行排序，得到出现序列；

5 408，根据出现序列从修正集合中选取候选实体，对待纠错文本中对应的命名实体进行修正，得到修正文本。

当得到各候选实体对应的出现概率后，对各出现概率进行数值大小的比较，并对各候选实体按照对应的出现概率的比较结果进行排序，生成出现序列。根据排序的结果，从出现序列中选取出现概率最大的候选实体对待纠错文本的命名实体进行修正，从而得到修正文本。

10 在本申请实施例中，步骤 401-404 与上述的文本纠错方法的第一个实施例中的步骤 101-104 一致，在此不做赘述。

15 在本申请实施例中，计算候选实体在待纠错文本中的出现概率，并根据出现概率选取出现概率最大的候选实体对待纠错文本进行修正，提高了对待纠错文本中命名实体修正的准确度。

上面对本申请实施例中的文本纠错方法进行了描述，下面对本申请实施例中的文本纠错装置进行描述，请参照图 5，本申请实施例中的文本纠错装置的一个实施例包括：

20 分词模块 501，用于获取待纠错文本，并对所述待纠错文本进行分词处理，得到命名实体集；

识别模块 502，用于将所述命名实体集输入至预设的卷积神经网络中进行领域识别，确定所述命名实体集中各命名实体的垂直领域及类型；

选取模块 503，用于从预设的领域知识图谱集中选取与所述垂直领域对应的领域知识图谱，并从所述领域知识图谱中选取与所述类型对应的候选实体；

25 计算模块 504，用于计算所述命名实体与对应的候选实体之间的匹配度，并根据所述匹配度生成修正集合；

修正模块 505，用于从所述修正集合中选取候选实体，对所述待纠错文本中对应的命名实体进行修正，得到修正文本。

30 在本申请实施例中，通过文本纠错装置对待纠错文本进行分词处理得到命名实体，并从待纠错文本的垂直领域对应的领域知识图谱中选取与命名实体类型一致的候选实体，计算命名实体与候选实体的匹配度从而生成修正集合，根据该修正集合对待纠错文本进行修正。本提案通过调用领域知识图谱，选取候选实体，对待纠错文本中出现的错误进行针对性的修正，从而提高了纠错效率和准确度。

35 请参阅图 6，本申请实施例中的文本纠错装置的另一个实施例包括：

分词模块 501，用于获取待纠错文本，并对所述待纠错文本进行分词处理，得到命名实体集；

识别模块 502，用于将所述命名实体集输入至预设的卷积神经网络中进行领域识别，确定所述命名实体集中各命名实体的垂直领域及类型；

40 选取模块 503，用于从预设的领域知识图谱集中选取与所述垂直领域对应的领域知识图谱，并从所述领域知识图谱中选取与类型对应的候选实体；

计算模块 504，用于计算所述命名实体与对应的候选实体之间的匹配度，并根据所述匹配度生成修正集合；

45 修正模块 505，用于从所述修正集合中选取候选实体，对所述待纠错文本中对应的命名实体进行修正，得到修正文本。

其中, 所述计算模块 504 包括:

第一计算单元 5041, 用于计算所述命名实体与对应的候选实体之间的字形相似度, 若所述字形相似度大于预设的字形相似阈值, 则汇集所述候选实体生成修正集合;

5 第二计算单元 5042, 用于计算所述命名实体与对应的候选实体之间的字音相似度, 若所述字音相似度大于预设的字音相似阈值, 则汇集所述候选实体生成修正集合;

第三计算单元 5043, 用于分析所述命名实体与对应的候选实体的字词结构, 并基于所述字词结构确定所述命名实体与所述候选实体之间的相似度, 若所述相似度大于预设的字词结构相似阈值, 则汇集所述候选实体生成修正集合。

其中, 所述计算模块 504 还包括转换单元 5044, 其具体用于:

10 判断所述命名实体是否为拼音汉字混合拼写;

若所述命名实体为拼音汉字混合拼写, 则基于预设的拼音转化算法, 将所述命名实体中的汉字对应转化为拼音。

其中, 所述第三计算单元 5043 具体用于:

15 分析所述命名实体与对应的候选实体之间的字词组合, 计算所述字词组合的相似度; 若所述字词组合的相似度大于预设的字词组合相似阈值, 则汇集所述候选实体生成修正集合;

分析所述命名实体与对应的候选实体之间的字序, 计算所述字序的相似度; 若所述字序的相似度大于预设的字序相似阈值, 则汇集所述候选实体生成修正集合。

其中, 所述修正模块 505 包括:

20 判断单元 5051, 用于判断所述修正集合是否包含多个所述候选实体;

计算单元 5052, 用于若所述修正集合包含多个所述候选实体, 则根据预设的领域语言模型, 计算所述候选实体在所述待纠错文本中的出现概率;

排序单元 5053, 用于对所述候选实体按照所述出现概率的大小进行排序, 得到出现序列;

25 修正单元 5054, 用于根据所述出现序列从所述修正集合中选取候选实体, 对所述待纠错文本中对应的命名实体进行修正, 得到修正文本。

其中, 所述分词模块 501 具体用于:

获取待纠错文本, 并根据预设的词典, 将所述待纠错文本生成前缀树; 对所述前缀树进行词图扫描, 生成有向无环图;

30 调用预设的动态规划查找最大概率路径算法, 从所述有向无环图中查找基于词频的最大切分组合;

将所述待纠错文本按照所述最大切分组合进行分词处理, 得到词序列; 将所述词序列输入预设的基于词序列的命名实体识别模型, 输出得到命名实体集。

其中, 所述识别模块 502 具体用于:

35 将所述命名实体集输入至预设的卷积神经网络中, 并调用所述卷积神经网络的逻辑回归函数, 计算所述命名实体集中各命名实体所涉及的各个领域的领域属性值;

对各个所述领域属性值进行比较, 将所述领域属性值最大的领域作为所述命名实体集中各命名实体的垂直领域;

40 基于所述卷积神经网络中的卷积层, 提取所述命名实体集中各命名实体的类型特征信息;

