

一种对词库进行处理的方法及装置

本申请要求于 2018 年 5 月 25 日提交的申请号为 201810515382.X、发明名称为“一种对词库进行处理的方法及装置”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

本申请涉及计算机领域，特别涉及一种对词库进行处理的方法及装置。

背景技术

日语包括汉字和假名，汉字的读音可以使用至一个假名来标注。例如，对于汉字“身体”，可以使用“しんたい”标注其读音，即“身体”的读音可以为“しんたい”。在日语文章或歌词等场景中常常在汉字上方标注其假名，例如“身体”，这样可以方便读者阅读。

目前存在用于给日语文本标注假名的工具，该工具中存在一个词库，该词库中包括多个数据记录，每个数据记录中包括一个词条和该词条中的各汉字对应的至少一个假名等。在标注假名时，可以使该工具读取待标注文本中的词条，根据该词条从该词库中找出该词条中的各汉字对应的假名，在日语文本中该词条包括的各汉字上方标注找出的假名。

发明人在实现本申请的过程中，发现上述方式至少存在如下缺陷：

目前词库的容量都比较大，词库中包括大量的数据记录，在标注假名时需要较长时间才能从词库中获取词条中的汉字对应的假名，降低了标注假名的效率。

发明内容

为了提高标注假名的效率，本申请实施例提供了一种对词库进行处理的方法及装置。所述技术方案如下：

第一方面，本申请实施例提供了一种对词库进行处理的方法，所述方法包

括:

获取第一词库中的第一数据记录,第一数据记录包括一个多词条和所述多词条中的每个汉字对应的第一假名集合,所述多词条为包括多个汉字的词条,汉字对应的第一假名集合包括所述汉字对应的至少一个假名;

在第二词库中查找所述第一数据记录对应的多个目标数据记录,每个目标数据记录中的目标词条是所述多词条的不同组成部分,所述每个目标数据记录中的目标词条组成所述多词条,目标数据记录中的目标词条中的每个汉字对应的第二假名集合分别与所述每个汉字对应的第一假名集合相同;

当在所述第二词库中没有查找出所述第一数据记录对应的多个目标数据记录,则将所述第一数据记录保存在所述第二词库中。

可选的,所述在第二词库中查找所述第一数据记录对应的多个目标数据记录,包括:

将所述第一数据记录中的多词条划分成N个单词条,N大于1的整数,所述单词条为包括一个汉字的词条;

在第二词库中查找所述每个单词条对应的目标数据记录,单词条对应的目标数据记录包括所述单词条和所述单词条中的汉字对应的第二假名集合,所述汉字对应的第二假名集合与所述汉字对应的第一假名集合相同;

当在所述第二词库中没有查找出所述每个单词条对应的目标数据记录,则确定从所述第二词库中查找不出所述第一数据记录对应的多个目标数据记录。

可选的,所述获取第一词库中的第一数据记录之前,还包括:

将所述第一词库中的各第二数据记录保存在所述第二词库中,所述第二数据记录包括一个单词条和所述单词条中的汉字对应的第一假名集合。

可选的,所述将所述第一词库中的各第二数据记录保存在所述第二词库中,包括:

获取所述第一词库中的任一条数据记录;

在所述数据记录中包括一个第一假名集合时,确定所述数据记录中的词条为单词条并将所述数据记录保存在所述第二词库中。

可选的，所述获取第一词库中的第一数据记录之前，还包括：

将第三词库中包括预设应用场景的数据记录保存在第一词库中，所述第三词库中的每个数据记录包括一个词条、所述词条中的每个汉字对应的假名集合和应用场景。

可选的，所述将第三词库中包括预设应用场景的数据记录保存在第一词库中，包括：

从第三词库中获取包括预设应用场景的数据记录，所述数据记录包括一个词条、所述第一词条中的每个汉字对应的至少一个假名集合、所述至少一个假名集合中的每个假名集合的使用频率和所述预设应用场景；

根据所述每个汉字对应的至少一个假名集合的使用频率，分别从所述每个汉字对应的至少一个假名集合中选择所述每个汉字对应的假名集合；

将所述一个词条、为所述每个汉字选择假名集合组成第一数据记录并保存在所述第一词库中。

第二方面，本申请实施例提供了一种对词库进行处理的装置，所述装置包括：

获取模块，用于获取第一词库中的第一数据记录，第一数据记录包括一个多词条和所述多词条中的每个汉字对应的第一假名集合，所述多词条为包括多个汉字的词条，汉字对应的第一假名集合包括所述汉字对应的至少一个假名；

查找模块，用于在第二词库中查找所述第一数据记录对应的多个目标数据记录，每个目标数据记录中的目标词条是所述多词条的不同组成部分，所述每个目标数据记录中的目标词条组成所述多词条，目标数据记录中的目标词条中的每个汉字对应的第二假名集合分别与所述每个汉字对应的第一假名集合相同；

保存模块，用于当在所述第二词库中没有查找出所述第一数据记录对应的多个目标数据记录，则将所述第一数据记录保存在所述第二词库中。

可选的，所述查找模块包括：

划分单元，用于将所述第一数据记录中的多词条划分成N个单词条，N大于1的整数，所述单词条为包括一个汉字的词条；

查找单元，用于在第二词库中查找所述每个单词条对应的目标数据记录，单词条对应的目标数据记录包括所述单词条和所述单词条中的汉字对应的第二假名集合，所述汉字对应的第二假名集合所述汉字对应的第一假名集合相同；

