(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 102135814 A
(43)申请公布日 2011.07.27

(21)申请号 201110079201.1

(22)申请日 2011.03.30

(71)申请人 北京搜狗科技发展有限公司
地址 100084 北京市海淀区中关村东路 1 号院 9 号楼搜狐网络大厦 9 层 01 房间

(72)发明人 张扬

(74)专利代理机构 北京集佳知识产权代理有限公司 11227
代理人 逯长明 王宝筠

(51)Int.Cl.
G06F 3/023(2006.01)
G06F 17/27(2006.01)
G06F 17/30(2006.01)

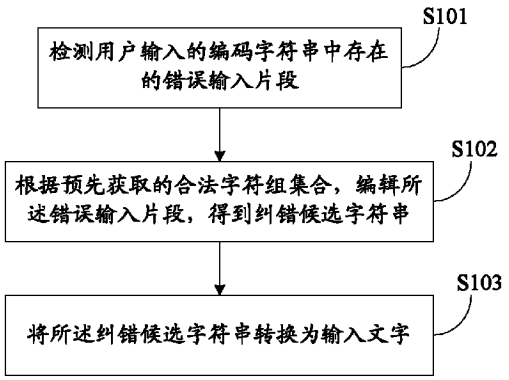
权利要求书 3 页 说明书 12 页 附图 1 页

(54)发明名称

一种字词输入方法及系统

(57)摘要

本发明公开了一种字词输入方法及系统，其中，所述方法包括：检测用户输入的编码字符串中存在的错误输入片段；根据预先获取的合法字符组集合，编辑所述错误输入片段，得到纠错候选字符串；将所述纠错候选字符串转换为输入文字。通过本发明，能够在在字词输入过程中，更有效地进行纠错，并且适用范围比较广泛。



CN 102135814 A

1. 一种字词输入方法，其特征在于，包括：

检测用户输入的编码字符串中存在的错误输入片段；

根据预先获取的合法字符组集合，编辑所述错误输入片段，得到纠错候选字符串；

将所述纠错候选字符串转换为输入文字。

2. 根据权利要求 1 所述的方法，其特征在于，所述检测用户输入的编码字符串中存在的错误输入片段包括：

根据当前语境，对所述用户输入的编码字符串进行分词，将得到的分词碎片确定为错误输入片段。

3. 根据权利要求 1 所述的方法，其特征在于，所述检测用户输入的编码字符串中存在的错误输入片段包括：

如果所述用户输入的编码字符串中存在不属于所述合法字符组集合的字符组，或者基于合法字符组集合统计的合法概率小于阈值的字符组，则该字符组为错误输入片段。

4. 根据权利要求 1 所述的方法，其特征在于，所述合法字符组集合通过以下方式获得：从至少两个文字的合法编码字符串连接而成的字符串中抽取合法字符组。

5. 根据权利要求 1 所述的方法，其特征在于，所述根据预先获取的合法字符组集合，编辑所述错误输入片段，得到纠错候选字符串包括：

对所述错误输入片段分别进行基于字符的替换、插入、删除及交换的处理；

如果处理后的片段属于所述合法字符组集合，则基于该处理后的片段生成纠错候选字符串。

6. 根据权利要求 5 所述的方法，其特征在于，所述根据预先获取的合法字符组集合，编辑所述错误输入片段，得到纠错候选字符串包括：

去除合法字符组中设定位置的至少一个字符，将其他剩余字符按设定顺序排列作为所述被去除字符的索引；

以合法字符组集合的若干个所述索引及对应的被去除字符组成反查字符组集合；

当需要对所述错误输入片段在所述设定位置进行字符的替换或插入时，根据所述错误输入片段以所述设定顺序生成查询串；

从所述反查字符组集合中获取以所述查询串为索引的字符，将该字符作为在所述设定位置替换或插入的字符，得到纠错候选字符串。

7. 根据权利要求 6 所述的方法，其特征在于，所述设定位置包括合法字符组中首字符位置以外的其他位置。

8. 根据权利要求 6 所述的方法，其特征在于，所述合法字符组集合及所述反查字符组集合以树形结构进行保存。

9. 根据权利要求 1 至 8 任一项所述的方法，其特征在于，所述将纠错候选字符串转换为输入文字包括：

对所述纠错候选字符串进行评估；

根据评估的结果，对所述编码字符串及符合预置条件的纠错候选字符串进行转换，并向用户展现转换的结果。

10. 根据权利要求 1 至 8 任一项所述的方法，其特征在于，还包括：

将所述用户输入的编码字符串发送到远端服务器，并接收所述远端服务器返回的纠错

候选字符串。

11. 根据权利要求 1 至 8 任一项所述的方法，其特征在于，还包括：

如果所述用户输入的编码字符串命中设定词库，根据所述设定词库将所述编码字符串转换为输入文字。

12. 一种字词输入系统，其特征在于，包括：

检错单元，用于检测用户输入的编码字符串中存在的错误输入片段；

纠错单元，用于根据预先获取的合法字符组集合，编辑所述错误输入片段，得到纠错候选字符串；

转换单元，用于将所述纠错候选字符串转换为输入文字。

13. 根据权利要求 12 所述的系统，其特征在于，所述检错单元包括：

第一检错子单元，用于根据当前语境，对所述用户输入的编码字符串进行分词，将得到的分词碎片确定为错误输入片段。

14. 根据权利要求 12 所述的系统，其特征在于，所述检错单元包括：

第二检错子单元，用于如果所述用户输入的编码字符串中存在不属于所述合法字符组集合的字符组，或者基于合法字符组集合统计的合法概率小于阈值的字符组，则该字符组为错误输入片段。