计算所述类型特征信息与预设的类型之间的匹配度, 根据所述匹配度确定所述命名实体的类型。

45 在本申请实施例中, 通过文本纠错装置对待纠错文本进行一系列处理生成词序列, 再输入至命名实体识别模型中基于词序列进行命名实体的识别, 从而可以根据命名实体准确、快速的定位错误, 对命名实体和候选实体分别进行字形分析、字音分析和字词结构分

析，从而可以识别文本的多种错误，对待纠错文本有针对性地进行修正，计算候选实体在待纠错文本中的出现概率，并根据出现概率选取出现概率最大的候选实体对待纠错文本进行修正，提高了对待纠错文本中命名实体修正的准确度。

请参阅图 7，下面从硬件处理的角度对本申请实施例中的文本纠错设备的一个实施例进行详细描述。

图 7 是本申请实施例提供的一种文本纠错设备的结构示意图，该文本纠错设备 700 可因配置或性能不同而产生比较大的差异，可以包括一个或一个以上处理器（central processing units, CPU）710（例如，一个或一个以上处理器）和存储器 720，一个或一个以上存储应用程序 733 或数据 732 的存储介质 730（例如一个或一个以上海量存储设备）。其中，存储器 720 和存储介质 730 可以是短暂存储或持久存储。存储在存储介质 730 的程序可以包括一个或一个以上模块（图示没标出），每个模块可以包括对文本纠错设备 700 中的一系列指令操作。更进一步地，处理器 710 可以设置为与存储介质 730 通信，在文本纠错设备 700 上执行存储介质 730 中的一系列指令操作。

文本纠错设备 700 还可以包括一个或一个以上电源 740，一个或一个以上有线或无线网络接口 750，一个或一个以上输入输出接口 760，和或或，一个或一个以上操作系统 731，例如 Windows Serve, Mac OS X, Unix, Linux, FreeBSD 等等。本领域技术人员可以理解，图 7 示出的文本纠错设备结构并不构成对文本纠错设备的限定，可以包括比图示更多或更少的部件，或者组合某些部件，或者不同的部件布置。

本申请所指区块链是分布式数据存储、点对点传输、共识机制、加密算法等计算机技术的新型应用模式。区块链（Blockchain），本质上是一个去中心化的数据库，是一串使用密码学方法相关联产生的数据块，每一个数据块中包含了一批次网络交易的信息，用于验证其信息的有效性（防伪）和生成下一个区块。区块链可以包括区块链底层平台、平台产品服务层以及应用服务层等。

本申请还提供一种计算机可读存储介质，该计算机可读存储介质可以为非易失性计算机可读存储介质，该计算机可读存储介质也可以为易失性计算机可读存储介质，所述计算机可读存储介质中存储有指令，当所述指令在计算机上运行时，使得计算机执行所述文本纠错方法的步骤。

所属领域的技术人员可以清楚地了解到，为描述的方便和简洁，上述描述的装置和单元的具体工作过程，可以参考前述方法实施例中的对应过程，在此不再赘述。

所述集成的单元如果以软件功能单元的形式实现并作为独立的产品销售或使用，可以存储在一个计算机可读存储介质中。基于这样的理解，本申请的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的全部或部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质中，包括若干指令用以使得一台计算机设备（可以是个人计算机，服务器，或者网络设备等）执行本申请各个实施例所述方法的全部或部分步骤。而前述的存储介质包括：U 盘、移动硬盘、只读存储器（read-only memory, ROM）、随机存取存储器（random access memory, RAM）、磁碟或者光盘等各种可以存储程序代码的介质。

以上所述，以上实施例仅用以说明本申请的技术方案，而非对其限制；尽管参照前述实施例对本申请进行了详细的说明，本领域的普通技术人员应当理解：其依然可以对前述各实施例所记载的技术方案进行修改，或者对其中部分技术特征进行等同替换；而这些修改或者替换，并不使相应技术方案的本质脱离本申请各实施例技术方案的精神和范围。

权 利 要 求 书

1、一种文本纠错方法，其中，所述文本纠错方法包括：

获取待纠错文本，并对所述待纠错文本进行分词处理，得到命名实体集；

5 将所述命名实体集输入至预设的卷积神经网络中进行领域识别，确定所述命名实体集中各命名实体的垂直领域及类型；

从预设的领域知识图谱集中选取与所述垂直领域对应的领域知识图谱，并从所述领域知识图谱中选取与所述类型对应的候选实体；

计算所述命名实体与对应的候选实体之间的匹配度，并根据所述匹配度生成修正集合；

10 从所述修正集合中选取候选实体，对所述待纠错文本中对应的命名实体进行修正，得到修正文本。

2、根据权利要求1所述的文本纠错方法，其中，所述计算所述命名实体与对应的候选实体之间的匹配度，并根据所述匹配度生成修正集合包括：

15 计算所述命名实体与对应的候选实体之间的字形相似度，若所述字形相似度大于预设的字形相似阈值，则汇集所述候选实体生成修正集合；

或，计算所述命名实体与对应的候选实体之间的字音相似度，若所述字音相似度大于预设的字音相似阈值，则汇集所述候选实体生成修正集合；

20 或，分析所述命名实体与对应的候选实体的字词结构，并基于所述字词结构确定所述命名实体与所述候选实体之间的相似度，若所述相似度大于预设的字词结构相似阈值，则汇集所述候选实体生成修正集合。

3、根据权利要求2所述的文本纠错方法，其中，在所述计算所述命名实体与对应的候选实体之间的字音相似度，若所述字音相似度大于预设的字音相似阈值，则汇集所述候选实体生成修正集合之前，还包括：

判断所述命名实体是否为拼音汉字混合拼写；

25 若是，则基于预设的拼音转化算法，将所述命名实体中的汉字对应转化为拼音。

4、根据权利要求2所述的文本纠错方法，其中，所述分析所述命名实体与对应的候选实体的字词结构，并基于所述字词结构确定所述命名实体与所述候选实体之间的相似度，若所述相似度大于预设的字词结构相似阈值，则汇集所述候选实体生成修正集合包括：