第一确定单元，用于当在所述第二词库中没有查找出所述每个单词条对应的目标数据记录，则确定从所述第二词库中查找不出所述第一数据记录对应的多个目标数据记录。

可选的，所述保存模块，还用于：

将所述第一词库中的各第二数据记录保存在所述第二词库中，所述第二数据记录包括一个单词条和所述单词条中的汉字对应的第一假名集合。

可选的，所述保存模块包括：

获取单元，用于获取所述第一词库中的任一条数据记录；

第二确定单元，用于在所述数据记录中包括一个第一假名集合时，确定所述数据记录中的词条为单词条并将所述数据记录保存在所述第二词库中。

可选的，所述保存模块，还用于：

将第三词库中包括预设应用场景的数据记录保存在第一词库中，所述第三词库中的每个数据记录包括一个词条、所述词条中的每个汉字对应的假名集合和应用场景。

可选的，所述保存模块还包括：

第二获取单元，用于从第三词库中获取包括预设应用场景的数据记录，所述数据记录包括一个词条、所述第一词条中的每个汉字对应的至少一个假名集合、所述至少一个假名集合中的每个假名集合的使用频率和所述预设应用场景；

选择单元，用于根据所述每个汉字对应的至少一个假名集合的使用频率，分别从所述每个汉字对应的至少一个假名集合中选择所述每个汉字对应的假名集合；

保存单元，用于将所述一个词条、为所述每个汉字选择假名集合组成第一

数据记录并保存在所述第一词库中。

第三方面，本申请实施例提供了一种非易失性计算机可读存储介质，用于存储计算机程序，所述计算机程序通过处理器进行加载来执行第一方面或第一方面的任意可选的实现方式的方法的指令。

本申请实施例提供的技术方案可以包括以下有益效果：

对于第一词库中的任一条第一数据记录，如果第二词库中存在多条目标数据记录，每个目标数据记录中的目标词条中的汉字对应的第二假名集合与第一数据记录中的该汉字对应的第一假名集合相同时，则不将该第一数据记录保存在第二词库中；如果第二词库中不存在该多条目标数据记录，才将第一数据记录保存在第二词库中，这样精简了第二词库中的数据记录的数目，减小了第二词库的容量。由于第二词库的容量较小，所以使用第二词库标注日文文本中的各汉字的假名时，可以提高标注假名的效率。

应当理解的是，以上的一般描述和后文的细节描述仅是示例性和解释性的，并不能限制本申请。

附图说明

此处的附图被并入说明书中并构成本说明书的一部分，示出了符合本申请的实施例，并与说明书一起用于解释本申请的原理。

图 1 是本申请实施例提供的一种对词库进行处理的方法流程图；

图 2 是本申请实施例提供的另一种对词库进行处理的方法流程图；

图 3-1 是本申请实施例提供的一种标注假名的方法流程图；

图 3-2 是本申请实施例提供的日文文本的示意图；

图 4 是本申请实施例提供的一种对词库进行处理的装置结构示意图；

图 5 是本申请实施例提供的一种终端结构示意图。

通过上述附图，已示出本申请明确的实施例，后文中将有更详细的描述。这些附图和文字描述并不是为了通过任何方式限制本申请构思的范围，而是通过参考特定实施例为本领域技术人员说明本申请的概念。

具体实施方式

这里将详细地对示例性实施例进行说明，其示例表示在附图中。下面的描述涉及附图时，除非另有表示，不同附图中的相同数字表示相同或相似的要素。以下示例性实施例中所描述的实施方式并不代表与本发明相一致的所有实施方式。相反，它们仅是与如所附权利要求书中所详述的、本发明的一些方面相一致的装置和方法的例子。

参见图 1，本发明实施例提供了一种对词库进行处理的方法，所述方法包括：

步骤 101：获取第一词库中的第一数据记录，第一数据记录包括一个多词条和该多词条中的每个汉字对应的第一假名集合，该多词条为包括多个汉字的词条，汉字对应的第一假名集合包括该汉字对应的至少一个假名。

步骤 102：在第二词库中查找第一数据记录对应的多个目标数据记录，每个目标数据记录中的目标词条是该多词条的不同组成部分，每个目标数据记录中的目标词条组成该多词条，目标数据记录中的目标词条中的每个汉字对应的第二假名集合分别与该每个汉字对应的第一假名集合相同。

步骤 103：当在第二词库中没有查找出第一数据记录对应的多个目标数据记录，则将第一数据记录保存在第二词库中。

在本发明实施例中，对于第一词库中的任一条第一数据记录，如果第二词库中存在多条目标数据记录，每个目标数据记录中的目标词条中的汉字对应的第二假名集合与第一数据记录中的该汉字对应的第一假名集合相同时，则不将该第一数据记录保存在第二词库中；如果第二词库中不存在该多条目标数据记录，才将第一数据记录保存在第二词库中，这样精简了第二词库中的数据记录的数目，减小了第二词库的容量，使用第二词库标注日文文本中的各汉字的假名时，可以提高标注假名的效率。

参见图 2，本发明实施例提供了一种对词库进行处理的方法，所述方法用于对第一词库进行精简得到第二词库，第二词库包括的数据记录数目小于第一词库包括的数据记录数目，所述方法包括：

