(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 111428522 A
(43)申请公布日 2020.07.17

(21)申请号 202010207061.0

(22)申请日 2020.03.23

(71)申请人 腾讯科技(深圳)有限公司
地址 518057 广东省深圳市南山区高新区科技中一路腾讯大厦35层

(72)发明人 袁星宇 黄杰

(74)专利代理机构 北京三高永信知识产权代理有限责任公司 11138
代理人 祝亚男

(51)Int.Cl.
G06F 40/58(2020.01)
G06F 40/42(2020.01)
G06F 40/211(2020.01)
G06F 40/30(2020.01)
G06F 9/50(2006.01)
G06F 16/951(2019.01)
G06F 16/955(2019.01)
G06N 20/00(2019.01)

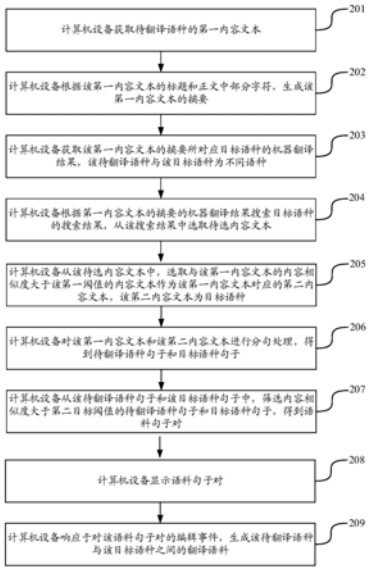
权利要求书2页 说明书20页 附图5页

(54)发明名称

翻译语料生成方法、装置、计算机设备及存储介质

(57)摘要

本申请公开了一种翻译语料生成方法、装置、计算机设备及存储介质，属于计算机技术领域。方法包括：获取待翻译语种的第一内容文本；确定与第一内容文本的内容相似度大于第一目标阈值的第二内容文本；对第一内容文本和第二内容文本进行分句处理，得到待翻译语种句子和目标语种句子；筛选内容相似度大于第二目标阈值的待翻译语种句子和目标语种句子，得到语料句子对；响应于对语料句子对的编辑事件，生成待翻译语种与目标语种之间的翻译语料。本申请直接向用户提供不同语种且内容相似度较高的待翻译语种句子和目标语种句子，使得用户只需对两者进行比对，并进行微调即可得到不同语种之间的翻译语料，提高了翻译语料生成的效率。



CN 111428522 A

1.一种翻译语料生成方法,其特征在于,所述方法包括:

获取待翻译语种的第一内容文本;

确定与所述第一内容文本的内容相似度大于第一目标阈值的第二内容文本,所述第二内容文本为目标语种,所述待翻译语种与所述目标语种为不同语种;

对所述第一内容文本和所述第二内容文本进行分句处理,得到待翻译语种句子和目标语种句子;

从所述待翻译语种句子和所述目标语种句子中,筛选内容相似度大于第二目标阈值的待翻译语种句子和目标语种句子,得到语料句子对;

响应于对所述语料句子对的编辑事件,生成所述待翻译语种与所述目标语种之间的翻译语料。

2.根据权利要求1所述的方法,其特征在于,所述从所述待翻译语种句子和所述目标语种句子中,筛选内容相似度大于第二目标阈值的待翻译语种句子和目标语种句子,得到语料句子对,包括:

从所述待翻译语种句子和所述目标语种句子中,确定顺序对应的待翻译语种句子和目标语种句子,得到待选句子对;

对于任一待选句子对,获取所述任一待选句子对中的待翻译语种句子和目标语种句子的内容相似度;

响应于所述任一待选句子对中的待翻译语种句子和目标语种句子的内容相似度大于所述第二目标阈值,将所述任一待选句子对作为所述语料句子对。

3.根据权利要求1所述的方法,其特征在于,所述从所述待翻译语种句子和所述目标语种句子中,筛选内容相似度大于第二目标阈值的待翻译语种句子和目标语种句子,得到语料句子对之后,所述方法还包括:

在语料生成界面中显示所述语料句子对,所述语料生成界面中设置有第一编辑区域和第二编辑区域,所述第一编辑区域用于对所述语料句子对中的待翻译语种句子进行编辑,所述第二编辑区域用于对所述语料句子对中的目标语种句子进行编辑。

4.根据权利要求3所述的方法,其特征在于,所述在语料生成界面中显示所述语料句子对,包括:

在所述语料生成界面中对所述语料句子对中的待翻译语种句子分行显示,不同待翻译语种句子位于不同行;

在所述语料生成界面中对所述语料句子对中的目标语种句子分行显示,不同目标语种句子位于不同行。

5.根据权利要求3所述的方法,其特征在于,所述响应于对所述语料句子对的编辑事件,生成所述待翻译语种与所述目标语种之间的翻译语料,包括下述任一项:

响应于在所述第一编辑区域内的编辑事件,获取编辑后的待翻译语种句子,基于所述编辑后的待翻译语种句子和所述语料句子对中的目标语种句子,生成所述翻译语料;或,

响应于在所述第二编辑区域内的编辑事件,获取编辑后的目标语种句子,基于所述编辑后的目标语种句子和所述语料句子对中的待翻译语种句子,生成所述翻译语料;或,

响应于在所述第一编辑区域和所述第二编辑区域内的编辑事件,获取编辑后的待翻译语种句子和编辑后的目标语种句子,基于所述编辑后的待翻译语种句子和所述编辑后的目

标语种句子,生成所述翻译语料。

6.根据权利要求3所述的方法,其特征在于,所述响应于对所述语料句子对的编辑事件,生成所述待翻译语种与所述目标语种之间的翻译语料之前,所述方法还包括下述至少一项:

响应于第一粘贴事件,在所述第一编辑区域内输入所述语料句子对中的待翻译语种句子;

响应于第二粘贴事件,在所述第二编辑区域内输入所述语料句子对中的目标语种句子。

7.根据权利要求1所述的方法,其特征在于,所述确定与所述第一内容文本的内容相似度大于第一目标阈值的第二内容文本,包括:

根据所述第一内容文本的标题和正文中部分字符,生成所述第一内容文本的摘要;

获取所述第一内容文本的摘要所对应目标语种的机器翻译结果;

根据所述机器翻译结果搜索所述目标语种的搜索结果,从所述搜索结果中选取待选内容文本;

从所述待选内容文本中,选取与所述第一内容文本的内容相似度大于所述第一目标阈值的内容文本作为所述第二内容文本。

8.一种翻译语料生成装置,其特征在于,所述装置包括:

获取模块,用于获取待翻译语种的第一内容文本;

确定模块,用于确定与所述第一内容文本的内容相似度大于第一目标阈值的第二内容文本,所述第二内容文本为目标语种,所述待翻译语种与所述目标语种为不同语种;

处理模块,用于对所述第一内容文本和所述第二内容文本进行分句处理,得到待翻译语种句子和目标语种句子;

筛选模块,用于从所述待翻译语种句子和所述目标语种句子中,筛选内容相似度大于第二目标阈值的待翻译语种句子和目标语种句子,得到语料句子对;