30 分析所述命名实体与对应的候选实体的字词组合，计算所述字词组合的相似度；若所述字词组合的相似度大于预设的字词组合相似阈值，则汇集所述候选实体生成修正集合；

或，分析所述命名实体与对应的候选实体的字序，计算所述字序的相似度；若所述字序的相似度大于预设的字序相似阈值，则汇集所述候选实体生成修正集合。

5、根据权利要求1-4中任一项所述的文本纠错方法，其中，所述从所述修正集合中选取候选实体，对所述待纠错文本中对应的命名实体进行修正，得到修正文本包括：

35 判断所述修正集合是否包含多个所述候选实体；

若是，则根据预设的领域语言模型，计算所述候选实体在所述待纠错文本中的出现概率；

对所述候选实体按照所述出现概率的大小进行排序，得到出现序列；

40 根据所述出现序列从所述修正集合中选取候选实体，对所述待纠错文本中对应的命名实体进行修正，得到修正文本。

6、根据权利要求5所述的文本纠错方法，其中，所述获取待纠错文本，并对所述待纠错文本进行分词处理，得到命名实体集包括：

获取待纠错文本，并根据预设的词典，将所述待纠错文本生成前缀树；

对所述前缀树进行词图扫描，生成有向无环图；

调用预设的动态规划查找最大概率路径算法,从所述有向无环图中查找基于词频的最大切分组合;

将所述待纠错文本按照所述最大切分组合进行分词处理,得到词序列;

将所述词序列输入预设的基于词序列的命名实体识别模型,输出得到命名实体集。

5

7、根据权利要求5所述的文本纠错方法,其中,所述将所述命名实体集输入至预设的卷积神经网络中进行领域识别,确定所述命名实体集中各命名实体的垂直领域及类型包括:

10 将所述命名实体集输入至预设的卷积神经网络中,并调用所述卷积神经网络的逻辑回归函数,计算所述命名实体集中各命名实体所涉及的各个领域的领域属性值;

对各个所述领域属性值进行比较,将所述领域属性值最大的领域作为所述命名实体集中各命名实体的垂直领域;

基于所述卷积神经网络中的卷积层,提取所述命名实体集中各命名实体的类型特征信息;

15 计算所述类型特征信息与预设的类型之间的匹配度,根据所述匹配度确定所述命名实体的类型。

8、一种文本纠错设备,其中,所述文本纠错设备包括:

存储器 and 至少一个处理器,所述存储器中存储有指令,所述存储器和所述至少一个处理器通过线路互连;

20 所述至少一个处理器调用所述存储器中的所述指令,以使得所述网络访问探测设备执行如下所述的文本纠错方法的步骤:

获取待纠错文本,并对所述待纠错文本进行分词处理,得到命名实体集;

将所述命名实体集输入至预设的卷积神经网络中进行领域识别,确定所述命名实体集中各命名实体的垂直领域及类型;

25 从预设的领域知识图谱集中选取与所述垂直领域对应的领域知识图谱,并从所述领域知识图谱中选取与所述类型对应的候选实体;

计算所述命名实体与对应的候选实体之间的匹配度,并根据所述匹配度生成修正集合;

30 从所述修正集合中选取候选实体,对所述待纠错文本中对应的命名实体进行修正,得到修正文本。

9、根据权利要求8所述的文本纠错设备,其中,所述文本纠错程序被所述处理器执行实现所述计算所述命名实体与对应的候选实体之间的匹配度,并根据所述匹配度生成修正集合的步骤时,还执行以下步骤:

35 计算所述命名实体与对应的候选实体之间的字形相似度,若所述字形相似度大于预设的字形相似阈值,则汇集所述候选实体生成修正集合;

或,计算所述命名实体与对应的候选实体之间的字音相似度,若所述字音相似度大于预设的字音相似阈值,则汇集所述候选实体生成修正集合;

40 或,分析所述命名实体与对应的候选实体的字词结构,并基于所述字词结构确定所述命名实体与所述候选实体之间的相似度,若所述相似度大于预设的字词结构相似阈值,则汇集所述候选实体生成修正集合。

10、根据权利要求9所述的文本纠错设备,其中,所述文本纠错程序被所述处理器执行实现所述计算所述命名实体与对应的候选实体之间的字音相似度,若所述字音相似度大于预设的字音相似阈值,则汇集所述候选实体生成修正集合的步骤之前,还执行以下步骤:

45 判断所述命名实体是否为拼音汉字混合拼写;

若是，则基于预设的拼音转化算法，将所述命名实体中的汉字对应转化为拼音。

11、根据权利要求 9 所述的文本纠错设备，其中，所述文本纠错程序被所述处理器执行实现所述分析所述命名实体与对应的候选实体的字词结构，并基于所述字词结构确定所述命名实体与所述候选实体之间的相似度，若所述相似度大于预设的字词结构相似阈值，
5 则汇集所述候选实体生成修正集合的步骤时，还执行以下步骤：

分析所述命名实体与对应的候选实体的字词组合，计算所述字词组合的相似度；若所述字词组合的相似度大于预设的字词组合相似阈值，则汇集所述候选实体生成修正集合；

或，分析所述命名实体与对应的候选实体的字序，计算所述字序的相似度；若所述字序的相似度大于预设的字序相似阈值，则汇集所述候选实体生成修正集合。

12、根据权利要求 8-11 中任一项所述的文本纠错设备，其中，所述文本纠错程序被所述处理器执行实现所述从所述修正集合中选取候选实体，对所述待纠错文本中对应的命名实体进行修正，得到修正文本的步骤时，还执行以下步骤：

判断所述修正集合是否包含多个所述候选实体；

若是，则根据预设的领域语言模型，计算所述候选实体在所述待纠错文本中的出现概
15 率；

对所述候选实体按照所述出现概率的大小进行排序，得到出现序列；

根据所述出现序列从所述修正集合中选取候选实体，对所述待纠错文本中对应的命名实体进行修正，得到修正文本。

13、根据权利要求 12 所述的文本纠错设备，其中，所述文本纠错程序被所述处理器执行实现所述获取待纠错文本，并对所述待纠错文本进行分词处理，得到命名实体集的步骤时，还执行以下步骤：