步骤 201：将第一词库中的各第二数据记录保存在第二词库中，第二数据记录包括一个单词条和该单词条中的汉字对应的第一假名集合，该单词条为包括一个汉字的词条。

可选的，在执行本步骤之前，还可以创建空白的第二词库。

可选的，在执行本步骤之前，还可以获取第一词库。获取第一词库的方式有多种，在本步骤中列举了获取第一词库的方式，分别为：

第一种方式，第一词库可以是用于标注假名的工具包括的词库，在本步骤可以获取某个工具包括的词库作为第一词库。该工具可以用于标注某一应用场景的日语文本中的汉字的假名。例如，应用场景可以为新闻或歌词等应用场景，对于新闻应用场景，使用新闻应用场景对应的第一工具，第一工具专用于标注新闻文稿中的汉字的假名；对于歌词应用场景，使用歌词应用场景对应的第二工具，第二工具专用于标注歌词文本中的汉字的假名。即在本步骤中可以获取第一工具包括的词库作为第一词库或者获取第二工具包括的词库作为第一词库。

由于日语中同一汉字在不同应用场景中该汉字的读音不同，即在不同应用场景中给该汉字标注的假名不同，所以第一工具的词库和第二工具的词库不同。

在本步骤中，除了采用第一种方式获取第一词库外，还可以采用其他方式获取第一词库。例如，接下来还列举了第二种方式，在第二种方式中，将第三词库中包括预设应用场景的数据记录保存在第一词库中。

其中，第三词库是个通用词库，可以应用于不同的应用场景，第三词库中的每个数据记录包括一个词条、该词条中的每个汉字对应的假名集合和应用场景。汉字对应的假名集合包括该汉字对应的至少一个假名。

可选的，对于第三词库中的任一条数据记录，对于该条数据记录中的每个汉字，该条数据记录中可能包括该汉字对应的一个假名集合，也可能包括该汉字对应的多个假名集合，即该条数据记录中的每个汉字对应至少一个假名集合，该条数据记录中还可以包括每个假名集合对应的使用频率，还可以包括该数据记录中的词条的词性等信息。

可选的，在从第三词库中获取到一条包括预设应用场景的数据记录时，根据该数据记录中的每个汉字对应的至少一个假名集合的使用频率，分别从每个汉字对应的至少一个假名集合中选择每个汉字对应的假名集合，将为每个汉字选择的假名集合分别作为每个汉字对应的第一假名集合，将该数据记录中的词条，为该词条中的每个汉字选择的第一假名集合和该数据记录中的该词条的词性等信息组成第一数据记录并保存在第一词库中。

可选的，对于每个汉字对应的至少一个假名集合，可以从该汉字对应的至少一个假名集合中选择使用频率最大的假名集合作为该汉字对应的一个第一假名集合。

可选的，第一词库中的每条数据记录包括一个词条和该词条中的每个汉字对应的第一假名集合，还可以包括该词条的词性等信息。

例如，参见下表 1 所示的第一词库，表 1 所示的第一词库中的第一条数据记录中包括词条“思い出す”，该词条中的汉字“思”对应的第一假名集合（おも），该词条“出”对应的第一假名集合（だ），以及该词条的词性为“五段动词”。表 1 所示的第一词库中的第二条数据记录中包括词条“思い”，该词条中的汉字“思”对应的第一假名集合（おも），以及该词条的词性为“五段动词”。表 1 所示的第一词库中的第三条数据记录中包括词条“出す”，该词条中的汉字“出”对应的第一假名集合（だ），以及该词条的词性为“五段动词”。

表 1：第一词库

词条	第一假名集合	词性
思い出す	（おも）、（だ）	五段动词
思い	（おも）	五段动词
出す	（だ）	五段动词
.....		

本步骤可以为：获取第一词库中的任一条数据记录；如果该数据记录中包括一个第一假名集合时，确定该数据记录中的词条为包括一个汉字的单词条并将该数据记录作为第二数据记录并保存在第二词库中。按上述方式获取第一第一词库中的每条第二数据记录保存在第二词库中。

例如，获取表 1 所示的第一词库中的第一条数据记录，第一条数据记录中包括两个第一假名集合，分别为（おも）和（だ），所以第一条数据记录中的词条“思い出す”包括两个汉字，并不是包括一个汉字的单词条。

获取表 1 所示的第一词库中的第二条数据记录，第二条数据记录中包括一个第一假名集合（おも），所以第二条数据记录中的词条“思い”包括一个汉字，是包括一个汉字的单词条，可以将第二条数据记录作为第二数据记录并保存在如表 2 所示的第二词库中。

获取表 1 所示的第一词库中的第三条数据记录，第三条数据记录中包括一

个第一假名集合 (だ), 所以第三条数据记录中的词条 “出す” 包括一个汉字, 是包括一个汉字的单词条, 可以将第三条数据记录作为第二数据记录并保存在如表 2 所示的第二词库中。

表 2: 第二词库

词条	第二假名集合	词性
思い	(おも)	五段动词
出す	(だ)	五段动词
思	(おも)	五段动词
.....		

步骤 202: 获取第一词库中的第一数据记录, 第一数据记录包括一个多词条和该多词条中的每个汉字对应的第一假名集合, 该多词条为包括多个汉字的词条, 汉字对应的第一假名集合包括该汉字对应的至少一个假名。

第一数据记录是第一词库中除第二数据记录以外的其他数据记录, 所以本步骤可以, 从第一词库中除第二数据记录以外的其他数据记录中获取一条数据记录作为第一数据记录。

例如, 在表 1 所示的第一词库中第二条数据记录和第三条数据记录为包括单词条的第二数据记录, 所以从表 1 所示的第一词库中除该两条第二数据记录以外的数据记录中获取一条数据记录, 假设获取第一条数据记录并作为第一数据记录, 该第一数据记录包括多词条 “思い出す”, 第一假名集合 (おも) 和 (だ), 以及该多词条的词性为 “五段动词”。