生成模块,用于响应于对所述语料句子对的编辑事件,生成所述待翻译语种与所述目标语种之间的翻译语料。

9.一种计算机设备,其特征在于,所述终端包括一个或多个处理器和一个或多个存储器,所述一个或多个存储器中存储有至少一条程序代码,所述程序代码由所述一个或多个处理器加载并执行以实现如权利要求1至7任一项所述的翻译语料生成方法。

10.一种计算机可读存储介质,其特征在于,所述计算机可读存储介质中存储有至少一条程序代码,所述至少一条程序代码由处理器加载并执行,以实现如权利要求1至7任一项所述的翻译语料生成方法。

翻译语料生成方法、装置、计算机设备及存储介质

技术领域

[0001] 本申请涉及计算机技术领域，尤其涉及一种翻译语料生成方法、装置、计算机设备及存储介质。

背景技术

[0002] 随着民汉交流更加频繁，所交流的内容也会更加广泛，然而由于不同民族的语种不同，民族之间的交流存在障碍，方便、快捷、准确的进行不同语种的翻译成为不同民族的共同需求。采用深度学习方法训练得到翻译模型，使用该翻译模型进行不同语种的翻译是一种广泛使用的技术手段，而翻译模型的训练需要大量的翻译语料，如何为模型提供翻译语料成为值得关注的问题。

[0003] 相关技术中，对于偏冷门的语种，依然需要专业领域的人士对每一句话或每一篇文章进行翻译，并且需要参考翻译字典或第三方翻译工具来辅助翻译，以维语和汉语之间的翻译为例，通过提供需翻译的维语文本，由维语专业领域的人士手动输入对应的汉语内容，从而得到维语文本对应的汉语文本，然后将维语文本和对应的汉语文本提供给模型作为翻译语料。

[0004] 上述技术在翻译语料的生成过程中，需要人工参考翻译字典或第三方翻译工具来进行不同语种的翻译，且手动输入全部的翻译内容，翻译语料生成的效率低。

发明内容

[0005] 本申请实施例提供了一种翻译语料生成方法、装置、计算机设备及存储介质，可以提高翻译语料生成的效率。所述技术方案如下：

[0006] 一方面，提供了一种翻译语料生成方法，所述方法包括：

[0007] 获取待翻译语种的第一内容文本；

[0008] 确定与所述第一内容文本的内容相似度大于第一目标阈值的第二内容文本，所述第二内容文本为目标语种，所述待翻译语种与所述目标语种为不同语种；

[0009] 对所述第一内容文本和所述第二内容文本进行分句处理，得到待翻译语种句子和目标语种句子；

[0010] 从所述待翻译语种句子和所述目标语种句子中，筛选内容相似度大于第二目标阈值的待翻译语种句子和目标语种句子，得到语料句子对；

[0011] 响应于对所述语料句子对的编辑事件，生成所述待翻译语种与所述目标语种之间的翻译语料。

[0012] 在一种可能实现方式中，所述获取所述任一待选句子对中的待翻译语种句子和目标语种句子的内容相似度，包括：

[0013] 获取所述任一待选句子对中的待翻译语种句子的机器翻译结果，所述机器翻译结果为所述目标语种；

[0014] 根据所述机器翻译结果和所述待选句子对中的目标语种句子，获取所述任一待选

句子对中的待翻译语种句子和目标语种句子的内容相似度。

[0015] 在一种可能实现方式中，所述从所述待翻译语种句子和所述目标语种句子中，筛选内容相似度大于第二目标阈值的待翻译语种句子和目标语种句子，得到语料句子对，包括：

[0016] 从所述待翻译语种句子和所述目标语种句子中，筛选内容相似度大于第二目标阈值的待翻译语种句子和目标语种句子，得到候选句子对；

[0017] 从所述候选句子对中确定内容相似度最大的句子对，将所述内容相似度最大的句子对作为所述语料句子对。

[0018] 在一种可能实现方式中，所述响应于对所述语料句子对的编辑事件，生成所述待翻译语种与所述目标语种之间的翻译语料之前，所述方法还包括：

[0019] 对所述语料生成界面进行光学字符识别，得到所述语料句子对中待翻译语种句子的文本坐标和目标语种句子的文本坐标；

[0020] 基于所述语料句子对中待翻译语种句子的文本坐标和目标语种句子的文本坐标，获取所述语料句子对中待翻译语种句子的内容和目标语种句子的内容；

[0021] 基于所述语料句子对中待翻译语种句子的内容和目标语种句子的内容，在所述语料生成界面的所述第一编辑区域内生成所述语料句子对中的待翻译语种句子，在所述语料生成界面的所述第二编辑区域内生成所述语料句子对中的目标语种句子。

[0022] 在一种可能实现方式中，所述从所述待选内容文本中，选取与所述第一内容文本的内容相似度大于所述第一目标阈值的内容文本作为所述第二内容文本，包括：

[0023] 获取所述第一内容文本的标题所对应目标语种的机器翻译结果；

[0024] 对于任一待选内容文本，根据所述第一内容文本的标题所对应目标语种的机器翻译结果、所述第一内容文本的摘要所对应目标语种的机器翻译结果、所述任一待选内容文本的标题和摘要，获取所述任一待选内容文本与所述第一内容文本的标题相似度和摘要相似度；

[0025] 对所述标题相似度和摘要相似度进行加权求和，得到所述任一待选内容文本与所述第一内容文本的内容相似度；

[0026] 当所述任一待选内容文本与所述第一内容文本的内容相似度大于所述第一目标阈值时，将所述述任一待选内容文本作为所述第二内容文本。

[0027] 在一种可能实现方式中，所述获取所述第一内容文本，包括：

[0028] 根据起始统一资源定位符URL，迭代爬取所述起始URL对应的起始页面上的至少一个URL，所述起始页面上的内容文本为所述待翻译语种；

[0029] 对于当前爬取到的URL，对所述当前爬取到的URL进行解析，得到所述当前爬取到的URL对应的内容文本作为所述第一内容文本。

[0030] 在一种可能实现方式中，所述方法还包括：

[0031] 在显示所述语料句子对时，显示所述语料句子对中的待翻译语种句子的机器翻译结果，所述机器翻译结果为所述目标语种。

[0032] 在一种可能实现方式中，所述响应于对所述语料句子对的编辑事件，生成所述待翻译语种与所述目标语种之间的翻译语料之后，所述方法还包括：

[0033] 基于所述翻译语料进行训练，得到翻译模型，所述翻译模型用于将所述待翻译语

种的内容文本翻译为所述目标语种的内容文本。

[0034] 一方面，提供了一种翻译语料生成装置，所述装置包括：

[0035] 获取模块，用于获取待翻译语种的第一内容文本；

[0036] 确定模块，用于确定与所述第一内容文本的内容相似度大于第一目标阈值的第二内容文本，所述第二内容文本为目标语种，所述待翻译语种与所述目标语种为不同语种；