获取待纠错文本，并根据预设的词典，将所述待纠错文本生成前缀树；

对所述前缀树进行词图扫描，生成有向无环图；

调用预设的动态规划查找最大概率路径算法，从所述有向无环图中查找基于词频的最
25 大切分组合；

将所述待纠错文本按照所述最大切分组合进行分词处理，得到词序列；

将所述词序列输入预设的基于词序列的命名实体识别模型，输出得到命名实体集。

14、一种计算机可读存储介质，所述计算机可读存储介质上存储有指令，其中，所述指令被处理器执行时实现如下所述的文本纠错方法的步骤：

获取待纠错文本，并对所述待纠错文本进行分词处理，得到命名实体集；

将所述命名实体集输入至预设的卷积神经网络中进行领域识别，确定所述命名实体集中各命名实体的垂直领域及类型；

从预设的领域知识图谱集中选取与所述垂直领域对应的领域知识图谱，并从所述领域知识图谱中选取与所述类型对应的候选实体；

35 计算所述命名实体与对应的候选实体之间的匹配度，并根据所述匹配度生成修正集合；

从所述修正集合中选取候选实体，对所述待纠错文本中对应的命名实体进行修正，得到修正文本。

15、根据权利要求 14 所述的计算机可读存储介质，其中，所述计算机程序被处理器执行所述计算所述命名实体与对应的候选实体之间的匹配度，并根据所述匹配度生成修正集合的步骤时，还执行如下步骤：

计算所述命名实体与对应的候选实体之间的字形相似度，若所述字形相似度大于预设的字形相似阈值，则汇集所述候选实体生成修正集合；

或，计算所述命名实体与对应的候选实体之间的字音相似度，若所述字音相似度大于
45 预设的字音相似阈值，则汇集所述候选实体生成修正集合；

或,分析所述命名实体与对应的候选实体的字词结构,并基于所述字词结构确定所述命名实体与所述候选实体之间的相似度,若所述相似度大于预设的字词结构相似阈值,则汇集所述候选实体生成修正集合。

16、根据权利要求15所述的计算机可读存储介质,其中,所述计算机程序被处理器执行在所述计算所述命名实体与对应的候选实体之间的字音相似度,若所述字音相似度大于预设的字音相似阈值,则汇集所述候选实体生成修正集合的步骤之前,还执行如下步骤:

判断所述命名实体是否为拼音汉字混合拼写;

若是,则基于预设的拼音转化算法,将所述命名实体中的汉字对应转化为拼音。

17、根据权利要求15所述的计算机可读存储介质,其中,所述计算机程序被处理器执行所述分析所述命名实体与对应的候选实体的字词结构,并基于所述字词结构确定所述命名实体与所述候选实体之间的相似度,若所述相似度大于预设的字词结构相似阈值,则汇集所述候选实体生成修正集合的步骤时,还执行如下步骤:

分析所述命名实体与对应的候选实体的字词组合,计算所述字词组合的相似度;若所述字词组合的相似度大于预设的字词组合相似阈值,则汇集所述候选实体生成修正集合;

或,分析所述命名实体与对应的候选实体的字序,计算所述字序的相似度;若所述字序的相似度大于预设的字序相似阈值,则汇集所述候选实体生成修正集合。

18、根据权利要求14-17中任一项所述的计算机可读存储介质,其中,所述计算机程序被处理器执行所述从所述修正集合中选取候选实体,对所述待纠错文本中对应的命名实体进行修正,得到修正文本的步骤时,还执行如下步骤:

判断所述修正集合是否包含多个所述候选实体;

若是,则根据预设的领域语言模型,计算所述候选实体在所述待纠错文本中的出现概率;

对所述候选实体按照所述出现概率的大小进行排序,得到出现序列;

根据所述出现序列从所述修正集合中选取候选实体,对所述待纠错文本中对应的命名实体进行修正,得到修正文本。

19、根据权利要求18所述的计算机可读存储介质,其中,所述计算机程序被处理器执行所述获取待纠错文本,并对所述待纠错文本进行分词处理,得到命名实体集的步骤时,还执行如下步骤:

获取待纠错文本,并根据预设的词典,将所述待纠错文本生成前缀树;

对所述前缀树进行词图扫描,生成有向无环图;

调用预设的动态规划查找最大概率路径算法,从所述有向无环图中查找基于词频的最大切分组合;

将所述待纠错文本按照所述最大切分组合进行分词处理,得到词序列;

将所述词序列输入预设的基于词序列的命名实体识别模型,输出得到命名实体集。

20、一种文本纠错装置,其中,所述文本纠错装置包括:

分词模块,用于获取待纠错文本,并对所述待纠错文本进行分词处理,得到命名实体;

识别模块,用于将所述命名实体集输入至预设的卷积神经网络中进行领域识别,确定所述命名实体集中各命名实体的垂直领域及类型;

选取模块,用于从预设的领域知识图谱集中选取与所述垂直领域对应的领域知识图谱,并从所述领域知识图谱中选取与所述类型对应的候选实体;

计算模块,用于计算所述命名实体与对应的候选实体之间的匹配度,并根据所述匹配度生成修正集合;