15. 根据权利要求 12 所述的系统，其特征在于，所述合法字符组集合通过以下方式获得：从至少两个文字的合法编码字符串连接而成的字符串中抽取合法字符组。

16. 根据权利要求 12 所述的系统，其特征在于，所述纠错单元包括：

编辑子单元，用于对所述错误输入片段分别进行基于字符的替换、插入、删除及交换的处理；

生成子单元，用于如果处理后的片段属于所述合法字符组集合，则基于该处理后的片段生成纠错候选字符串。

17. 根据权利要求 16 所述的系统，其特征在于，所述纠错单元包括：

去除子单元，用于去除合法字符组中设定位置的至少一个字符，将其他剩余字符按设定顺序排列作为所述被去除字符的索引；

组合子单元，用于以合法字符组集合的若干个所述索引及对应的被去除字符组成反查字符组集合；

查询串生成子单元，用于当需要对所述错误输入片段在所述设定位置进行字符的替换或插入时，根据所述错误输入片段以所述设定顺序生成查询串；

查询子单元，用于从所述反查字符组集合中获取以所述查询串为索引的字符，将该字符作为在所述设定位置替换或插入的字符，得到纠错候选字符串。

18. 根据权利要求 17 所述的系统，其特征在于，所述设定位置包括合法字符组中首字符位置以外的其他位置。

19. 根据权利要求 17 所述的系统，其特征在于，所述合法字符组集合及所述反查字符组集合以树形结构进行保存。

20. 根据权利要求 12 至 19 任一项所述的系统，其特征在于，所述转换单元包括：

评估子单元，用于对所述纠错候选字符串进行评估；

选择转换子单元，用于根据评估的结果，对所述编码字符串及符合预置条件的纠错候

选字符串进行转换，并向用户展现转换的结果。

21. 根据权利要求 12 至 19 任一项所述的系统，其特征在于，还包括：

云计算单元，用于将所述用户输入的编码字符串发送到远端服务器，并接收所述远端服务器返回的纠错候选字符串。

22. 根据权利要求 12 至 19 任一项所述的系统，其特征在于，还包括：

词库匹配单元，用于如果所述用户输入的编码字符串命中设定词库，根据所述设定词库将所述编码字符串转换为输入文字。

一种字词输入方法及系统

技术领域

[0001] 本发明涉及输入法技术领域，特别是涉及一种字词输入方法及系统。

背景技术

[0002] 随着计算机、互联网等技术应用的越来越广泛，人们很多的日常工作和娱乐都在计算机上进行，用户越来越频繁地需要通过计算机输入信息而完成人机交互。对于中文、日文、韩文等用户而言，一般需要通过输入法程序与计算机进行交互。以中文用户为例，一般来说，用户输入的是一串字母（通常是汉字的拼音）或笔画等编码字符串，系统需要把它转换成相应的中文字符。

[0003] 然而用户在进行字词输入的过程中可能存在较多的错误，这里涉及的输入错误，大致可以分为认知错误和非认知错误两类。认知错误是那些不知道欲输入的字词如何正确拼写，造成的输入错误，模糊音就属于认知错误这个范畴。而非认知错误是指，知道字词如何拼写，但是由于输入时手忙脚乱或者受制于输入设备等而造成了输入错误。

[0004] 目前，一些输入法系统提供了纠错设置，参见图 1，这种方法通常根据大量的用户输入数据训练生成纠错列表；在生成候选之前根据纠错列表中的规则进行强制纠错，例如，根据图 1 所示的纠错列表，如果用户的输入序列中出现了 gn，便直接将其转换为 ng。这种方法虽然在一定程度上实现了自动纠错，但也存在一些缺点，例如，纠错列表是默认预置的，并且列表中的错误片段与正确片段之间是一一对应的。在用户输入的过程中，会将命中了纠错列表的输入片段作为错误输入片段，纠错时，只能将纠错列表中与该错误输入片段对应的片段作为纠错结果。这种方法仅针对一些常见的输入错误比较有效，但是，对于实际应用中的一些不常见的输入错误，这种方法的有效性比较低。

[0005] 因此，需要本领域技术人员迫切解决的一个技术问题就是：如何提供一种更有效、应用范围更广的纠错方案。

发明内容

[0006] 本发明提供一种字词输入方法及系统，能够在在字词输入过程中，更有效地进行纠错，并且适用范围比较广泛。

[0007] 本发明提供了如下方案：

[0008] 一种字词输入方法，包括：

[0009] 检测用户输入的编码字符串中存在的错误输入片段；

[0010] 根据预先获取的合法字符组集合，编辑所述错误输入片段，得到纠错候选字符串；

[0011] 将所述纠错候选字符串转换为输入文字。

[0012] 其中，所述检测用户输入的编码字符串中存在的错误输入片段包括：

[0013] 根据当前语境，对所述用户输入的编码字符串进行分词，将得到的分词碎片确定为错误输入片段。

[0014] 其中，所述检测用户输入的编码字符串中存在的错误输入片段包括：

[0015] 如果所述用户输入的编码字符串中存在不属于所述合法字符组集合的字符组，或者基于合法字符组集合统计的合法概率小于阈值的字符组，则该字符组为错误输入片段。