[0037] 处理模块，用于对所述第一内容文本和所述第二内容文本进行分句处理，得到待翻译语种句子和目标语种句子；

[0038] 筛选模块，用于从所述待翻译语种句子和所述目标语种句子中，筛选内容相似度大于第二目标阈值的待翻译语种句子和目标语种句子，得到语料句子对；

[0039] 生成模块，用于响应于对所述语料句子对的编辑事件，生成所述待翻译语种与所述目标语种之间的翻译语料。

[0040] 在一种可能实现方式中，所述筛选模块用于：

[0041] 从所述待翻译语种句子和所述目标语种句子中，确定顺序对应的待翻译语种句子和目标语种句子，得到待选句子对；

[0042] 对于任一待选句子对，获取所述任一待选句子对中的待翻译语种句子和目标语种句子的内容相似度；

[0043] 响应于所述任一待选句子对中的待翻译语种句子和目标语种句子的内容相似度大于所述第二目标阈值，将所述任一待选句子对作为所述语料句子对。

[0044] 在一种可能实现方式中，所述筛选模块用于：

[0045] 获取所述任一待选句子对中的待翻译语种句子的机器翻译结果，所述机器翻译结果为所述目标语种；

[0046] 根据所述机器翻译结果和所述待选句子对中的目标语种句子，获取所述任一待选句子对中的待翻译语种句子和目标语种句子的内容相似度。

[0047] 在一种可能实现方式中，所述筛选模块用于：

[0048] 从所述待翻译语种句子和所述目标语种句子中，筛选内容相似度大于第二目标阈值的待翻译语种句子和目标语种句子，得到候选句子对；

[0049] 从所述候选句子对中确定内容相似度最大的句子对，将所述内容相似度最大的句子对作为所述语料句子对。

[0050] 在一种可能实现方式中，所述装置还包括：

[0051] 显示模块，用于在语料生成界面中显示所述语料句子对，所述语料生成界面中设置有第一编辑区域和第二编辑区域，所述第一编辑区域用于对所述语料句子对中的待翻译语种句子进行编辑，所述第二编辑区域用于对所述语料句子对中的目标语种句子进行编辑。

[0052] 在一种可能实现方式中，所述显示模块用于：

[0053] 在所述语料生成界面中对所述语料句子对中的待翻译语种句子分行显示，不同待翻译语种句子位于不同行；

[0054] 在所述语料生成界面中对所述语料句子对中的目标语种句子分行显示，不同目标语种句子位于不同行。

[0055] 在一种可能实现方式中，所述生成模块用于执行下述任一项：

[0056] 响应于在所述第一编辑区域内的编辑事件，获取编辑后的待翻译语种句子，基于所述编辑后的待翻译语种句子和所述语料句子对中的目标语种句子，生成所述翻译语料；或，

[0057] 响应于在所述第二编辑区域内的编辑事件，获取编辑后的目标语种句子，基于所述编辑后的目标语种句子和所述语料句子对中的待翻译语种句子，生成所述翻译语料；或，

[0058] 响应于在所述第一编辑区域和所述第二编辑区域内的编辑事件，获取编辑后的待翻译语种句子和编辑后的目标语种句子，基于所述编辑后的待翻译语种句子和所述编辑后的目标语种句子，生成所述翻译语料。

[0059] 在一种可能实现方式中，所述装置还包括下述至少一项：

[0060] 第一输入模块，用于响应于第一粘贴事件，在所述第一编辑区域内输入所述语料句子对中的待翻译语种句子；

[0061] 第二输入模块，用于响应于第二粘贴事件，在所述第二编辑区域内输入所述语料句子对中的目标语种句子。

[0062] 在一种可能实现方式中，所述装置还包括：

[0063] 识别模块，用于对所述语料生成界面进行光学字符识别，得到所述语料句子对中待翻译语种句子的文本坐标和目标语种句子的文本坐标；

[0064] 所述获取模块还用于基于所述语料句子对中待翻译语种句子的文本坐标和目标语种句子的文本坐标，获取所述语料句子对中待翻译语种句子的内容和目标语种句子的内容；

[0065] 生成模块，用于基于所述语料句子对中待翻译语种句子的内容和目标语种句子的内容，在所述语料生成界面的所述第一编辑区域内生成所述语料句子对中的待翻译语种句子，在所述语料生成界面的所述第二编辑区域内生成所述语料句子对中的目标语种句子。

[0066] 在一种可能实现方式中，所述确定模块用于：

[0067] 根据所述第一内容文本的标题和正文中部分字符，生成所述第一内容文本的摘要；

[0068] 获取所述第一内容文本的摘要所对应目标语种的机器翻译结果；

[0069] 根据所述机器翻译结果搜索所述目标语种的搜索结果，从所述搜索结果中选取待选内容文本；

[0070] 从所述待选内容文本中，选取与所述第一内容文本的内容相似度大于所述第一目标阈值的内容文本作为所述第二内容文本。

[0071] 在一种可能实现方式中，所述确定模块用于：

[0072] 获取所述第一内容文本的标题所对应目标语种的机器翻译结果；

[0073] 对于任一待选内容文本，根据所述第一内容文本的标题所对应目标语种的机器翻译结果、所述第一内容文本的摘要所对应目标语种的机器翻译结果、所述任一待选内容文本的标题和摘要，获取所述任一待选内容文本与所述第一内容文本的标题相似度和摘要相似度；

[0074] 对所述标题相似度和摘要相似度进行加权求和，得到所述任一待选内容文本与所述第一内容文本的内容相似度；

[0075] 当所述任一待选内容文本与所述第一内容文本的内容相似度大于所述第一目标

阈值时，将所述述任一待选内容文本作为所述第二内容文本。

[0076] 在一种可能实现方式中，所述获取模块用于：

[0077] 根据起始统一资源定位符URL，迭代爬取所述起始URL对应的起始页面上的至少一个URL，所述起始页面上的内容文本为所述待翻译语种；

[0078] 对于当前爬取到的URL，对所述当前爬取到的URL进行解析，得到所述当前爬取到的URL对应的内容文本作为所述第一内容文本。

[0079] 在一种可能实现方式中，所述显示模块还用于：

[0080] 在显示所述语料句子对时，显示所述语料句子对中的待翻译语种句子的机器翻译结果，所述机器翻译结果为所述目标语种。

[0081] 在一种可能实现方式中，所述装置还包括：

[0082] 训练模块，用于基于所述翻译语料进行训练，得到翻译模型，所述翻译模型用于将所述待翻译语种的内容文本翻译为所述目标语种的内容文本。

[0083] 一方面，提提供了一种计算机设备，所述计算机设备包括一个或多个处理器和一个或多个存储器，所述一个或多个存储器中存储有至少一条程序代码，所述程序代码由所述一个或多个处理器加载并执行以实现上述翻译语料生成方法。

[0084] 一方面，提供了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有至少一条程序代码，所述至少一条程序代码由处理器加载并执行以实现上述翻译语料生成方法。