修正模块,用于从所述修正集合中选取候选实体,对所述待纠错文本中对应的命名实体进行修正,得到修正文本。

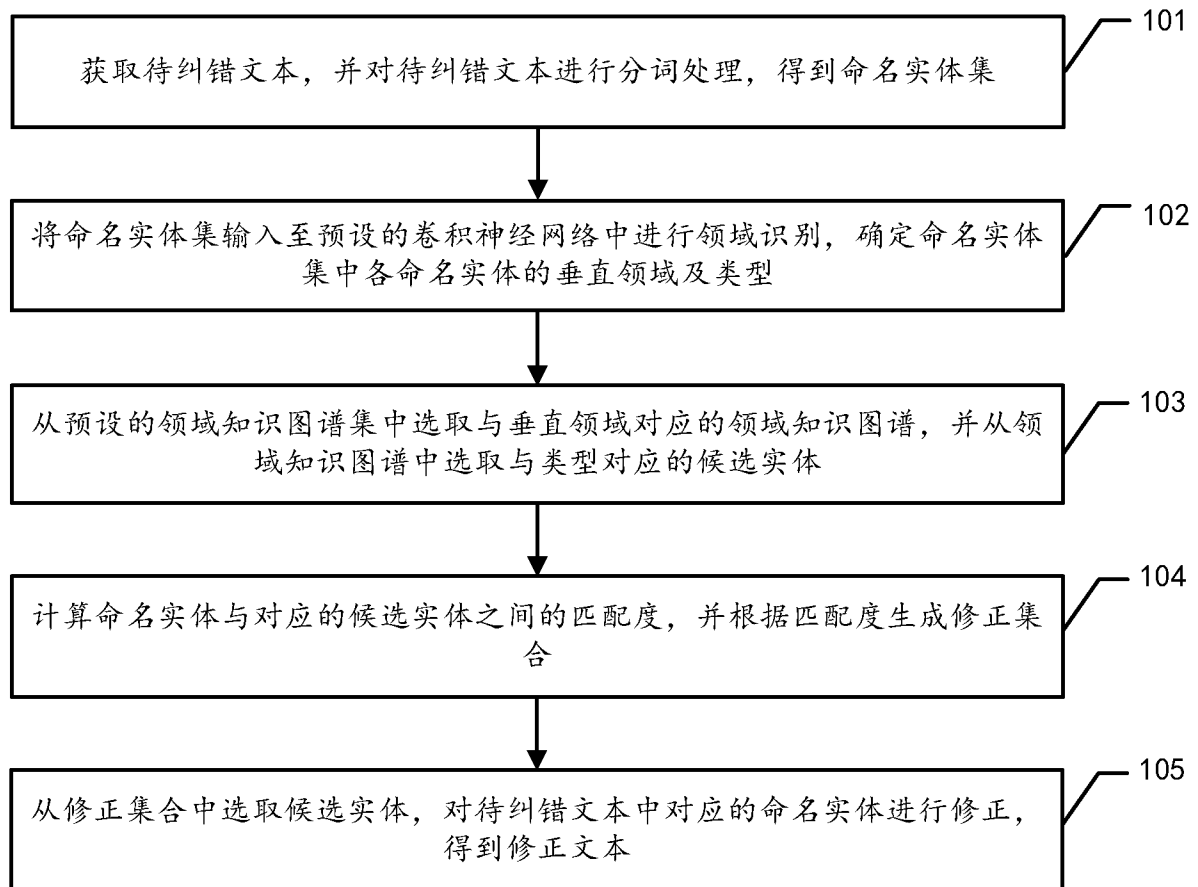


图 1

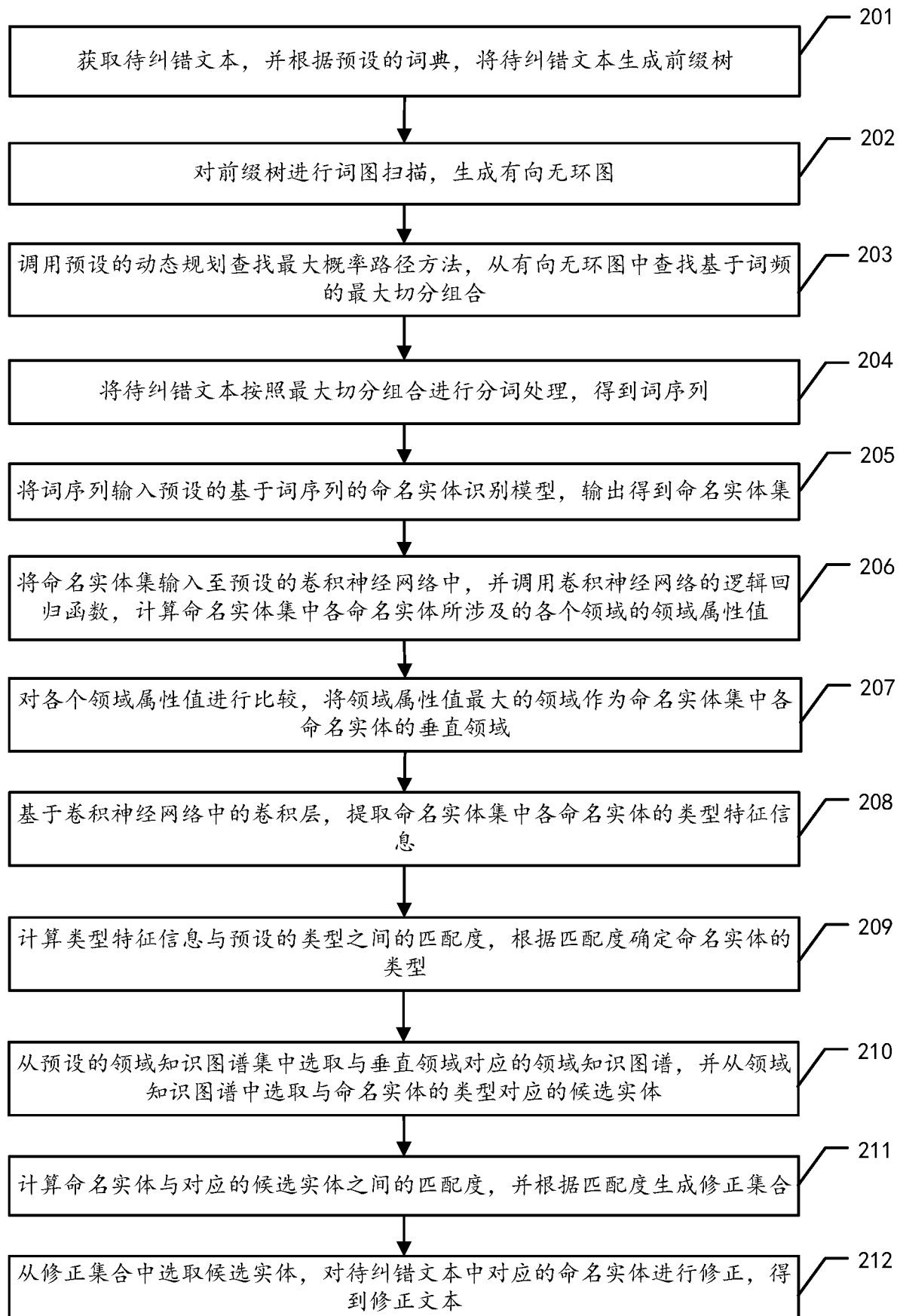


图 2

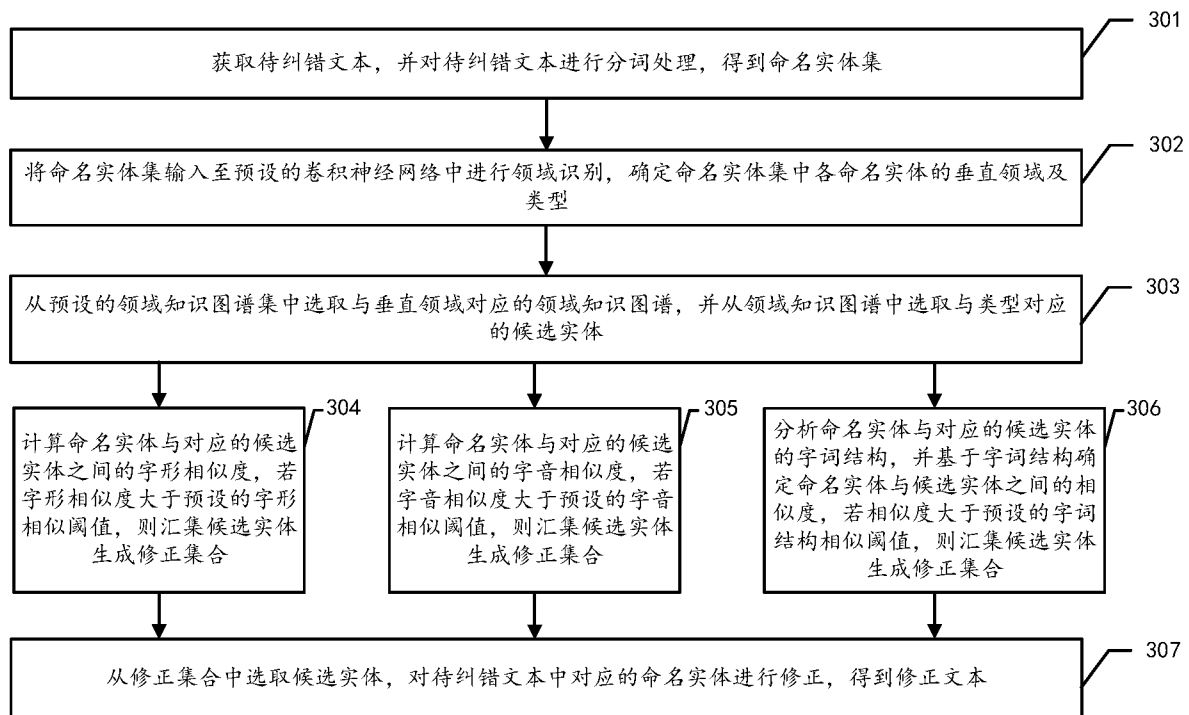


图 3

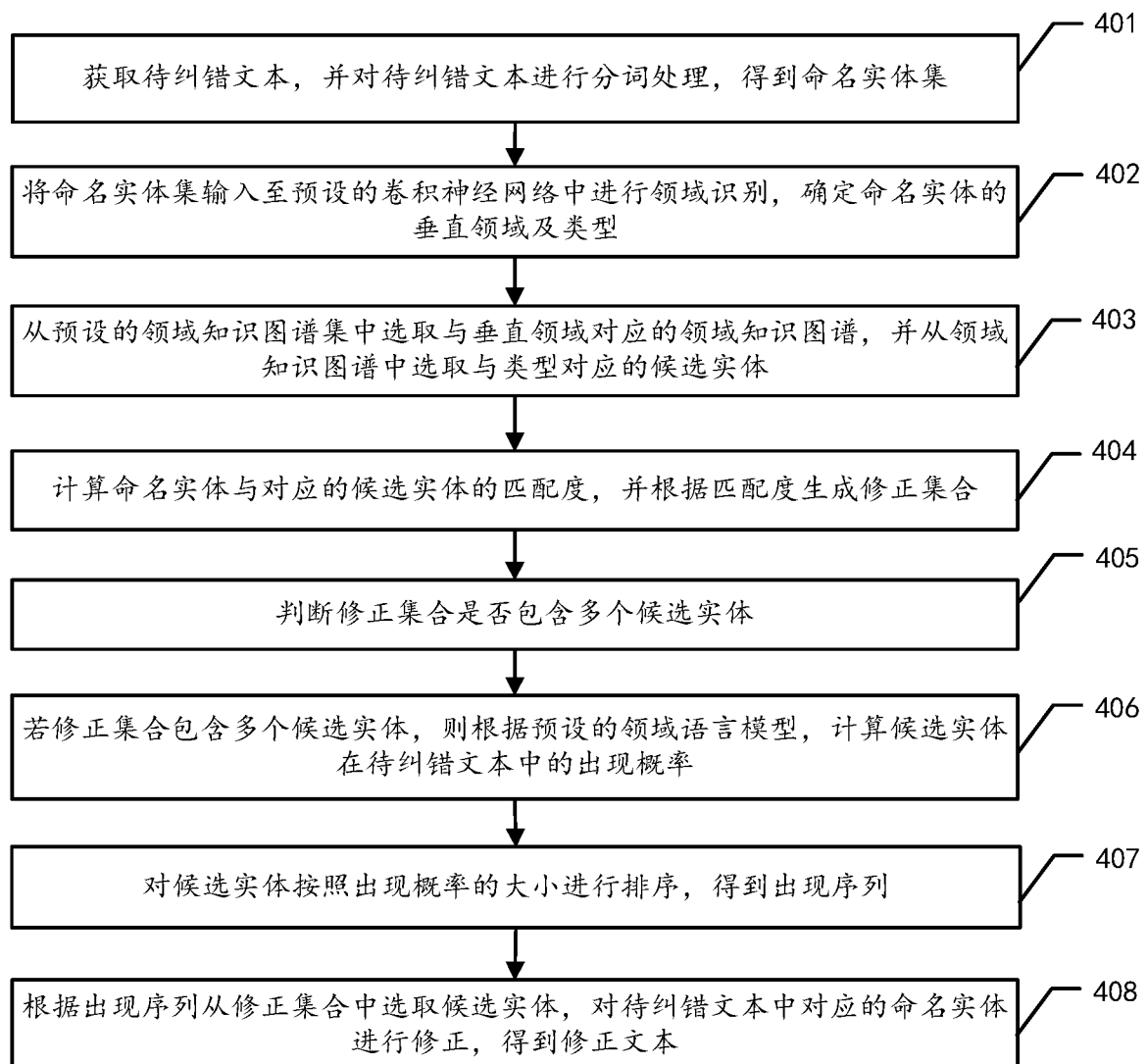


图 4

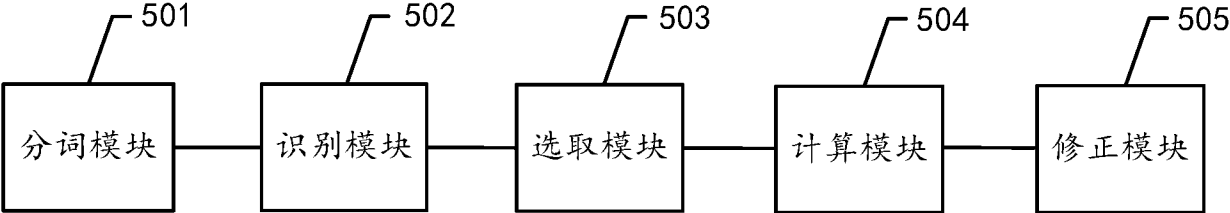


图 5

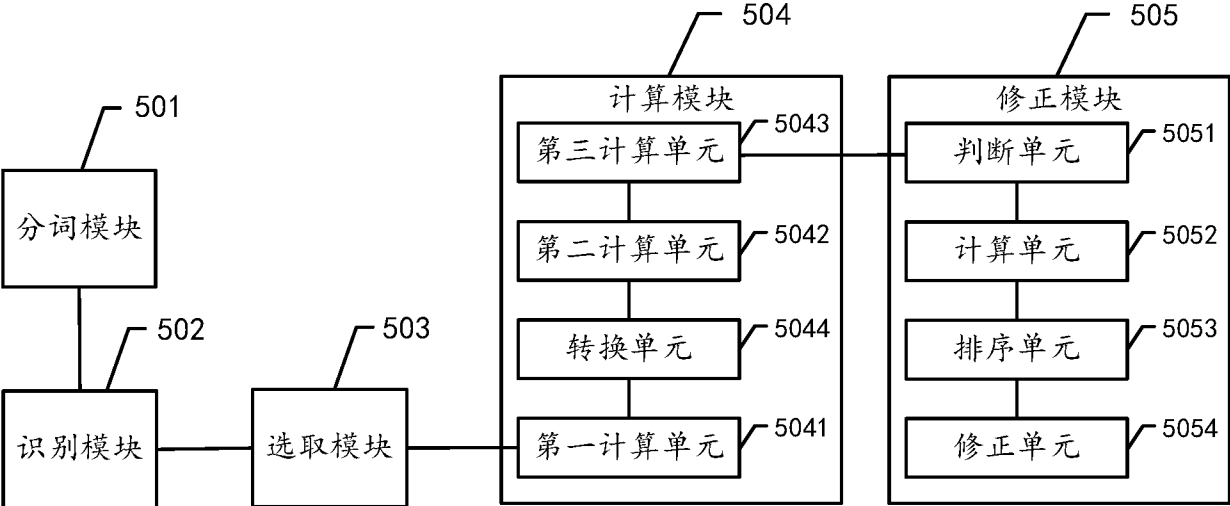


图 6

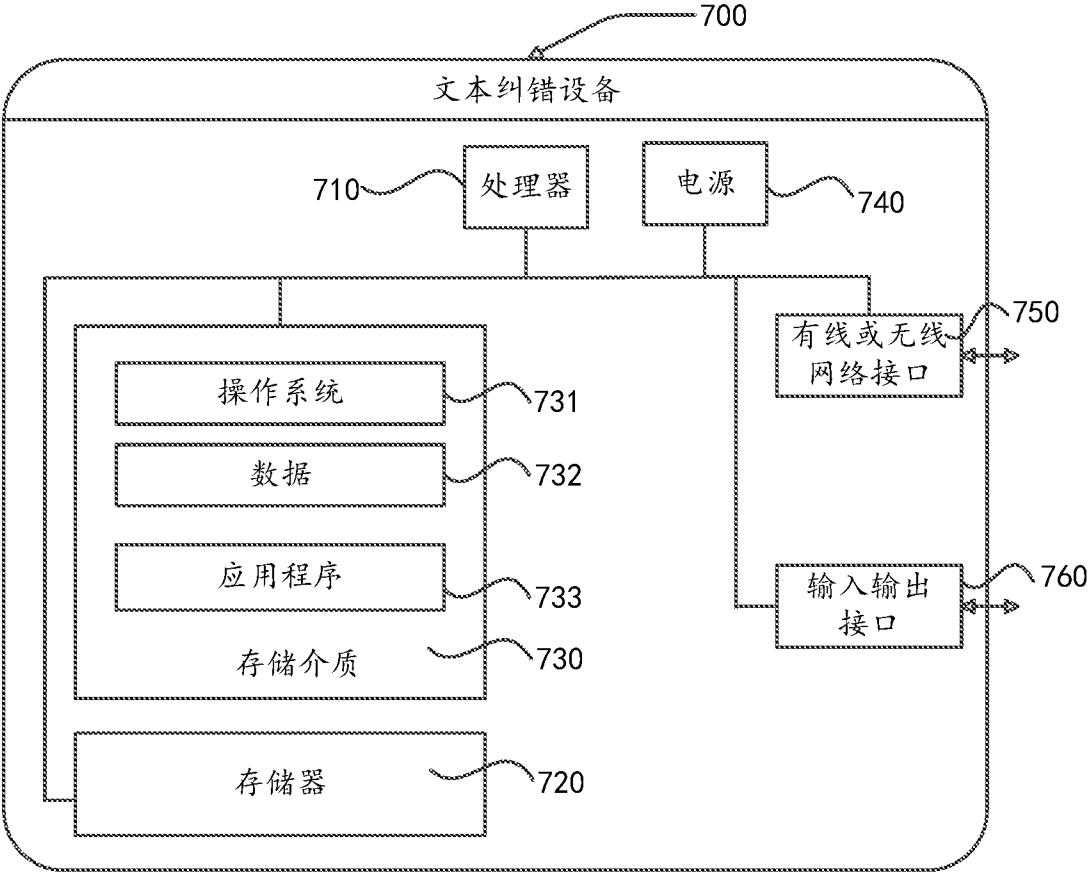


图 7

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2022/088892

A. CLASSIFICATION OF SUBJECT MATTER

G06F 40/00(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)

G06F40/-

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)

WPABSC; CNTXT; CNABS; WPABS; ENTXT; CJFD; DWPI; ENTXTC; VEN, CNKI: 互联网; 文本, 文档, 卷积网络, 神经网络, 深度学习, 纠正, 纠错, 校正, 切词, 分词, 实体, 领域, 图谱, word, text, neural network, CNN, DNN, correct, modify, entity, NER, domain, knowledge graph

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 113591457 A (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) 02 November 2021 (2021-11-02) claims 1-10	1-20