步骤 203: 查找第二词库中是否存在第一数据记录对应的多个目标数据记录, 每个目标数据记录中的目标词条是该多词条的不同组成部分, 每个目标数据记录中的目标词条组成该多词条, 目标数据记录中的目标词条中的每个汉字对应的第二假名集合分别与该每个汉字对应的第一假名集合相同。

可选的, 在本步骤中可以通过如下 2031 至的步骤实现, 分别为:

2031: 将第一数据记录中的多词条划分成 N 个单词条, N 大于 1 的整数。

可选的, 在本步骤中可以将该多词条划分成 N 个包括一个汉字的单词条。例如, 可以将第一数据记录包括多词条 “思い出す” 划分成两个单词条, 分别为 “思い” 和 “出す”。

2032: 查找第二词库中是否存在每个单词条对应的目标数据记录, 单词条

对应的目标数据记录包括该单词条和该单词条中的汉字对应的第二假名集合，该汉字对应的第二假名集合与该汉字对应的第一假名集合相同。

可选的，对于该 N 个单词条中的任一个单词条，从第二词库中查找是否包括该单词条的数据记录。如果查找出该数据记录，判断在该数据记录中的该单词条中的汉字对应的第二假名集合是否与第一数据记录中的该单词条中的汉字对应的第一假名集合相同；如果判断出该单词条中的汉字对应的第二假名集合与该汉字对应的第一假名集合相同，则将该数据记录作为该单词条对应的目标数据记录，否则，确定出第二词库中不存在该单词条对应的目标数据记录。

例如，对于单词条“思い”，从表 2 所示的第二词库中查找包括单词条“思い”的第一条数据记录，该第一条数据记录中包括该单词条“思い”、第二假名集合（おも）和词性“五段动词”。在该第一条数据记录中该单词条“思い”中的汉字“思”对应的第二假名集合（おも）和在第一数据记录中该单词条“思い”中的汉字“思”对应的第一假名集合（おも）相同，所以该第一条数据记录为该单词条“思い”对应的目标数据记录。

再例如，对于单词条“出す”，从表 2 所示的第二词库中查找包括单词条“出す”的第二条数据记录，该第二条数据记录中包括该单词条“出す”、第二假名集合（だ）和词性“五段动词”。在该第二条数据记录中该单词条“出す”中的汉字“出”对应的第二假名集合（だ）和在第一数据记录中该单词条“出す”中的汉字“出”对应的第一假名集合（だ）相同，所以该第二条数据记录为该单词条“出す”对应的目标数据记录。

2033: 如果第二词库中存在每个单词条对应的目标数据记录，则确定第二词库中存在第一数据记录对应的多个目标数据记录。

例如，对于该多词条“思い出す”划分成的两个单词条，该两个单词条均有对应的目标数据记录，从而可以确定第二词库中存在第一数据记录对应的多个目标数据记录。

可选的，如果第二词库中不存在每个单词条对应的目标数据记录，则可以确定第二词库中不存在第一数据记录对应的多个目标数据记录。

如果确定第二词库中存在第一数据记录对应的多个目标数据记录，则不需要将该第一数据记录添加到第二词库中，即在第二词库中可以使用该多个目标数据记录代替第一数据记录，从而减少第二词库中的数据记录数目。

步骤 204: 如果第二词库不存在该多个目标数据记录，则将第一数据记录

保存在第二词库中。

在本申请实施例中，对于第一词库中的任一条第一数据记录，如果第二词库中存在多条目标数据记录，每个目标数据记录中的目标词条中的汉字对应的第二假名集合与第一数据记录中的该汉字对应的第一假名集合相同时，则不将该第一数据记录保存在第二词库中；如果第二词库中不存在该多条目标数据记录，才将第一数据记录保存在第二词库中，这样精简了第二词库中的数据记录的数目，减小了第二词库的容量，使用第二词库标注日文文本中的各汉字的假名时，可以提高标注假名的效率。由于第二词库的容量较小，可以将第二词库安装在移动终端上，可以在移动终端上标注日文文本中的各汉字的假名。

参见图 3-1，本申请实施例提供了一种标注假名的方法，该方法可以使用上述任一实施例得到的第二词库标注日文文本中的汉字的假名，该方法包括：

步骤 301：获取日文文本中的第一个汉字，将该第一个汉字作为待处理词条。

例如，对于日文文本“思い出す”，将第一个汉字“思”作为待处理词条。

步骤 302：查找第二词库中是否存在包括待处理词条的数据记录，如果存在，执行 303，如果不存在，则执行 305。

查找如表 2 所示的第二词库是否存在包括待处理词条“思”的数据记录，查找出的结果为表 2 中存在第三条数据记录包括待处理词条“思”。

步骤 303：将该数据记录中的待处理词条中的每个汉字对应的假名集合中的假名分别确定为待处理词条中的每个汉字对应的假名。

例如，将第三条数据记录中的待处理词条中的汉字“思”对应的假名集合（おも）中的假名确定为汉字“思”对应的假名，即汉字“思”对应的假名为おも。

步骤 304：获取该日文文本中下一个未获取的文字，将该文字与待处理词条组成一个新的待处理词条，返回执行步骤 302。

从该日文文本中“思い出す”获取下一个未获取的文字“い”，将待处理词条“思”和获取的文字“い”组成新的待处理词条“思い”。查找如表 2 所示的第二词库是否存在包括待处理词条“思い”的数据记录，查找出的结果为表 2 中存在第一条数据记录包括待处理词条“思い”。将第一条数据记录中的待处理词条中的汉字“思い”中的汉字“思”对应的假名集合（おも）中的假名确定为汉