[0085] 本申请实施例提供的技术方案带来的有益效果至少包括：

[0086] 通过获取与待翻译语种的第一内容文本为不同语种且内容相似度大于第一目标阈值的第二内容文本，然后分别对内容文本进行分句处理，得到待翻译语种句子和目标语种句子，从中筛选内容相似度大于第二目标阈值的待翻译语种句子和目标语种句子，得到语料句子对，使得用户可以对语料句子对进行编辑，从而响应于对语料句子对的编辑事件，生成待翻译语种与该目标语种之间的翻译语料。上述技术方案直接向用户提供不同语种且内容相似度较高的待翻译语种句子和目标语种句子，使得用户只需对两者进行比对，并进行微调即可得到不同语种之间的翻译语料，提高了翻译语料生成的效率。

附图说明

[0087] 为了更清楚地说明本申请实施例中的技术方案，下面将对实施例描述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申请的一些实施例，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。

[0088] 图1是本申请实施例提供的一种翻译语料生成方法的实施环境示意图；

[0089] 图2是本申请实施例提供的一种翻译语料生成方法的流程图；

[0090] 图3是本申请实施例提供的一种翻译语料生成过程的示意图；

[0091] 图4是本申请实施例提供的一种语料生成界面的示意图；

[0092] 图5是本申请实施例提供的一种语料生成界面的示意图；

[0093] 图6是本申请实施例提供的一种翻译语料生成装置的结构示意图；

[0094] 图7是本申请实施例提供的一种终端的结构示意图；

[0095] 图8是本申请实施例提供的一种服务器的结构示意图。

具体实施方式

[0096] 为使本申请的目的、技术方案和优点更加清楚，下面将结合附图对本申请实施方式作进一步地详细描述。

[0097] 在对本申请实施例进行详细地解释说明之前，先对本申请实施例涉及到的一些名词进行解释说明。

[0098] 域名：又称网域，是由一串用点分隔的名字组成的Internet(互联网)上某一台计算机或计算机组的名称，用于在数据传输时对计算机的定位标识。

[0099] Goose3：最初是用Java编写的一篇文章提取器，最近将它(Auff2011)转换成Scala项目，这是Python中的完全重写。该软件的目标是获取任何新闻文章或文章类型的网页，不仅提取文章的主体，而且还提取所有元数据和图片。

[0100] Selenium：是一个用于Web应用程序测试的工具。Selenium测试直接运行在浏览器中，就像真正的用户在操作一样。支持的浏览器包括IE(Internet Explorer，网页浏览器)(IE7、IE8、IE9、IE10、IE11)，Mozilla Firefox，Safari，Google Chrome，Opera等。这个工具的主要功能包括：测试与浏览器的兼容性，包括测试应用程序看是否能够很好的工作在不同浏览器和操作系统之上。测试系统功能包括创建回归测试检验软件功能和用户需求。支持自动录制动作和自动生成Net、Java、Perl等不同语言的测试脚本。

[0101] Gensim：是一款开源的第三方Python工具包，用于从原始的非结构化的文本中，无监督地学习到文本隐层的主题向量表达。它支持包括TF-IDF(Term Frequency-Inverse Document Frequency，词频和逆向文档频率)，LSA(Latent Semantic Analysis，潜在语义分析)，LDA(Linear Discriminant Analysis，线性判别分析)，和Word2vec在内的多种主题模型算法，支持流式训练，并提供了诸如相似度计算，信息检索等一些常用任务的API(Application Programming Interface，应用程序编程接口)。

[0102] 人工智能(Artificial Intelligence，AI)：是利用数字计算机或者数字计算机控制的机器模拟、延伸和扩展人的智能，感知环境、获取知识并使用知识获得最佳结果的理论、方法、技术及应用系统。换句话说，人工智能是计算机科学的一个综合技术，它企图了解智能的实质，并生产出一种新的能以人类智能相似的方式做出反应的智能机器。人工智能也就是研究各种智能机器的设计原理与实现方法，使机器具有感知、推理与决策的功能。人工智能技术是一门综合学科，涉及领域广泛，既有硬件层面的技术也有软件层面的技术。人工智能基础技术一般包括如传感器、专用人工智能芯片、云计算、分布式存储、大数据处理技术、操作/交互系统、机电一体化等技术。

[0103] 深度学习/机器学习(Machine Learning，ML)：是一门多领域交叉学科，涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科。专门研究计算机怎样模拟或实现人类的学习行为，以获取新的知识或技能，重新组织已有的知识结构使之不断改善自身的性能。机器学习是人工智能的核心，是使计算机具有智能的根本途径，其应用遍及人工智能的各个领域。机器学习和深度学习通常包括人工神经网络、置信网络、强化学习、迁移学习、归纳学习、示教学习等技术。

[0104] 本申请实施例提供的方案涉及人工智能的深度学习技术，通过生成用于进行模型

训练的语料，可以采用深度学习方法训练得到翻译模型，从而使用该翻译模型进行不同语种的翻译。具体内容将通过下述实施例进行说明。

[0105] 图1是本申请实施例提供的一种翻译语料生成方法的实施环境示意图，参见图1，该实施环境中可以包括终端101和服务器102。

[0106] 终端101通过无线网络或有线网络与服务器102相连。终端101可以是智能手机、平板电脑、便携计算机等设备。终端101安装和运行有支持语料生成的应用程序。示例性的，终端101是用户使用的终端，终端101中运行的应用程序内登录有用户账号。

[0107] 服务器102可以是云计算平台、虚拟化中心等。服务器102用于为支持语料生成的应用程序提供后台服务。可选地，服务器102承担主要语料生成工作，终端101承担次要语料生成工作；或者，服务器102承担次要语料生成工作，终端101承担主要语料生成工作；或者，服务器102或终端101分别可以单独承担语料生成工作。

[0108] 可选地，服务器102包括：接入服务器、语料生成服务器和数据库。接入服务器用于为终端101提供接入服务。语料生成服务器用于提供语料生成有关的后台服务。该数据库可以包括语料数据库以及用户信息数据库等，基于服务器所提供的不同服务可以对应于不同数据库，语料生成服务器可以是一台或多台。当语料生成服务器是多台时，存在至少两台语料生成服务器用于提供不同的服务，和/或，存在至少两台语料生成服务器用于提供相同的服务，比如以负载均衡方式提供同一种服务，本申请实施例对此不加以限定。

[0109] 终端101可以泛指多个终端中的一个，本实施例仅以终端101来举例说明。

[0110] 本领域技术人员可以知晓，上述终端的数量可以更多或更少。比如上述终端可以仅为一个，或者上述终端为几十个或几百个，或者更多数量，此时上述实施环境中还包括其他终端。本申请实施例对终端的数量和设备类型不加以限定。

[0111] 图2是本申请实施例提供的一种翻译语料生成方法的流程图。该方法由计算机设备执行，该计算机设备可以是终端，也可以是服务器，参见图2，该方法可以包括：

[0112] 201、计算机设备获取待翻译语种的第一内容文本。

[0113] 其中，待翻译语种可以为任一种需要翻译或较难翻译的语种，例如，待翻译语种可以为维语，相应地，第一内容文本可以为维语文章。

[0114] 在一种可能实现方式中，该获取待翻译语种的第一内容文本，包括：根据起始URL(Uniform Resource Locator，统一资源定位符)，迭代爬取该起始URL对应的页面上的至少一个URL，该页面上的内容文本为待翻译语种；对于当前爬取到的URL，对该当前爬取到的URL进行解析，得到该当前爬取到的URL对应的内容文本作为该第一内容文本。