Y	CN 111291571 A (HUAWEI TECHNOLOGIES CO., LTD.) 16 June 2020 (2020-06-16) description, paragraphs [0002]-[0136]	1-20
Y	CN 110597992 A (TENCENT TECHNOLOGY SHENZHEN CO., LTD.) 20 December 2019 (2019-12-20) description, paragraphs [0087]-[0152]	1-20
A	CN 109800407 A (TENCENT TECHNOLOGY (SHENZHEN) CO., LTD.; TENCENT CLOUD COMPUTING (BEIJING) CO., LTD.) 24 May 2019 (2019-05-24) entire document	1-20
A	CN 110147549 A (ALIBABA GROUP HOLDING LTD.) 20 August 2019 (2019-08-20) entire document	1-20
A	CN 110750993 A (CHENGDU BUSINESS BIG DATA TECHNOLOGY CO., LTD.) 04 February 2020 (2020-02-04) 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

13 July 2022

Date of mailing of the international search report

21 July 2022

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

International application No.

PCT/CN2022/088892

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2021050017 A1 (SAMSUNG ELECTRONICS CO., LTD.) 18 February 2021 (2021-02-18) entire document	1-20

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2022/088892

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	113591457	A	02 November 2021	None			
CN	111291571	A	16 June 2020	WO	2021143299		22 July 2021
CN	110597992	A	20 December 2019	None			
CN	109800407	A	24 May 2019	CN	109800407	B	16 November 2021
CN	110147549	A	20 August 2019	None			
CN	110750993	A	04 February 2020	None			
US	2021050017	A1	18 February 2021	WO	2021029643	A1	18 February 2021
				EP	3790002	A1	10 March 2021

国际检索报告

国际申请号

PCT/CN2022/088892

A. 主题的分类

G06F 40/00 (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)

G06F40/-

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))

WPABSC;CNTXT;CNABS;WPABS;ENTXT;CJFD;DWPI;ENTXTC;VEN, CNKI, 互联网:文本, 文档, 卷积网络, 神经网络, 深度学习, 纠正, 纠错, 校正, 切词, 分词, 实体, 领域, 图谱, word, text, neural network, CNN, DNN, correct, modify, entity, NER, domain, knowledge graph

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 113591457 A (平安科技深圳有限公司) 2021年11月2日 (2021 - 11 - 02) 权利要求1-10	1-20
Y	CN 111291571 A (华为技术有限公司) 2020年6月16日 (2020 - 06 - 16) 说明书第[0002]-[0136]段	1-20