字“思”对应的假名，即汉字“思”对应的假名为お和も。

从该日文文本中“思い出す”获取下一个未获取的文字“出”，将待处理词条“思い”和获取的文字“出”组成新的待处理词条“思い出”。查找如表 2 所示的第二词库是否存在包括待处理词条“思い出”的数据记录，查找出的结果为表 2 中不存在包括待处理词条“思い出”的数据记录，执行步骤 305。

步骤 305: 在日文文本中使用待处理词条中的汉字对应的假名标注该汉字。

例如，待处理词条中“思い出”中只有汉字“思”有对应的假名，分别为お和も，因此可以在日文文本“思い出す”中，使用汉字“思”对应的假名标注汉字“思”，得到的结果为如图 3-2 所示。

步骤 306: 如果该日文文本中还存在未标注假名的汉字，则从未标注假名的汉字中获取第一个汉字，将该第一个汉字作为待处理词条，执行步骤 302。

如果该日文文本中不存在未标注假名的汉字，则结束操作，实现给该日文文本中的每个汉字标注假名。

下述为本申请装置实施例，可以用于执行本申请方法实施例。对于本申请装置实施例中未披露的细节，请参照本申请方法实施例。

参见图 4，本申请实施例提供了一种对词库进行处理的装置 400，所述装置 400 包括：

获取模块 401，用于获取第一词库中的第一数据记录，第一数据记录包括一个多词条和所述多词条中的每个汉字对应的第一假名集合，所述多词条为包括多个汉字的词条，汉字对应的第一假名集合包括所述汉字对应的至少一个假名；

查找模块 402，用于在第二词库中查找第二词库中是否存在所述第一数据记录对应的多个目标数据记录，每个目标数据记录中的目标词条是所述多词条的不同组成部分，所述每个目标数据记录中的目标词条组成所述多词条，目标数据记录中的目标词条中的每个汉字对应的第二假名集合分别与所述每个汉字对应的第一假名集合相同；

保存模块 403，用于当在所述第二词库中没有查找出所述第一数据记录对应的多个目标数据记录，则将所述第一数据记录保存在所述第二词库中。

可选的，所述查找模块 402 包括：

划分单元，用于将所述第一数据记录中的多词条划分成N个单词条，N大于1的整数，所述单词条为包括一个汉字的词条；

查找单元，用于在第二词库中查找所述每个单词条对应的目标数据记录，单词条对应的目标数据记录包括所述单词条和所述单词条中的汉字对应的第二假名集合，所述汉字对应的第二假名集合与所述汉字对应的第一假名集合相同；

第一确定单元，用于当在所述第二词库中没有查找出所述每个单词条对应的目标数据记录，则确定从所述第二词库中查找不出所述第一数据记录对应的多个目标数据记录。

可选的，所述保存模块403，还用于：

将所述第一词库中的各第二数据记录保存在所述第二词库中，所述第二数据记录包括一个单词条和所述单词条中的汉字对应的第一假名集合。

可选的，所述保存模块403包括：

获取单元，用于获取所述第一词库中的任一条数据记录；

第二确定单元，用于在所述数据记录中包括一个第一假名集合时，确定所述数据记录中的词条为单词条并将所述数据记录保存在所述第二词库中。

可选的，所述保存模块403，还用于：

将第三词库中包括预设应用场景的数据记录保存在第一词库中，所述第三词库中的每个数据记录包括一个词条、所述词条中的每个汉字对应的假名集合和应用场景。

可选的，所述保存模块403还包括：

第二获取单元，用于从第三词库中获取包括预设应用场景的数据记录，所述数据记录包括一个词条、所述第一词条中的每个汉字对应的至少一个假名集合、所述至少一个假名集合中的每个假名集合的使用频率和所述预设应用场景；

选择单元，用于根据所述每个汉字对应的至少一个假名集合的使用频率，分别从所述每个汉字对应的至少一个假名集合中选择所述每个汉字对应的假名集合；

保存单元，用于将所述一个词条、为所述每个汉字选择假名集合组成第一数据记录并保存在所述第一词库中。

在本申请实施例中，对于第一词库中的任一条第一数据记录，如果第二词

库中存在多条目标数据记录，每个目标数据记录中的目标词条中的汉字对应的第二假名集合与第一数据记录中的该汉字对应的第一假名集合相同时，则不将该第一数据记录保存在第二词库中；如果第二词库中不存在该多条目标数据记录，才将第一数据记录保存在第二词库中，这样精简了第二词库中的数据记录的数目，减小了第二词库的容量，使用第二词库标注日文文本中的各汉字的假名时，可以提高标注假名的效率。

关于上述实施例中的装置，其中各个模块执行操作的具体方式已经在有关该方法的实施例中进行了详细描述，此处将不做详细阐述说明。

图5示出了本发明一个示例性实施例提供的终端500的结构框图。该终端500可以是便携式移动终端，比如：智能手机、平板电脑、MP3播放器（Moving Picture Experts Group Audio Layer III，动态影像专家压缩标准音频层面3）、MP4（Moving Picture Experts Group Audio Layer IV，动态影像专家压缩标准音频层面4）播放器、笔记本电脑或台式电脑。终端500还可能被称为用户设备、便携式终端、膝上型终端、台式终端等其他名称。