[0115] 其中，该起始URL可以由用户预先指定，起始URL对应的页面可以称为起始页面，该页面上的至少一个URL可以是该页面的至少一个子页面的URL。

[0116] 计算机设备可以通过Scrapy爬虫框架，限定爬取的域名(domin)，迭代爬取起始页面上的所有URL并存入爬取队列中。同时，将爬取到的URL通过Goose3库进行解析，提取页面的内容文本作为第一内容文本。重复以上过程，直到此域名下的所以URL都已经进入到爬取队列中。

[0117] 参见图3，图3是本申请实施例提供的一种翻译语料生成过程的示意图，如图3中的流程301所示，计算机设备可以通过Scrapy爬取全站文章，起始URL可以为一个网站，如维语网站，该起始URL对应的页面可以为该维语网站的首页，计算机设备可以提取该网站的首页

作为起始URL，将其加入Scrapy爬取队列，进行迭代爬取，判断Scrapy爬取队列是否爬取完毕，也即是，判断起始URL对应的页面上的所有URL是否爬取完毕，如果否(未爬取完毕)，则继续爬取，同时打开当前爬取到的URL，提取文本内容，得到第一内容文本，如维语文章文本；如果是(爬取完毕)，则终止爬取。

[0118] 通过根据一个起始URL，迭代爬取对应页面上的所有URL，对URL进行解析，得到第一内容文本，提供了一种获取第一内容文本的有效方式，可以提高第一内容文本的获取效率。

[0119] 202、计算机设备根据该第一内容文本的标题和正文中部分字符，生成该第一内容文本的摘要。

[0120] 在一种可能实现方式中，该部分字符可以是正文的前第一数量的字符，计算机设备可以提取第一内容文本的标题和正文，将标题和正文的前第一数量的字符组装成摘要。例如，第一内容文本为维语文章文本，则计算机设备可以提取维语文章文本的标题和正文，将标题和正文的前150个字符组装成文章摘要。通过根据内容文本的标题和正文中前一定数量的字符生成摘要，使得该摘要能够很好的代表该内容文本，提高了摘要的准确性和可靠性，从而便于后续的文本搜索。在一些可能实施例中，该部分字符也可以是正文的中间目标数量的字符，或正文的结尾目标数量的字符，本申请实施例对此不做限定。

[0121] 计算机设备获取的第一内容文本可以有多个，计算机设备每获取一个第一内容文本，则执行步骤202，生成一个第一内容文本的摘要。

[0122] 203、计算机设备获取该第一内容文本的摘要所对应目标语种的机器翻译结果，该待翻译语种与该目标语种为不同语种。

[0123] 计算机设备可以对第一内容文本的摘要进行目标语种的机器翻译处理，将待翻译语种的摘要翻译成目标语种的摘要，如通过翻译引擎进行翻译，得到第一内容文本的摘要的机器翻译结果，该机器翻译结果为目标语种，例如，目标语种可以为汉语。

[0124] 以第一内容文本为维语文章文本为例，如图3所示，计算机设备可以对维语文章文本的摘要进行维汉翻译，具体地，计算机设备可以通过维汉翻译引擎，将维语文章文本的摘要翻译成汉语，得到汉语摘要，如接入“民汉翻译”接口，进行汉语翻译，将维语摘要翻译成汉语摘要。

[0125] 204、计算机设备根据第一内容文本的摘要的机器翻译结果搜索目标语种的搜索结果，从该搜索结果中选取待选内容文本。

[0126] 计算机设备可以将第一内容文本的摘要的机器翻译结果作为搜索关键词，进行搜索处理，如过Selenium模拟浏览器打开搜索引擎的首页，将摘要的机器翻译结果作为搜索关键词进行搜索，得到至少一个搜索结果，从至少一个搜索结果中选取待选内容文本。

[0127] 在一种可能实现方式中，计算机设备从搜索结果中从搜索结果中选取待选内容文本，包括：从搜索结果中选取前第二数量的搜索结果作为待选内容文本。通过将搜索结果中前一定数量的搜索结果作为待选内容文本，由于搜索结果越靠前表明与搜索关键词越相关，因此可以保证待选内容文本与第一内容文本的内容相似度较高。如图3所示，计算机设备可以将汉语摘要作为搜索关键词进行搜索，将搜索结果的前三位作为待选汉语文章文本。

[0128] 205、计算机设备从该待选内容文本中，选取与该第一内容文本的内容相似度大于

该第一目标阈值的内容文本作为该第一内容文本对应的第二内容文本，该第二内容文本为目标语种，该待翻译语种与该目标语种为不同语种。

[0129] 其中，该第一内容文本与该第二内容文本的语种不同且内容相似度大于第一目标阈值，该内容相似度用于指示描述同一事件的可能性，也即是，两个文本为同一内容采用不同语种描述的可能性。

[0130] 在一种可能实现方式中，从该待选内容文本中，选取与该第一内容文本的内容相似度大于该第一目标阈值的内容文本作为该第二内容文本，包括下述步骤一至步骤四：

[0131] 步骤一、获取该第一内容文本的标题所对应目标语种的机器翻译结果和该第一内容文本的摘要所对应目标语种的机器翻译结果。

[0132] 计算机设备可以将第一内容文本的标题和摘要分别通过翻译引擎进行翻译，将待翻译语种的标题和摘要翻译为目标语种的标题和摘要，例如，第一内容文本为维语文章文本，其标题和摘要分别为维语标题和维语摘要，则计算机设备可以将维语标题和维语摘要分别通过维汉翻译引擎，翻译成汉语标题和汉语摘要。

[0133] 步骤二、对于任一待选内容文本，根据该第一内容文本的标题所对应目标语种的机器翻译结果、该第一内容文本的摘要所对应目标语种的机器翻译结果、该任一待选内容文本的标题和摘要，获取该任一待选内容文本与该第一内容文本的标题相似度和摘要相似度。

[0134] 计算机设备可以待选内容文本的标题和摘要提取出来作为待选标题和待选摘要，使用相似度算法，分别计算待选内容文本的标题与第一内容文本的标题的机器翻译结果之间的相似度，将其作为待选内容文本与第一内容文本的标题相似度，以及计算待选内容文本的摘要与第一内容文本的摘要的机器翻译结果之间的相似度，将其作为待选内容文本与第一内容文本的摘要相似度。其中，相似度算法可以是Gensim相似度算法或其他文本相似度算法。

[0135] 步骤三、对该标题相似度和摘要相似度进行加权求和，得到该任一待选内容文本与该第一内容文本的内容相似度。

[0136] 计算机设备可以分别以第一权重和第二权重，对标题相似度和摘要相似度进行加权求和，将加权求和结果作为待选内容文本与该第一内容文本的内容相似度。例如，第一权重可以为0.7，第二权重可以为0.3，内容相似度计算公式可以表示如下：

[0137] total simi＝np.array([i[1]for i in title_simi])*0.7+np.array([i[1]for i in descripytion_simi])*0.3