[0016] 优选地，所述合法字符组集合通过以下方式获得：从至少两个文字的合法编码字符串连接而成的字符串中抽取合法字符组。

[0017] 其中，所述根据预先获取的合法字符组集合，编辑所述错误输入片段，得到纠错候选字符串包括：

[0018] 对所述错误输入片段分别进行基于字符的替换、插入、删除及交换的处理；

[0019] 如果处理后的片段属于所述合法字符组集合，则基于该处理后的片段生成纠错候选字符串。

[0020] 优选地，所述根据预先获取的合法字符组集合，编辑所述错误输入片段，得到纠错候选字符串包括：

[0021] 去除合法字符组中设定位置的至少一个字符，将其他剩余字符按设定顺序排列作为所述被去除字符的索引；

[0022] 以合法字符组集合的若干个所述索引及对应的被去除字符组成反查字符组集合；

[0023] 当需要对所述错误输入片段在所述设定位置进行字符的替换或插入时，根据所述错误输入片段以所述设定顺序生成查询串；

[0024] 从所述反查字符组集合中获取以所述查询串为索引的字符，将该字符作为在所述设定位置替换或插入的字符，得到纠错候选字符串。

[0025] 优选地，所述设定位置包括合法字符组中首字符位置以外的其他位置。

[0026] 优选地，所述合法字符组集合及所述反查字符组集合以树形结构进行保存。

[0027] 优选地，所述将纠错候选字符串转换为输入文字包括：

[0028] 对所述纠错候选字符串进行评估；

[0029] 根据评估的结果，对所述编码字符串及符合预置条件的纠错候选字符串进行转换，并向用户展现转换的结果。

[0030] 优选地，还包括：

[0031] 将所述用户输入的编码字符串发送到远端服务器，并接收所述远端服务器返回的纠错候选字符串。

[0032] 优选地，还包括：

[0033] 如果所述用户输入的编码字符串命中设定词库，根据所述设定词库将所述编码字符串转换为输入文字。

[0034] 一种字词输入系统，包括：

[0035] 检错单元，用于检测用户输入的编码字符串中存在的错误输入片段；

[0036] 纠错单元，用于根据预先获取的合法字符组集合，编辑所述错误输入片段，得到纠错候选字符串；

[0037] 转换单元，用于将所述纠错候选字符串转换为输入文字。

[0038] 其中，所述检错单元包括：

[0039] 第一检错子单元，用于根据当前语境，对所述用户输入的编码字符串进行分词，将

得到的分词碎片确定为错误输入片段。

[0040] 其中，所述检错单元包括：

[0041] 第二检错子单元，用于如果所述用户输入的编码字符串中存在不属于所述合法字符组集合的字符组，或者基于合法字符组集合统计的合法概率小于阈值的字符组，则该字符组为错误输入片段。

[0042] 优选地，所述合法字符组集合通过以下方式获得：从至少两个文字的合法编码字符串连接而成的字符串中抽取合法字符组。

[0043] 优选地，所述纠错单元包括：

[0044] 编辑子单元，用于对所述错误输入片段分别进行基于字符的替换、插入、删除及交换的处理；

[0045] 生成子单元，用于如果处理后的片段属于所述合法字符组集合，则基于该处理后的片段生成纠错候选字符串。

[0046] 优选地，所述纠错单元包括：

[0047] 去除子单元，用于去除合法字符组中设定位置的至少一个字符，将其他剩余字符按设定顺序排列作为所述被去除字符的索引；

[0048] 组合子单元，用于以合法字符组集合的若干个所述索引及对应的被去除字符组成反查字符组集合；

[0049] 查询串生成子单元，用于当需要对所述错误输入片段在所述设定位置进行字符的替换或插入时，根据所述错误输入片段以所述设定顺序生成查询串；

[0050] 查询子单元，用于从所述反查字符组集合中获取以所述查询串为索引的字符，将该字符作为在所述设定位置替换或插入的字符，得到纠错候选字符串。

[0051] 优选地，所述设定位置包括合法字符组中首字符位置以外的其他位置。

[0052] 优选地，所述合法字符组集合及所述反查字符组集合以树形结构进行保存。

[0053] 优选地，所述转换单元包括：

[0054] 评估子单元，用于对所述纠错候选字符串进行评估；

[0055] 选择转换子单元，用于根据评估的结果，对所述编码字符串及符合预置条件的纠错候选字符串进行转换，并向用户展现转换的结果。

[0056] 优选地，还包括：

[0057] 云计算单元，用于将所述用户输入的编码字符串发送到远端服务器，并接收所述远端服务器返回的纠错候选字符串。

[0058] 优选地，还包括：

[0059] 词库匹配单元，用于如果所述用户输入的编码字符串命中设定词库，根据所述设定词库将所述编码字符串转换为输入文字。

[0060] 根据本发明提供的具体实施例，本发明公开了以下技术效果：

[0061] 本发明实施例在用户进行字词输入的过程中，在发现编码字符串中存在的错误输入片段之后，可以根据预先获取的合法字符组集合，通过对错误输入片段进行编辑操作，得到纠错候选字符串，这样，获取纠错候选字符串的方式更加灵活，进而在根据用户输入的编码字符串以及纠错候选字符串进行音字转换时，也可以获得更多可能的候选项。因此，该方法更加有效，且适用范围比较广泛。

[0062] 另外，本发明在对所述错误输入片段进行编辑操作时，使用了反查字符组，这样，针对一些替换或插入操作而言，可以缩小查找范围，从而减少操作次数，提高纠错的效率。

附图说明

[0063] 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的前提下，还可以根据这些附图获得其他的附图。

[0064] 图 1 是本发明实施例提供的方法的流程图；

[0065] 图 2 是本发明实施例提供的系统的示意图。

具体实施方式

[0066] 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所获得的所有其他实施例，都属于本发明保护的范围。

[0067] 参见图 1，本发明实施例提供的字词输入方法包括以下步骤：

[0068] S101：检测用户输入的编码字符串中存在的错误输入片段；

[0069] S102：根据预先获取的合法字符组集合，编辑所述错误输入片段，得到纠错候选字符串；

[0070] S103：将所述纠错候选字符串转换为输入文字。

[0071] 其中，在检测编码字符串中存在的错误输入片段时，可以有多种方法。例如，其中一种方法可以是：使用当前语境下的词典对用户输入的编码字符串进行分词（如，对于汉语而言，每个音节即是一个词），查看是否存在分词碎片，或者是否存在词典中没有出现过的输入片段；如果存在，则证明存在错误输入片段。例如，假设用户输入的编码字符串是“shenem”，按照音节进行分词时，“em”就会成为分词碎片，因为“nem”或“em”都不能构成一个音节，因此，就可以将其作为一个可能的错误输入片段检测出来，等等。

[0072] 另一种方法可以是预先采用对语料库进行统计的方法，基于合法的编码字符串建立合法字符组集合，然后利用合法字符组集合，检测用户输入的编码字符串中是否存在错误输入片段。所谓合法的编码字符串是指符合语言规则的字符串，例如，符合汉语拼音规则的拼音，或者符合五笔规则的字符串，符合笔画规则的字符串等等。当然，对于不同的拼音方案（包括简拼、全拼等），拼音规则可能会有所不同，对应的合法编码字符串也会有所不同，在实际应用中，可以根据实际的需要分别建立合法编码字符串。从合法的编码字符串中取出的字符组合片段，就可以作为合法的字符组合。例如，基于全拼的拼音方案提取合法三元组时，对于“zuzhi”这一合法拼音串，其中的任意三个连续字符都可以组成合法三元组，也即，“zuz”、“uzh”、“zhi”都是合法的字符组合。