通常，终端500包括有：处理器501和存储器502。

处理器501可以包括一个或多个处理核心，比如4核心处理器、8核心处理器等。处理器501可以采用DSP（Digital Signal Processing，数字信号处理）、FPGA（Field-Programmable Gate Array，现场可编程门阵列）、PLA（Programmable Logic Array，可编程逻辑阵列）中的至少一种硬件形式来实现。处理器501也可以包括主处理器和协处理器，主处理器是用于对在唤醒状态下的数据进行处理的处理器，也称CPU（Central Processing Unit，中央处理器）；协处理器是用于对在待机状态下的数据进行处理的低功耗处理器。在一些实施例中，处理器501可以在集成有GPU（Graphics Processing Unit，图像处理器），GPU用于负责显示屏所需要显示的内容的渲染和绘制。一些实施例中，处理器501还可以包括AI（Artificial Intelligence，人工智能）处理器，该AI处理器用于处理有关机器学习的计算操作。

存储器502可以包括一个或多个计算机可读存储介质，该计算机可读存储介质可以是非暂态的。存储器502还可包括高速随机存取存储器，以及非易失性存储器，比如一个或多个磁盘存储设备、闪存存储设备。在一些实施例中，

存储器 502 中的非暂态的计算机可读存储介质用于存储至少一个指令，该至少一个指令用于被处理器 501 所执行以实现上述任一实施例提供的方法。

在一些实施例中，终端 500 还可选包括有：外围设备接口 503 和至少一个外围设备。处理器 501、存储器 502 和外围设备接口 503 之间可以通过总线或信号线相连。各个外围设备可以通过总线、信号线或电路板与外围设备接口 503 相连。具体地，外围设备包括：射频电路 504、触摸显示屏 505、摄像头 506、音频电路 507、定位组件 508 和电源 509 中的至少一种。

外围设备接口 503 可被用于将 I/O (Input /Output, 输入/输出) 相关的至少一个外围设备连接到处理器 501 和存储器 502。在一些实施例中，处理器 501、存储器 502 和外围设备接口 503 被集成在同一芯片或电路板上；在一些其他实施例中，处理器 501、存储器 502 和外围设备接口 503 中的任意一个或两个可以在单独的芯片或电路板上实现，本实施例对此不加以限定。

射频电路 504 用于接收和发射 RF (Radio Frequency, 射频) 信号，也称电磁信号。射频电路 504 通过电磁信号与通信网络以及其他通信设备进行通信。射频电路 504 将电信号转换为电磁信号进行发送，或者，将接收到的电磁信号转换为电信号。可选地，射频电路 504 包括：天线系统、RF 收发器、一个或多个放大器、调谐器、振荡器、数字信号处理器、编解码芯片组、用户身份模块卡等等。射频电路 504 可以通过至少一种无线通信协议来与其它终端进行通信。该无线通信协议包括但不限于：万维网、城域网、内联网、各代移动通信网络 (2G、3G、4G 及 5G)、无线局域网和/或 WiFi(Wireless Fidelity, 无线保真) 网络。在一些实施例中，射频电路 504 还可以包括 NFC (Near Field Communication, 近距离无线通信) 有关的电路，本申请对此不加以限定。

显示屏 505 用于显示 UI (User Interface, 用户界面)。该 UI 可以包括图形、文本、图标、视频及其它们的任意组合。当显示屏 505 是触摸显示屏时，显示屏 505 还具有采集在显示屏 505 的表面或表面上方的触摸信号的能力。该触摸信号可以作为控制信号输入至处理器 501 进行处理。此时，显示屏 505 还可以用于提供虚拟按钮和/或虚拟键盘，也称软按钮和/或软键盘。在一些实施例中，显示屏 505 可以为一个，设置终端 500 的前面板；在另一些实施例中，显示屏 505 可以为至少两个，分别设置在终端 500 的不同表面或呈折叠设计；在再一些实施例中，显示屏 505 可以是柔性显示屏，设置在终端 500 的弯曲表面上或折叠面上。甚至，显示屏 505 还可以设置成非矩形的不规则图形，也即异形屏。

显示屏 505 可以采用 LCD(Liquid Crystal Display, 液晶显示屏)、OLED(Organic Light-Emitting Diode, 有机发光二极管)等材质制备。

摄像头组件 506 用于采集图像或视频。可选地, 摄像头组件 506 包括前置摄像头和后置摄像头。通常, 前置摄像头设置在终端的前面板, 后置摄像头设置在终端的背面。在一些实施例中, 后置摄像头为至少两个, 分别为主摄像头、景深摄像头、广角摄像头、长焦摄像头中的任意一种, 以实现主摄像头和景深摄像头融合实现背景虚化功能、主摄像头和广角摄像头融合实现全景拍摄以及 VR (Virtual Reality, 虚拟现实) 拍摄功能或者其它融合拍摄功能。在一些实施例中, 摄像头组件 506 还可以包括闪光灯。闪光灯可以是单色温闪光灯, 也可以是双色温闪光灯。双色温闪光灯是指暖光闪光灯和冷光闪光灯的组合, 可以用于不同色温下的光线补偿。

音频电路 507 可以包括麦克风和扬声器。麦克风用于采集用户及环境的声波, 并将声波转换为电信号输入至处理器 501 进行处理, 或者输入至射频电路 504 以实现语音通信。出于立体声采集或降噪的目的, 麦克风可以为多个, 分别设置在终端 500 的不同部位。麦克风还可以是阵列麦克风或全向采集型麦克风。扬声器则用于将来自处理器 501 或射频电路 504 的电信号转换为声波。扬声器可以是传统的薄膜扬声器, 也可以是压电陶瓷扬声器。当扬声器是压电陶瓷扬声器时, 不仅可以将电信号转换为人类可听见的声波, 也可以将电信号转换为人类听不见的声波以进行测距等用途。在一些实施例中, 音频电路 507 还可以包括耳机插孔。