[0138] 其中，total simi表示内容相似度，np.array([i[1]for i in title_simi])表示标题相似度，np.array([i[1]for i in descripytion_simi])表示摘要相似度。

[0139] 步骤四、当该任一待选内容文本与该第一内容文本的内容相似度大于该第一目标阈值时，将该述任一待选内容文本作为该第二内容文本。

[0140] 如图3所示，计算机设备可以通过相似度计算，将内容相似度大于第一目标阈值的待选内容文本作为第一内容文本对应的第二内容文本，将该第一内容文本和第二内容文本匹配为平行语料。例如，第一内容文本为维语文章文本，待选内容文本为汉语文章文本，则可以将内容相似度大于0.45的汉语文章文本，匹配为维语文章文本对应的汉语文章文本。

[0141] 通过根据第一内容文本的标题和摘要的机器翻译结果，以及待选内容文本的标题

和摘要，分别计算标题之间的相似度，以及对应摘要之间的相似度，然后计算一个综合相似度，根据综合相似度选取对应的第二内容文本，提供了一种计算不同内容文本的内容相似度的有效方式，保证了准确性。

[0142] 需要说明的是，步骤201至步骤205是确定与该第一内容文本的内容相似度大于目标阈值的第二内容文本的一种可能实现方式，其中，该第一内容文本与该第二内容文本的语种不同且内容相似度大于第一目标阈值。在一些可能实施例中，计算机设备在获取第一内容文本后，可以提取第一内容文本的标题，获取标题的机器翻译结果，基于标题的机器翻译结果进行搜索，从搜索结果中选取第一内容文本对应的第二内容文本。在另一些可能实施例中，计算机设备在获取第一内容文本后，也可以直接对第一内容文本进行机器翻译处理，得到第一内容文本的机器翻译结果，根据第一内容文本的机器翻译结果进行搜索，从搜索结果中选取第一内容文本对应的第二内容文本。

[0143] 上述步骤201至步骤205的过程也即为图3中的流程302所示的平行语料匹配环节的一部分。通过获取第一内容文本后，生成该第一内容文本的摘要，根据摘要的翻译结果进行搜索，从搜索结果中选取第一内容文本对应的第二内容文本，由于摘要根据内容文本的标题和正文中前第一数量的字符生成，根据摘要的翻译结果搜索到的内容文本可能与第一内容文本为同篇文章，进而从中选择与第一内容文本内容相似度大的内容文本作为第二内容文本，可以保证准确性。

[0144] 206、计算机设备对该第一内容文本和该第二内容文本进行分句处理，得到待翻译语种句子和目标语种句子。

[0145] 计算机设备可以根据第一内容文本中的标点符号，对第一内容文本进行分句处理，得到待翻译语种句子，该待翻译语种句子的数量可以为一个或多个，计算机设备可以根据第二内容文本中的标点符号，对第二内容文本进行分句处理，得到目标语种句子，该目标语种句子的数量可以为一个或多个，如图3所示。

[0146] 207、计算机设备从该待翻译语种句子和该目标语种句子中，筛选内容相似度大于第二目标阈值的待翻译语种句子和目标语种句子，得到语料句子对。

[0147] 在一种可能实现方式中，该步骤207包括：从待翻译语种句子和目标语种句子中，确定顺序对应的待翻译语种句子和目标语种句子，得到待选句子对；对于任一待选句子对，获取任一待选句子对中的待翻译语种句子和目标语种句子的内容相似度；响应于任一待选句子对中的待翻译语种句子和目标语种句子的内容相似度大于第二目标阈值，将任一待选句子对作为语料句子对。通过获取包括顺序对应的不同语种句子的候选句子对，从中筛选出所包括句子之间的内容相似度较大的句子对作为语料句子对，提供了一种获取语料句子对的有效方式，提高了语料句子对的获取效率。

[0148] 该待翻译语种句子和目标语种句子可以按照顺序形成对应关系，例如，第一个待翻译语种句子与第一个目标语种句子对应，顺序对应的两个句子可以作为一个句子对，这样可以得到一个或多个候选句子对，每个候选句子对包括不同语种的两个句子。计算机设备可以使用相似度算法，计算每个候选句子对的内容相似度。具体地，对于任一候选句子对，计算机设备可以获取该任一候选句子对中的待翻译语种句子的机器翻译结果，根据该待翻译语种句子的机器翻译结果和该任一句子对中的目标语种句子，获取该任一候选句子对的内容相似度。计算机设备可以使用相似度算法，计算任一候选句子对中待翻译语种句

子的机器翻译结果与该任一候选句子对中目标语种句子之间的相似度，将其作为该任一候选句子对的内容相似度。其中，该相似度算法可以是Gensim相似度算法或其他文本相似度算法。

[0149] 计算机设备可以根据候选句子对的内容相似度，对候选句子对进行相似度匹配，如图3所示。对于任一候选句子对，如果该候选句子对的内容相似度大于第二目标阈值，则计算机设备可以认为该候选句子对为匹配的句子对，因此计算机设备可以将该候选句子对作为语料句子对。如果该候选句子对的内容相似度小于第二目标阈值，则计算机设备可以认为该候选句子对为不匹配的句子对，因此，计算机设备可以舍弃该句子对。

[0150] 在一种可能实现方式中，该步骤207包括：计算机设备从待翻译语种句子和目标语种句子中，筛选内容相似度大于第二目标阈值的待翻译语种句子和目标语种句子，得到候选句子对；从候选句子对中确定内容相似度最大的句子对，将内容相似度最大的句子对作为语料句子对。通过筛选出所包括句子之间的内容相似度最大的句子对作为语料句子对，可以保证语料句子对的准确性。

[0151] 计算机设备可以将筛选出的内容相似度大于第二目标阈值的待翻译语种句子和目标语种句子作为待选句子对，计算机设备可以进一步地，从候选句子对中，确定待翻译语种句子和目标语种句子的内容相似度最大的句子对，将该句子对作为语料句子对。

[0152] 在一种可能实现方式中，该步骤207包括：对于每个待翻译语种句子，计算机设备获取该待翻译语种句子与每个目标语种句子之间的内容相似度，从目标语种句子中，筛选与该待翻译语种句子的内容相似度大于第二目标阈值的目标语种句子，将该待翻译语种句子与筛选出的目标语种句子作为语料句子对。

[0153] 计算机设备可以计算任意两个待翻译语种句子和目标语种句之间的内容相似度，从中筛选出内容相似度大于第二目标阈值的待翻译语种句子和目标语种句子作为语料句子对，可以保证语料句子对的准确性。

[0154] 上述步骤206至步骤207的过程也即为图3中的流程302所示的平行语料匹配环节的另一部分。通过选取内容相似度较大的句子对作为语料句子对，而丢弃内容相似度较小的句子对，可以避免无用信息的干扰，提高语料句子对所包括的句子的内容相似度。