[0073] 其中，在选取合法的字符组合时，每个字符组合中的字母个数通常可以取两个、三个或四个等等，分别对应合法二元组、合法三元组、合法四元组，这些都可以成为合法字符组集合的组成部分。为了方便描述，下面仅以合法三元组为例进行介绍。

[0074] 首先需要说明的是，对于英文输入而言，用户在输入的过程中，每输入一个单词之后，会自动输入空格作为分隔；此时输入法就可以将该空格作为单词的边界。因此，在建立合法三元组时，仅需在单词内部进行三元组合。例如，对于英文单词 tutor，可以拆分为 tut、uto、tor 三个合法的三元组。其他单词也可以进行类似处理，从而得到合法三元组集合。在对用户输入的单词进行纠错时，也只需要将用户输入的空格作为单词边界，在每个单词内部使用各个合法三元组进行检错即可。

[0075] 对于中文等文字的输入而言，中文输入的最小粒度是字，如果用户同样习惯进行单字输入，也即，每次仅输入一个字的拼音，则同样可以采用与英文类似的处理，即基于单字的拼音建立合法三元组。但是，更为普遍的情况是，用户在输入编码字符串时，通常会连续输入一个词、短语或句子的拼音。而无论是词、短语还是句子，可能都是由多个字组成的；并且在一次输入过程中，可能并不会在不同字的拼音之间输入空格等符号进行分隔（很多输入法系统都支持这种输入），因此如果仅基于单个字的拼音来创建合法三元组可能是不适用的。因此，在本发明实施例中，合法字符组可以通过以下方式获得：从至少两个文字的合法编码字符串连接而成的字符串中抽取合法字符组。也就是说，针对中文建立合法三元组时，使用的语料库可以是用户可能作为一个整体输入的词、短语或句子，从而从多个合法拼音连接而成的字符串中提取合法三元组。

[0076] 例如，如果根据单字的合法拼音提取合法三元组，则“zuz”是不合法的，这样，在用户想要输入“组织”时，对应输入的拼音串为“zuzhi”，此时，就可能会将“zuz”作为不合法三元组提取出来，但是，显然用户的此次输入中是不存在错误输入的。因此，为了避免这种情况的发生，就应该在建立合法三元组时，就考虑多个拼音组合出现时的情况。例如，假设语料中存在“组织”这个词，则就可以根据其对应的拼音串“zuzhi”提取出以下合法三元组：“zuz”、“uzh”、“zhi”，这样，在用这样的合法三元组进行检错时，就不会将“zuzhi”中的“zuz”或“uzh”作为错误输入片段提取出来。

[0077] 当然，也可以根据用户的实际输入习惯，选择训练语料，并创建相应的合法三元组。例如，某用户习惯一个词一个词地输入，则语料库中的训练语料就可以尽量以词语为主，进而根据各个词语的拼音串建立合法三元组；如果某用户习惯输入长的句子，则语料库中的训练语料就可以尽量以长的句子为主，进而根据各个句子的拼音串建立合法三元组，等等。

[0078] 在建立了合法三元组之后，就可以使用合法三元组对用户输入的编码字符串进行检错，如果编码字符串中某字符组合片段没有出现在合法三元组集合中，则可以将这三个字母作为一个错误输入片段提取出来。此外，还可以根据基于合法字符组集合，基于海量数据以及用户的输入习惯等因素，统计出字符组合片段的合法概率，如果该概率小于某预置的阈值，则可以将该字符组合片段作为一个错误输入片段提取出来。

[0079] 在提取出错误输入片段之后，就可以针对错误输入片段进行纠错。具体在进行纠错时，可以采用替换、插入、删除、交换这四种编辑方式，在一定的编辑距离范围内对该错误输入片段进行处理，然后根据合法字符组合（仍以合法三元组为例），判断处理后的片段是否为合法的字符组合，如果合法，则可以作为纠错候选。其中，编辑距离是指将一个字符串转变为另一个字符串所需要的编辑开销；如果将一次编辑操作的开销全部定义为 1，那么将 zipo 转变为 zippo 的编辑距离是 1，将 englsi 转变为 english 的编辑距离是 2，反过来

也是如此。

[0080] 其中，对于删除及交换这两种操作而言，比较简单，通常也不会具有大的工作量。例如，对于删除操作，就可以尝试删除其中的任意一个字母，然后判断剩余的两个字母是否为合法二元组即可（在进行纠错的过程中，可以同时借助于合法二元组或合法四元组等等）。对于利用合法三元组检出的错误输入片段而言，由于一个片段中最多只有三个字母，因此最多尝试三次操作即可。对于交换操作，可以尝试将任意两个相邻的字母之间的位置交换，然后判断位置交换后的三个字母是否是合法三元组即可。同样的，对于利用合法三元组检出的错误输入片段而言，由于一个片段中最多只有三个字母，因此最多尝试两次交换操作即可。

[0081] 但是，对于替换及插入操作而言，由于需要利用其他的字母替换当前字母，或者插入一个未知的字母，因此，通常可能会产生比较大的工作量，使得获得纠错候选字符串的速度比较低。例如，最简单的方法可以采用穷举的方式，这样，对于替换操作而言，就要对错误输入片段中的每个字母分别进行以下操作：分别用该字母以外的其他 25 个字母中的任意一个替换该字母，然后分别判断替换后的字母片段是否是合法三元组。相当于对于一个具有三个字母的错误输入片段而言，要进行 25*3 = 75 次的替换及判断操作。