定位组件 508 用于定位终端 500 的当前地理位置, 以实现导航或 LBS (Location Based Service, 基于位置的服务)。定位组件 508 可以是基于美国的 GPS (Global Positioning System, 全球定位系统)、中国的北斗系统或俄罗斯的伽利略系统的定位组件。

电源 509 用于为终端 500 中的各个组件进行供电。电源 509 可以是交流电、直流电、一次性电池或可充电电池。当电源 509 包括可充电电池时, 该可充电电池可以是有线充电电池或无线充电电池。有线充电电池是通过有线线路充电的电池, 无线充电电池是通过无线线圈充电的电池。该可充电电池还可以用于支持快充技术。

在一些实施例中, 终端 500 还包括有一个或多个传感器 510。该一个或多个传感器 510 包括但不限于: 加速度传感器 511、陀螺仪传感器 512、压力传

感器 513、指纹传感器 514、光学传感器 515 以及接近传感器 516。

加速度传感器 511 可以检测以终端 500 建立的坐标系的三个坐标轴上的加速度大小。比如，加速度传感器 511 可以用于检测重力加速度在三个坐标轴上的分量。处理器 501 可以根据加速度传感器 511 采集的重力加速度信号，控制触摸显示屏 505 以横向视图或纵向视图进行用户界面的显示。加速度传感器 511 还可以用于游戏或者用户的运动数据的采集。

陀螺仪传感器 512 可以检测终端 500 的机体方向及转动角度，陀螺仪传感器 512 可以与加速度传感器 511 协同采集用户对终端 500 的 3D 动作。处理器 501 根据陀螺仪传感器 512 采集的数据，可以实现如下功能：动作感应（比如根据用户的倾斜操作来改变 UI）、拍摄时的图像稳定、游戏控制以及惯性导航。

压力传感器 513 可以设置在终端 500 的侧边框和/或触摸显示屏 505 的下层。当压力传感器 513 设置在终端 500 的侧边框时，可以检测用户对终端 500 的握持信号，由处理器 501 根据压力传感器 513 采集的握持信号进行左右手识别或快捷操作。当压力传感器 513 设置在触摸显示屏 505 的下层时，由处理器 501 根据用户对触摸显示屏 505 的压力操作，实现对 UI 界面上的可操作性控件进行控制。可操作性控件包括按钮控件、滚动条控件、图标控件、菜单控件中的至少一种。

指纹传感器 514 用于采集用户的指纹，由处理器 501 根据指纹传感器 514 采集到的指纹识别用户的身份，或者，由指纹传感器 514 根据采集到的指纹识别用户的身份。在识别出用户的身份为可信身份时，由处理器 501 授权该用户执行相关的敏感操作，该敏感操作包括解锁屏幕、查看加密信息、下载软件、支付及更改设置等。指纹传感器 514 可以被设置终端 500 的正面、背面或侧面。当终端 500 上设置有物理按键或厂商 Logo 时，指纹传感器 514 可以与物理按键或厂商 Logo 集成在一起。

光学传感器 515 用于采集环境光强度。在一个实施例中，处理器 501 可以根据光学传感器 515 采集的环境光强度，控制触摸显示屏 505 的显示亮度。具体地，当环境光强度较高时，调高触摸显示屏 505 的显示亮度；当环境光强度较低时，调低触摸显示屏 505 的显示亮度。在另一个实施例中，处理器 501 还可以根据光学传感器 515 采集的环境光强度，动态调整摄像头组件 506 的拍摄参数。

接近传感器 516，也称距离传感器，通常设置在终端 500 的前面板。接近

传感器 516 用于采集用户与终端 500 的正面之间的距离。在一个实施例中，当接近传感器 516 检测到用户与终端 500 的正面之间的距离逐渐变小时，由处理器 501 控制触摸显示屏 505 从亮屏状态切换为息屏状态；当接近传感器 516 检测到用户与终端 500 的正面之间的距离逐渐变大时，由处理器 501 控制触摸显示屏 505 从息屏状态切换为亮屏状态。

本领域技术人员可以理解，图 5 中示出的结构并不构成对终端 500 的限定，可以包括比图示更多或更少的组件，或者组合某些组件，或者采用不同的组件布置。

本领域技术人员在考虑说明书及实践这里公开的申请后，将容易想到本申请的其它实施方案。本申请旨在涵盖本申请的任何变型、用途或者适应性变化，这些变型、用途或者适应性变化遵循本申请的一般性原理并包括本申请未公开的本技术领域中的公知常识或惯用技术手段。说明书和实施例仅被视为示例性的，本申请的真正范围和精神由下面的权利要求指出。

应当理解的是，本申请并不局限于上面已经描述并在附图中示出的精确结构，并且可以在不脱离其范围进行各种修改和改变。本申请的范围仅由所附的权利要求来限制。

权 利 要 求 书

1、一种对词库进行处理的方法，其特征在于，所述方法包括：

获取第一词库中的第一数据记录，第一数据记录包括一个多词条和所述多词条中的每个汉字对应的第一假名集合，所述多词条为包括多个汉字的词条，汉字对应的第一假名集合包括所述汉字对应的至少一个假名；