Y	CN 110597992 A (腾讯科技深圳有限公司) 2019年12月20日 (2019 - 12 - 20) 说明书第[0087]-[0152]段	1-20
A	CN 109800407 A (腾讯科技深圳有限公司 腾讯云计算北京有限责任公司) 2019年5月24日 (2019 - 05 - 24) 全文	1-20
A	CN 110147549 A (阿里巴巴集团控股有限公司) 2019年8月20日 (2019 - 08 - 20) 全文	1-20
A	CN 110750993 A (成都数联铭品科技有限公司) 2020年2月4日 (2020 - 02 - 04) 全文	1-20
A	US 2021050017 A1 (SAMSUNG ELECTRONICS CO LTD) 2021年2月18日 (2021 - 02 - 18) 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

2022年7月13日

国际检索报告邮寄日期

2022年7月21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

易建琼

电话号码 (86-28)-62967650

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2022/088892

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	113591457	A	2021年11月2日	无			
CN	111291571	A	2020年6月16日	W0	2021143299		2021年7月22日
CN	110597992	A	2019年12月20日	无			
CN	109800407	A	2019年5月24日	CN	109800407	B	2021年11月16日
CN	110147549	A	2019年8月20日	无			
CN	110750993	A	2020年2月4日	无			
US	2021050017	A1	2021年2月18日	W0	2021029643	A1	2021年2月18日
				EP	3790002	A1	2021年3月10日