[0082] 例如，对于某错误输入片段“zuu”，在使用替换的操作时，首先要用除“z”以外的其他 25 个字母分别替换“z”，然后判断替换之后的片段是否是合法三元组；例如替换为“a”之后，要判断“auu”是否合法，再替换为“b”，再判断“buu”是否合法，等等；然后针对第一个“u”，首先要用除“u”以外的其他 25 个字母分别替换“u”，然后判断替换之后的片段是否是合法三元组；例如替换为“a”之后，要判断“zau”是否合法，再替换为“b”，再判断“zbu”是否合法，等等。最后再针对第二个“u”，进行类似的替换及判断操作，最后将所有合法的替换后的三元组提取出来，作为纠错结果。

[0083] 类似的，对于插入操作而言，则需要在错误输入片段中的具有相邻关系的任意两个字母之间尝试插入“a”到“z”共 26 个字母，并且每插入一次，都需要判断插入之后的字符组合片段是否合法。因此，对于一个具有三个字母的错误输入片段而言，要进行 26*2 = 52 次的插入及判断操作。

[0084] 例如，同样假设错误输入片段是“zuu”，则在使用插入操作进行纠错时，首先需要在“zu”之间分别插入“a”到“z”共 26 个字母中的任意一个，然后判断插入后的字符组合片段是否合法；例如，在插入“a”后，判断“zauu”是否合法，然后再插入“b”，判断“zbuu”是否合法，等等。之后，还要再尝试在“uu”之间分别插入 26 个字母中的任意一个，然后判断插入后的片段是否合法；例如，在插入“a”后，判断“zuau”是否合法，然后再插入“b”，判断“zubu”是否合法，等等。当然，对于插入操作，还可能尝试在片段最前或最后进行插入，此时，又会进一步增加计算量。

[0085] 可见，在纠错过程，在使用替换或插入操作进行纠错时，会存在计算量大、耗费时间长的问题。因此，本发明实施例针对该问题也提供了相应的解决方案，主要是针对在错误输入片段中间部分插入字母的操作步骤，以及将错误输入片段中间部分的字母进行替换的操作步骤进行了简化，从而使得整体上的纠错效率提高。下面进行详细地介绍。

[0086] 为了达到使得上述操作步骤简化的目的，本发明实施例首先根据建立的合法三元组作为正查三元组，然后将正查三元组中的各个字符按照指定的顺序进行排列，例如，进行

倒排等，得到反查三元组。当然，由于实际应用中第一个字母就输错的可能性较低，因此，可以仅对首字符之后的两个字符按照指定顺序进行排列，得到反查三元组。例如，对合法拼音串 tubiao，每 3 个连续字母进行统计获得正查三元组为 tub、ubi、bia、iao ；而将 3 个字母中的后两个字母的顺序交换之后，即可得到反查三元组 tbu、uib、bai、ioa。

[0087] 在针对在错误输入片段中间部分插入字母的操作，以及将错误输入片段中间部分的字母进行替换的操作时，就可以通过查询反查三元组集合，来找出合法的插入或替换后的三元组。例如，对于错误输入片段“sho”，当需要尝试替换中间的“h”时，就可以查询反查三元组列表存在哪些第一个字母为 s，第二个字母为 o 的三元组，找出来之后，直接将后两个字母交换顺序，即可作为依据替换操作得到的纠错结果。例如，发现反查三元组集合中存在“sou”、“soa”等，将后两位倒排后为“suo”、“sao”，因此，就可以直接将“suo”、“sao”等作为纠错结果即可，不用再尝试其他的字母，也不需要进行判断是否合法的操作。

[0088] 同样的，对于错误输入片段“sho”，如果需要尝试在“ho”之间插入某字母时，则可以在反查三元组集合中搜索存在哪些第一个字符为 h、第二个字符为 o 的三元组，找出来之后，直接将后两个字符交换顺序，即可作为依据插入操作得到的纠错结果。例如，发现反查三元组集合中存在“hou”、“hoi”等，后两个字母倒排后为“huo”、“hio”，因此，直接将“shuo”、“shio”等作为纠错结果即可，不用再尝试其他的字母，也不需要进行判断是否合法的操作。

[0089] 总之，在建立反查字符组集合时，可以采用如下方法 ：去除合法字符组中设定位置的至少一个字符，将其他剩余字符按设定顺序排列作为该被去除字符的索引 ；然后，以合法字符组集合的若干个索引及对应的被去除字符组成反查字符组集合。之后，在利用反查三元组获取纠错候选字符串时，可以如下进行 ：当需要对错误输入片段在前述设定位置进行字符的替换或插入时，根据该错误输入片段以前述设定顺序生成查询串，然后，从反查字符组集合中获取以该查询串为索引的字符，将该字符作为在前述设定位置替换或插入的字符，得到纠错候选字符串。

[0090] 需要说明的是，在生成纠错候选的过程中，涉及到的“预定位置”以及“预定顺序”，与建立反查字符组集合时涉及到的“预定位置”及“预定顺序”是相同的。例如，对于合法三元组“tub”，其设定位置的至少一个字符可以是指 ：“t”与“b”之间的一个字符“u”，将该字符 u 去除之后，剩余字符也就是“t”和“b”，假设“设定顺序”就是这两个字符在原合法三元组中的先后顺序，则按设定顺序排列之后就可以是“tb”，则该“tb”就可以作为字符“u”的索引。方便起见，可以将该索引与对应的被去除字符连接在一起成为“tbu”，则该“tbu”就成为一个反查三元组。当然，按照上述方式，根据其他合法三元组生成反查三元组时，“tb”还可能作为其他字符的索引。

[0091] 在进行生成纠错候选时，假设错误输入片段为“tb”，需要在“t”与“b”之间进行插入操作时，就可以将“tb”作为查询串，然后在反查字符组集合中获取以“tb”为索引的字符，例如包括前述例子中的字符“u”（还可能包括其他字符），然后，就可以将该字符“u”作为插入到“t”与“b”之间的字符，生成“tub”，将该“tub”替换原编码字符串中的“tb”，就可以得到纠错候选字符串。

[0092] 类似的，假设错误输入片段是“ttb”，如果需要将“t”与“b”之间进行替换操作时，此时，同样可以将“tb”作为查询串，然后在反查字符组集合中获取以“tb”为索引的字符，