[0155] 通过借助模拟浏览器、文本相似度的技术，使用机器自动化全流程替代人工手工搜索维汉平行语料，更具体地，根据一个已知的维语网站的URL，获取维语文章的标题与正文，并生成摘要，通过文本相似度算法匹配相关汉语文章，然后对维语文章和对应的汉语文章进行切句，并逐一通过相似度算法进行匹配，选取匹配的句子对，丢弃不匹配的句子对，也即是丢弃无关信息，可以避免因文章搜索，带来的文章超长内容冗余的问题。

[0156] 在一种可能实现方式中，计算机设备从待翻译语种句子和目标语种句子中，筛选内容相似度大于第二目标阈值的待翻译语种句子和目标语种句子，得到语料句子对之后，本申请实施例提供的方法还包括：计算机设备将语料句子对存储到目标文件中。通过将语料句子对进行存储，使得计算机设备可以在需要时，直接从目标文件中获取语料句子对，提高语料句子对的获取效率。

[0157] 208、计算机设备显示语料句子对。

[0158] 在一种可能实现方式中，计算机设备显示该语料句子对，包括：计算机设备在语料生成界面中显示语料句子对，该语料生成界面中设置有第一编辑区域和第二编辑区域，该

第一编辑区域用于对该语料句子对中的待翻译语种句子进行编辑，该第二编辑区域用于对该语料句子对中的目标语种句子进行编辑。通过在界面上显示语料句子对以及不同语种句子对应的编辑区域，使得用户在看到不同语种句子后，如果认为需要任一语种句子，则可以在对应的编辑区域内进行编辑，提升了人工标注的效率。

[0159] 语料生成界面中可以设置有用于显示语料句子对的区域，包括用于显示语料句子对中的待翻译语种句子的区域和用于显示语料句子对中的目标语种句子的区域，除了这两个区域以外，语料生成界面中还可以设置有用于编辑待翻译语种句子的第一编辑区域和用于编辑目标语种句子的第二编辑区域。

[0160] 参见图4，图4是本申请实施例提供的一种语料生成界面的示意图，如图4所示，计算机设备可以在语料生成界面上显示语料句子对中的待翻译语种句子，如在图4中维语内容对应的区域401内显示语料句子对中的维语句子，计算机设备可以在语料生成界面上显示语料句子对中的目标语种句子，如在图4中相似内容对应的区域402内显示语料句子对中的汉语句子，除了显示语料句子对以外，该语料生成界面上还可以提供有第一编辑区域(如图4中的维语输入区403)和第二编辑区域(如图4中的汉语输入区404)，分别用于对待翻译语种句子和目标语种句子进行编辑。

[0161] 在一种可能实现方式中，计算机设备在语料生成界面中显示该语料句子对，包括：计算机设备在语料生成界面中对语料句子对中的待翻译语种句子分行显示，不同待翻译语种句子位于不同行；在语料生成界面中对该语料句子对中的目标语种句子分行显示，不同目标语种句子位于不同行。如图5所示，计算机设备可以在区域501内对维语句子分行显示，在区域502内对汉语句子分行显示。通过将句子分行显示，便于用户更直观的进行句子比对，提高标注效率。

[0162] 在一种可能实现方式中，本申请实施例提供的方法还包括：在显示该语料句子对时，显示语料句子对中的待翻译语种句子的机器翻译结果，该机器翻译结果为该目标语种。通过在界面中显示待翻译语种句子的机器翻译结果，使得用户可以参考机器翻译结果，对待翻译语种句子或目标语种句子进行编辑，可以提升人工标注的效率。

[0163] 如图4所示，计算机设备可以在语料生成界面上显示待翻译语种句子的机器翻译结果，如在图4中机翻结果对应的区域405内显示待翻译语种句子的机器翻译结果。在一种可能实现方式中，计算机设备可以对待翻译语种句子的机器翻译结果中的句子分行显示，该待翻译语种句子的机器翻译结果中的不同句子位于不同行。如图5所示，在图5中机翻结果对应的区域505内，对待翻译语种句子的机器翻译结果中的句子分行显示。通过将句子分行显示，便于用户更直观的参考和对比，提高标注效率。

[0164] 在一种可能实现方式中，计算机设备显示语料句子对之前，本申请实施例提供的方法还包括：响应于语料标注请求，从该目标文件中获取该语料句子对。通过在接收到语料标注请求时，从目标文件中获取预先存储的语料句子对，使得计算机设备可以在用户需要标注语料时，快速地为用户提供语料句子对进行标注，从而提高标注效率。

[0165] 其中，该语料标注请求用于请求生成语料，该语料标注请求可以由用户操作触发，如计算机设备可以提供一个界面，该界面中可以包括语料标注版块，用户可以点击该语料标注版块，触发该语料标注请求。计算机设备可以在接收到语料标注请求时，从目标文件中获取预先存储的语料句子对，然后进行显示。

[0166] 209、计算机设备响应于对语料句子对的编辑事件，生成待翻译语种与目标语种之间的翻译语料。

[0167] 计算机设备响应于对语料句子对的编辑事件，生成待翻译语种与目标语种之间的翻译语料，包括下述任一种可能实现方式：

[0168] 第一种方式、响应于在该第一编辑区域内的编辑事件，获取编辑后的待翻译语种句子，基于编辑后的待翻译语种句子和语料句子对中的目标语种句子，生成翻译语料。

[0169] 如果用户仅在第一编辑区域内对待翻译语种句子进行编辑，则计算机设备可以获取到编辑后的待翻译语种句子，从而基于编辑后的待翻译语种句子和未编辑的目标语种句子生成翻译语料。例如，待翻译语种句子为维语句子，目标语种句子为汉语句子，如果用户认为维语句子的内容需要调整，则可以在第一编辑区域内对维语句子的内容进行编辑，得到编辑后的维语句子，与原始的汉语句子生成翻译语料。

[0170] 第二种方式、响应于在该第二编辑区域内的编辑事件，获取编辑后的目标语种句子，基于编辑后的目标语种句子和语料句子对中的待翻译语种句子，生成翻译语料。

[0171] 如果用户仅在第二编辑区域内对目标语种句子进行编辑，则计算机设备可以获取到编辑后的目标语种句子，从而基于编辑后的目标语种句子和未编辑的待翻译语种句子生成翻译语料。例如，待翻译语种句子为维语句子，目标语种句子为汉语句子，如果用户认为汉语句子的内容需要调整，则可以在第二编辑区域内对汉语句子的内容进行编辑，得到编辑后的汉语句子，与原始的维语句子生成翻译语料。

[0172] 第三种方式、响应于在该第一编辑区域和该第二编辑区域内的编辑事件，获取编辑后的待翻译语种句子和编辑后的目标语种句子，基于编辑后的待翻译语种句子和编辑后的目标语种句子，生成翻译语料。

[0173] 如果用户既在第一编辑区域内对待翻译语种句子进行编辑，又在第二编辑区域内对目标语种句子进行编辑，则计算机设备可以获取到编辑后的待翻译语种句子和编辑后的目标语种句子，从而基于编辑后的待翻译语种句子和编辑后的目标语种句子生成翻译语料。例如，待翻译语种句子为维语句子，目标语种句子为汉语句子，如果用户认为维语句子和汉语句子的内容均需要调整，则可以分别在第一编辑区域内对维语句子的内容进行编辑，在第二编辑区域内对汉语句子的内容进行编辑，得到编辑后的维语句子本和汉语句子，将其作为生成的翻译语料。