在第二词库中查找所述第一数据记录对应的多个目标数据记录，每个目标数据记录中的目标词条是所述多词条的不同组成部分，所述每个目标数据记录中的目标词条组成所述多词条，目标数据记录中的目标词条中的每个汉字对应的第二假名集合分别与所述每个汉字对应的第一假名集合相同；

当在所述第二词库中没有查找出所述第一数据记录对应的多个目标数据记录，则将所述第一数据记录保存在所述第二词库中。

2、如权利要求1所述的方法，其特征在于，所述在第二词库中查找所述第一数据记录对应的多个目标数据记录，包括：

将所述第一数据记录中的多词条划分成N个单词条，N大于1的整数，所述单词条为包括一个汉字的词条；

在第二词库中查找所述每个单词条对应的目标数据记录，单词条对应的目标数据记录包括所述单词条和所述单词条中的汉字对应的第二假名集合，所述汉字对应的第二假名集合与所述汉字对应的第一假名集合相同；

当在所述第二词库中没有查找出所述每个单词条对应的目标数据记录，则确定从所述第二词库中查找不出所述第一数据记录对应的多个目标数据记录。

3、如权利要求1或2所述的方法，其特征在于，所述获取第一词库中的第一数据记录之前，还包括：

将所述第一词库中的各第二数据记录保存在所述第二词库中，所述第二数据记录包括一个单词条和所述单词条中的汉字对应的第一假名集合。

4、如权利要求3所述的方法，其特征在于，所述将所述第一词库中的各第二数据记录保存在所述第二词库中，包括：

获取所述第一词库中的任一条数据记录；

在所述数据记录中包括一个第一假名集合时，确定所述数据记录中的词条为单词条并将所述数据记录保存在所述第二词库中。

5、如权利要求1或2所述的方法，其特征在于，所述获取第一词库中的第一数据记录之前，还包括：

将第三词库中包括预设应用场景的数据记录保存在第一词库中，所述第三词库中的每个数据记录包括一个词条、所述词条中的每个汉字对应的假名集合和应用场景。

6、如权利要求5所述的方法，其特征在于，所述将第三词库中包括预设应用场景的数据记录保存在第一词库中，包括：

从第三词库中获取包括预设应用场景的数据记录，所述数据记录包括一个词条、所述第一词条中的每个汉字对应的至少一个假名集合、所述至少一个假名集合中的每个假名集合的使用频率和所述预设应用场景；

根据所述每个汉字对应的至少一个假名集合的使用频率，分别从所述每个汉字对应的至少一个假名集合中选择所述每个汉字对应的假名集合；

将所述一个词条、为所述每个汉字选择假名集合组成第一数据记录并保存在所述第一词库中。

7、一种对词库进行处理的装置，其特征在于，所述装置包括：

获取模块，用于获取第一词库中的第一数据记录，第一数据记录包括一个多词条和所述多词条中的每个汉字对应的第一假名集合，所述多词条为包括多个汉字的词条，汉字对应的第一假名集合包括所述汉字对应的至少一个假名；

查找模块，用于在第二词库中查找所述第一数据记录对应的多个目标数据记录，每个目标数据记录中的目标词条是所述多词条的不同组成部分，所述每个目标数据记录中的目标词条组成所述多词条，目标数据记录中的目标词条中的每个汉字对应的第二假名集合分别与所述每个汉字对应的第一假名集合相同；

保存模块，用于当在所述第二词库中没有查找出所述第一数据记录对应的多个目标数据记录，则将所述第一数据记录保存在所述第二词库中。

8、如权利要求 7 所述的装置，其特征在于，所述查找模块包括：

划分单元，用于将所述第一数据记录中的多词条划分成 N 个单词条， N 大于 1 的整数，所述单词条为包括一个汉字的词条；

查找单元，用于在第二词库中查找所述每个单词条对应的目标数据记录，单词条对应的目标数据记录包括所述单词条和所述单词条中的汉字对应的第二假名集合，所述汉字对应的第二假名集合与所述汉字对应的第一假名集合相同；

第一确定单元，用于当在所述第二词库中没有查找出所述每个单词条对应的目标数据记录，则确定从所述第二词库中查找不出所述第一数据记录对应的多个目标数据记录。

9、如权利要求 7 或 8 所述的装置，其特征在于，所述保存模块，还用于：

将所述第一词库中的各第二数据记录保存在所述第二词库中，所述第二数据记录包括一个单词条和所述单词条中的汉字对应的第一假名集合。

10、如权利要求 9 所述的装置，其特征在于，所述保存模块包括：

第一获取单元，用于获取所述第一词库中的任一条数据记录；

第二确定单元，用于在所述数据记录中包括一个第一假名集合时，确定所述数据记录中的词条为单词条并将所述数据记录保存在所述第二词库中。

11、如权利要求 7 或 8 所述的装置，其特征在于，所述保存模块，还用于：

将第三词库中包括预设应用场景的数据记录保存在第一词库中，所述第三词库中的每个数据记录包括一个词条、所述词条中的每个汉字对应的假名集合和应用场景。

12、如权利要求 11 所述的装置，其特征在于，所述保存模块还包括：

第二获取单元，用于从第三词库中获取包括预设应用场景的数据记录，所述数据记录包括一个词条、所述第一词条中的每个汉字对应的至少一个假名集合、所述至少一个假名集合中的每个假名集合的使用频率和所述预设应用场景；

选择单元，用于根据所述每个汉字对应的至少一个假名集合的使用频率，分别从所述每个汉字对应的至少一个假名集合中选择所述每个汉字对应的假名集合；

保存单元，用于将所述一个词条、为所述每个汉字选择假名集合组成第一数据记录并保存在所述第一词库中。