例如包括前述例子中的字符“u”（还可能包括其他字符），然后，就可以将该字符“u”作为可以替换中间的“t”的字符，生成“tub”，将该“tub”替换原编码字符串中的“ttb”，就可以得到纠错候选字符串。

[0093] 可见，通过建立反查三元组，可以减少尝试操作的次数，缩小查找的范围，从而可以提高纠错的速度。需要说明的是，对于删除操作及交换操作，可以直接用正查三元组进行纠错即可。另外，在没有将正查三元组中的首字母参与倒排的情况下，如果需要在错误输入片段的首字母之前的插入操作、在错误输入片段的尾字母之后的插入操作，或者，在将错误输入片段的首字母或尾字母进行替换操作，则也可以直接依据正查三元组进行纠错。

[0094] 其中，关于正查三元组和反查三元组，可以直接以列表的方式进行保存，而在本发明实施例中，为了进一步提高查询的速度，可以以树形结构进行保存，下面进行详细地介绍。

[0095] 首先针对各个合法三元组，建立正查树，正查树可以共有 26 棵（当然，也可以在同一根节点下建立 26 棵子树），分别以 26 个字母之一为根节点，且每棵正查树中最多共有三层节点。例如，合法三元组中包括“zuu”，则在以“z”为根节点的正查树中，其第一级子节点可能有很多，其中会包括字母“u”，这个“u”的下一级子节点也可能有很多，其中会包括字母“u”。也就是说，正查树中同一个路径上的节点对应的字母能够组成合法三元组。

[0096] 在建立正查树之后，还可以建立反查树，相当于是将合法三元组中的各个字母进行倒排之后，重新组建树型结构。当然，如前文所述，由于第一个字母出错的可能性比较小，因此，第一个字母可以不必倒排，仅将后两个字母进行倒排即可。例如，合法三元组“ibu”，则将后两个字母倒排之后，就变为“iub”，合法三元组“zuz”，则将后两个字母进行倒排之后就变为“zzu”。在将所有的合法三元组进行倒排之后，就可以建立反查树。同样，反查树也可以有 26 棵（同样，也可以在同一根节点下建立 26 棵子树），每棵树分别以 26 个字母为根节点。

[0097] 需要说明的是，在第一个字母不参与倒排的情况下，正查三元组与对应的反查三元组的第一个字母是相同的，第一个字母相当于一个前缀，因此，正查树与反查树可以分别称为正查前缀树和反查前缀树。

[0098] 在建立了正查树和反查树之后，就可以利用正查树和反查树进行纠错。具体的过程可以与前文所述类似。例如，假设某错误输入片段为“xd”，在需要使用插入操作对其进行纠错时，就可以遍历反查前缀树，获知“xd”后存在“i”和“u”两个分支（没有分支表示对应的三元组不合法），与之对应的正查三元组即为“xid”(xd 中间插入 i) 和“xud”(xd 中间插入 u)。显然，这样做的好处是免除了对其它“i”、“u”以外的其他 24 个字母的枚举操作。

[0099] 需要说明的是，在前文所述的各个例子中，均是以合法全拼三元组为例进行介绍的，使用这种合法三元组进行检错及纠错时，如果用户也是习惯使用全拼的方式，则是比较适用的；但是有些用户可能习惯简拼，此时，如果使用基于合法全拼建立的合法三元组对其进行检错纠错就不合适了。因此，在实际应用中，建立合法三元组时，也不限于基于合法全拼来建立，也可以考虑简拼时的合法三元组。例如，有一部分用户在想要输入“什么”或“怎么”时，都习惯输入“sm”、“zm”，而基于合法全拼建立的合法二元组中，不包括“sm”、“zm”，此时就可能将其作为错误输入片段检测出来，显然这相当于是一种误判。而此时，如果基于合法简拼建立了合法二元组，并且其中包括“sm”、“zm”，则就不会将其作为错误输入片段被

检测出来，直接依据简拼词库给出相应的音字转换结果即可。

[0100] 另外，在建立合法三元组时，还可以给予海量训练数据获得三元组的可信度概率，这相对于只有合法或不合法两个结果的方式而言，更加有利于降低误判的可能。

[0101] 通过以上所述，介绍了本发明实施例提供的检错及纠错方法，在检错的过程中，通过使用分词的方法或合法字符组合的统计方法，可以使得检错的过程更加灵活，也能够更加全面地检测出编码字符串中存在的错误输入片段。在纠错的过程中，通过使用合法字符组合，进行字母的替换、插入、删除、交换操作，可以更加全面地获取纠错候选字符串。另外，通过反查字符组合的使用，可以缩小查找的范围，减少操作次数，从而提高纠错的效率。

[0102] 当然，利用以上方法获得的纠错候选字符串的数目可能会有很多，如果全部进行音字转换，则，工作量可能会比较大，并且得到的转换结果过多，也可能增加噪音，反而降低候选项的质量。因此，在本发明实施例中，在获得纠错候选字符串之后，还可以对各个纠错候选字符串进行评估，根据评估的结果来选择质量最高的一个或几个纠错候选字符串进行音字转换，并向用户提供转换结果。

[0103] 其中，具体在进行评估时，可以有多种方法。例如，其中一种方法可以是预先建立规则模型，根据规则模型进行评估及选择。如，可以是将音节数目最少的纠错候选字符串作为最终的纠错结果，或者，将对应转换结果的词频最高的纠错候选字符串作为最终的纠错结果等。此外，还可以是基于噪音信道模型进行评估，或者基于决策树模型进行评估，等等，这里不再一一列举。此外，在利用各种模型进行评估时，还可以同时考虑转换后的词条的系统词频、用户词频、用户的输入习惯等多方面的有效因素综合考虑，使得最终选出的纠错结果更加理想。

[0104] 另外，实际应用中的情况可能会是多种多样的，即便对于具有明显输入习惯的用户而言，也可能存在偶尔不按照习惯进行输入的时候。例如，某用户习惯于输入全拼，因此一般情况下，可以使用基于全拼的合法字符组对该用户输入的编码字符串进行检错及纠错。但是，该用户经常输入自己的住址“回龙观”，并且他知道输入“huilg”就能得到该候选项，因此，就会直接输入“huilg”。此时，如果直接利用基于全拼的合法字符组进行检错，则可能会发现其中存在错误输入片段，然后对其进行纠错时，可能就会得到很多其他的候选项，而不是“回龙观”。显然，这相当于是一种误判，不仅使得最终的候选项的质量下降，还白白浪费了检错及纠错过程中的计算资源。

[0105] 因此，为了降低造成误判的可能性，本发明实施例还可以这样进行：在对用户输入的编码字符串进行检错及纠错之前，首先判断该编码字符串是否命中设定词库，如包括用户词库、系统词库及细胞词库等的输入法词库，如果没有命中设定词库，再进行后续的检错及纠错操作。其中，在采用输入法词库时，可以优先匹配用户词库。当然，如果是为了丰富候选项，则即使编码字符串命中了设定词库，也是可以按照本发明实施例提供的方法进行检错及纠错的，此时既提供直接命中设定词库的候选项，又提供进行纠错后得到的候选项。

[0106] 在进行具体的字词转换结果的展现时，为了体现出针对纠错候选字符串的转换结果（简称纠错后的转换结果）与其他转换结果之间的区别，可以以相区别的方式进行展现。例如，可以在其他转换结果上以悬浮框的方式展现纠错后的转换结果，或者，将纠错后的转换结果显示为与其他的转换结果不同的颜色，等等。其中，其他转换结果是指依据用户实际输入的编码字符串本身转换得到的转换结果。

[0107] 需要说明的是，本发明实施例提供的字词输入方法可以应用于客户端，也可以应用于服务器，也即可以通过云计算的方式为用户提供字词候选。其中，当应用于客户端时，由于用户在使用输入法系统进行字词输入时，除了候选项的质量以外，系统内存的占用量也是体现输入法系统性能的很重要的因素，也即用户通常希望输入法系统在运行的过程中能够尽可能少地占用内存空间，以避免影响其他应用程序的运行。然而，如果要对用户输入的编码字符串进行纠错，并且还要进行打分排序等操作，则在提高候选项质量的同时，可能会损失部分内存占用量上的性能，并且打分排序时参考的依据越高、参考的模型越复杂，由于计算量的增大、复杂度的提高，内存的占用量可能就会越大（当然，如果客户端本地的计算机系统足够强大，这种内存空间上的占用可以忽略）。因此，在本发明实施例中，为了避免在过多的占用内存空间，还可以在对用户输入进行纠错的过程中引入云计算的概念，也即借助于远端服务器，以降低对客户端本地资源的依赖。

[0108] 为此，具体实现时，可以仅在客户端本地进行一些低复杂度的纠错处理，其他高复杂度的纠错处理可以通过云计算来实现。例如，对于一些纠错可信度较高、长度适中、较低阶数的模型就能完成的纠错任务，可以在客户端本地进行；而那些较复杂或较为不常用的纠错任务，则可以放到远端服务器进行，例如，用户输入错误中以模糊音为代表的认知错误，由于在形式上往往表现为合法的拼音串，例如“cifan（吃饭）”、“huiji（飞机）”、“wobuzidao（我不知道）”等等，因此，如果采用合法二元组、合法三元组等，就无法识别这种错误输入，只能采用其他的较为复杂、计算量偏大的方式来识别及评估，例如，可能需要采用强制纠错的方式，并且可能需要在每个位置尝试删除、交换以及基于所有字母的插入、替换操作，找出所有可能的纠错候选字符串，然后根据用户的输入习惯等等，对纠错候选字符串进行评估。这种情况下，就可以将用户输入的编码字符串发送到远端服务器，由远端服务器进行采用强制纠错等方式进行纠错，并对纠错候选结果评估后返回给客户端，由客户端进行统一的排序。

[0109] 需要说明的是，在本发明实施例中，均是以拼音输入法为例进行的介绍，但是，本发明实施例同样可以适用于五笔、笔画等其他的输入法。另外，由于输入法平台可以运行在多种计算设备上，例如，个人电脑、个人数字助理、移动终端设备等等，所以本发明实施例提供的方案也可以适用在上述各种计算设备中。并且，在上述各种计算设备中，可以具有全字母键盘，也可以是能够进行字符输入的数字键盘，或者触摸屏，等等。本发明实施例对编码字符串的构成也没有限制，可以是字母、数字、笔画等形式的一种或者几种的组合。

[0110] 与本发明实施例提供的字词输入方法相对应，本发明实施例还提供了一种字词输入系统，参见图 2，该系统包括：

[0111] 检错单元 201，用于检测用户输入的编码字符串中存在的错误输入片段；

[0112] 纠错单元 202，用于根据预先获取的合法字符组集合，编辑所述错误输入片段，得到纠错候选字符串；

[0113] 转换单元 203，用于将所述纠错候选字符串转换为输入文字。

[0114] 在本发明实施例中，检错的具体实现方式可以有多种，例如，在一种实现方式下，检错单元 201 可以包括：

[0115] 第一检错子单元，用于根据当前语境，对所述用户输入的编码字符串进行分词，将得到的分词碎片确定为错误输入片段。

[0116] 在另一种实现方式下，检错单元 201 可以包括：

[0117] 第二检错子单元，用于如果所述用户输入的编码字符串中存在不属于所述合法字符组集合的字符组，或者基于合法字符组集合统计的合法概率小于阈值的字符组，则将该字符组确定为错误输入片段。

[0118] 其中，为了适应用户连续输入一个词、短语或句子的编码字符串，并且不会主动在各个字的编码字符串之间输入分隔符的情况，在获取合法字符组集合时，可以通过以下方式获得：从至少两个文字的合法编码字符串连接而成的字符串中抽取合法字符组。

[0119] 具体实现时，纠错单元 202 可以包括：

[0120] 编辑子单元，用于对所述错误输入片段分别进行基于字符的替换、插入、删除及交换的处理；

[0121] 生成子单元，用于如果处理后的片段属于所述合法字符组集合，则基于该处理后的片段生成纠错候选字符串。

[0122] 由于在进行替换及插入操作时，如果采用将每个字母在每个位置都尝试以便的话，会使得工作量非常大，为此，本发明实施例中可以采用建立反查字符组的方式，来简化上述两种编辑操作。此时，纠错单元 202 具体可以包括：

[0123] 去除子单元，用于去除合法字符组中设定位置的至少一个字符，将其他剩余字符按设定顺序排列作为所述被去除字符的索引；

[0124] 组合子单元，用于以合法字符组集合的若干个所述索引及对应的被去除字符组成反查字符组集合；

[0125] 查询串生成子单元，用于当需要对所述错误输入片段在所述设定位置进行字符的替换或插入时，根据所述错误输入片段以所述设定顺序生成查询串；

[0126] 查询子单元，用于从所述反查字符组集合中获取以所述查询串为索引的字符，将该字符作为在所述设定位置替换或插入的字符，得到纠错候选字符串。

[0127] 其中，由于第一个字母就输错的可能性较小，因此，建立反查字符组时的设定位置可以是合法字符组中首字符位置以外的其他位置。

[0128] 为了更加便于查询，提高生产纠错候选字符串的效率，合法字符组集合及反查字符组集合可以以树形结构进行保存，也即生产正查树及反查树。

[0129] 根据本发明实施例的方法生成纠错候选字符串时，可能会生成多个，如果将这么多的纠错候选字符串不加区分地全部进行转换，则候选项的数量可能会非常多，质量也会下降，因此，本发明实施例中，还可以对得到的纠错候选字符串进行评估，根据评估结果来选择对哪个或哪些纠错候选字符串进行转换。此时，转换单元 203 可以包括：

[0130] 评估子单元，用于对所述纠错候选字符串进行评估；

[0131] 选择转换子单元，用于根据评估的结果，对所述编码字符串及符合预置条件的纠错候选字符串进行转换，并向用户展现转换的结果。

[0132] 本发明实施例提供的字词输入系统可以应用于客户端，也可以应用于服务器。当应用于客户端时，由于检错及纠错过程可能会耗费较多的计算资源，对计算机性能的要求可能会比较高。为了降低这种对计算机性能的要求，本发明实施例可以采用客户端本地与云计算相结合的方式实现。即，对于一些简单的检错及纠错的情况，可以在客户端本地进行；而对于复杂的情况，则可以将用户输入的编码字符串发送到远端服务器，通过云计算的

方式，获取纠错候选字符串。此时，该系统还包括：

[0133] 云计算单元，用于将所述用户输入的编码字符串发送到远端服务器，并接收所述远端服务器返回的纠错候选字符串。

[0134] 当然，由于实际应用中的情况可能是多种多样的，因此，为了进一步提高候选项的质量，在进行检错及纠错之前或过程中，还可以借助于设定词库（如输入法词库），来判断是否需要进行纠错，或者是否需要进行强制检错，等等。此时，该系统还可以包括：

[0135] 词库匹配单元，用于如果所述用户输入的编码字符串命中设定词库，根据所述设定词库将所述编码字符串转换为输入文字。

[0136] 总之，通过本发明实施例提供的字词输入系统，在用户进行字词输入的过程中，在发现编码字符串中存在的错误输入片段之后，可以根据预先获取的合法字符组集合，通过对错误输入片段进行编辑操作，得到纠错候选字符串，这样，获取纠错候选字符串的方式更加灵活，进而在根据用户输入的编码字符串以及纠错候选字符串进行音字转换时，也可以获得更多可能的候选项。因此，该方法更加有效，且适用范围比较广泛。另外，本发明实施例在对所述错误输入片段进行编辑操作时，使用了反查字符组，这样，针对一些替换或插入操作而言，可以缩小查找范围，从而减少操作次数，提高纠错的效率。

[0137] 以上对本发明所提供的一种字词输入方法及系统，进行了详细介绍，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。

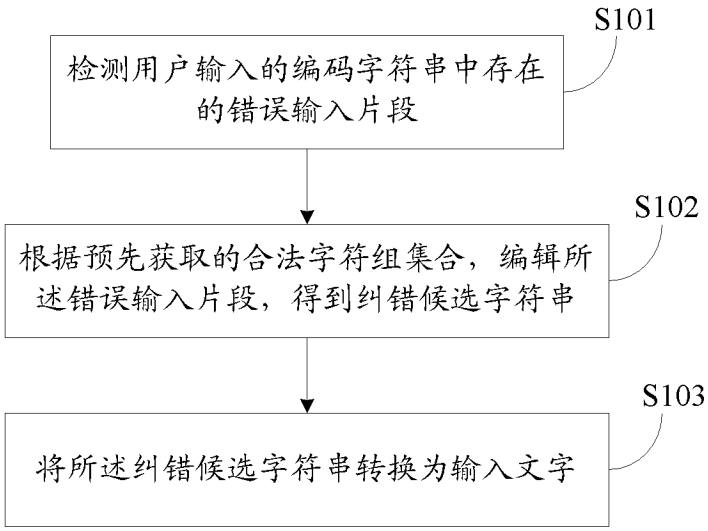

S101
检测用户输入的编码字符串中存在的错误输入片段
S102
根据预先获取的合法字符组集合，编辑所述错误输入片段，得到纠错候选字符串
S103
将所述纠错候选字符串转换为输入文字

图 1

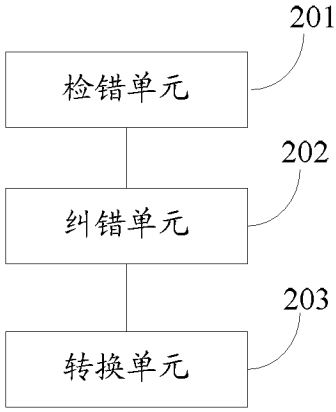

201
检错单元
202
纠错单元
203
转换单元

图 2