[0174] 上述三种方式使得用户可以仅对某一语种的句子进行编辑，也可以对不同语种的句子均进行编辑，提高了灵活性。

[0175] 在一种可能实现方式中，计算机设备响应于对语料句子对的编辑事件，生成待翻译语种与目标语种之间的翻译语料之前，本申请实施例提供的方法还包括下述至少一项：响应于第一粘贴事件，在该第一编辑区域内输入语料句子对中的待翻译语种句子；响应于第二粘贴事件，在该第二编辑区域内输入语料句子对中的目标语种句子。通过在检测到粘贴事件时，响应于粘贴事件，在编辑区域内输入对应的语种句子，使得用户可以进一步对句子进行编辑。

[0176] 其中，第一粘贴事件为将待翻译语种句子粘贴到第一编辑区域内的事件，第二粘贴事件为将目标语种句子粘贴到第二编辑区域内的事件。

[0177] 计算机设备在显示语料句子对时，用户可以对语料句子对中的待翻译语种句子进

行复制操作，然后在第一编辑区域内进行粘贴操作，触发该第一粘贴事件，计算机设备在检测到该第一粘贴事件时，作为响应，可以在第一编辑区域内输入待翻译语种句子，如在图4中的区域403内输入维语句子，如图5所示，计算机设备可以在区域503内对维语句子分行显示。用户可以对语料句子对中的目标语种句子进行复制操作，然后在第二编辑区域内进行粘贴操作，触发该第二粘贴事件，计算机设备在检测到该第二粘贴事件时，作为响应，可以在第二编辑区域内输入目标语种句子，如在图4中的区域404内输入汉语句子，如图5所示，计算机设备可以在区域504内对汉语句子分行显示。

[0178] 在另一种可能实现方式中，计算机设备响应于对该语料句子对的编辑事件，生成该待翻译语种与该目标语种之间的翻译语料之前，本申请实施例提供的方法还包括：对该语料生成界面进行光学字符识别，得到该语料句子对中待翻译语种句子的文本坐标和目标语种句子；基于该语料句子对中待翻译语种句子的文本坐标和目标语种句子的文本坐标，获取该语料句子对中待翻译语种句子的内容和目标语种句子的内容；基于该语料句子对中待翻译语种句子的内容和目标语种句子的内容，在该语料生成界面的该第一编辑区域内生成该语料句子对中的待翻译语种句子，在该语料生成界面的该第二编辑区域内生成该语料句子对中的目标语种句子。

[0179] 计算机设备可以使用OCR(Optical Character Recognition，光学字符识别)技术获取待翻译语种句子和该目标语种句子的文本坐标，基于该文本坐标获取具体的内容，从而直接在编辑区域内生成对应的句子，这样可以取消复制粘贴环节，简化了用户操作。

[0180] 在一种可能实现方式中，计算机设备响应于对该语料句子对的编辑事件，生成该待翻译语种与该目标语种之间的翻译语料之后，本申请实施例提供的方法还包括：基于该翻译语料进行训练，得到翻译模型，该翻译模型用于将待翻译语种的内容文本翻译为目标语种的内容文本。

[0181] 计算机设备在生成翻译语料后，由于该翻译语料包括不同语种的句子，因而计算机设备可以基于该翻译语料，对翻译模型进行训练，使得基于不同语种的句子得到的翻译模型可以将待翻译语种的句子翻译为目标语种的句子。

[0182] 上述技术方案通过内容相似度爬取不同语种的内容文本，并对内容文本再进行切句，逐句比对后将符合翻译内容的句子保留，组成平行语料，推送到前端显示，由人工进行判断，对不同语种的句子进行对应匹配，对于翻译不准确的内容进行微调即可得到最终语料，可以降低人工对语种(如维语)掌握能力的要求，同时，人工无需做大面积的调整动作，人工标注效率得到有效提升。

[0183] 本申请实施例提供的技术方案可以应用于基于维汉翻译的深度学习中。针对翻译领域，相关技术依靠人工手动进行翻译存在输入效率低，或使用机器翻译(其他翻译工具)又会存在翻译结果生硬，准确度差。本申请可以通过维语网站获取维语新闻，同时找到对应的汉语新闻，通过文章标题与内容进行比对是否为同篇文章，再对内容进行切句，逐句比对后删除其中的无用信息，消除人工检验多次无用信息的障碍，避免文章翻译过程中因多句或少句带来翻译语料不对称的问题，然后直接提供语料句子对，人工只需对语料句子对中的维语句子和汉语句子进行比对，并对结果进行微调即可，可以解决人工纯手动输入全部翻译内容的问题以及语言翻译中经常存在的生硬翻译问题，提升翻译准确度，提升机器学习中人工标注维语语料的效率，可将相关技术中的人工输入效率30条/h，提升至200条/h，

具体提升效果可以根据网站翻译质量具体评估。

[0184] 本申请实施例提供的方法，通过获取与待翻译语种的第一内容文本为不同语种且内容相似度大于第一目标阈值的第二内容文本，然后分别对内容文本进行分句处理，得到待翻译语种句子和目标语种句子，从中筛选内容相似度大于第二目标阈值的待翻译语种句子和目标语种句子，得到语料句子对，使得用户可以对语料句子对进行编辑，从而响应于对语料句子对的编辑事件，生成待翻译语种与该目标语种之间的翻译语料。上述技术方案直接向用户提供不同语种且内容相似度较高的待翻译语种句子和目标语种句子，使得用户只需对两者进行比对，并进行微调即可得到不同语种之间的翻译语料，提高了翻译语料生成的效率。

[0185] 图6是本申请实施例提供的一种翻译语料生成装置的结构示意图。参照图6，该装置包括：

[0186] 获取模块601，用于获取待翻译语种的第一内容文本；

[0187] 确定模块602，用于确定与该第一内容文本的内容相似度大于第一目标阈值的第二内容文本，该第二内容文本为目标语种，该待翻译语种与该目标语种为不同语种；

[0188] 处理模块603，用于对该第一内容文本和该第二内容文本进行分句处理，得到待翻译语种句子和目标语种句子；

[0189] 筛选模块604，用于从该待翻译语种句子和该目标语种句子中，筛选内容相似度大于第二目标阈值的待翻译语种句子和目标语种句子，得到语料句子对；

[0190] 生成模块605，用于响应于对该语料句子对的编辑事件，生成该待翻译语种与该目标语种之间的翻译语料。

[0191] 在一种可能实现方式中，该筛选模块604用于：

[0192] 从该待翻译语种句子和该目标语种句子中，确定顺序对应的待翻译语种句子和目标语种句子，得到待选句子对；

[0193] 对于任一待选句子对，获取该任一待选句子对中的待翻译语种句子和目标语种句子的内容相似度；

[0194] 响应于该任一待选句子对中的待翻译语种句子和目标语种句子的内容相似度大于该第二目标阈值，将该任一待选句子对作为该语料句子对。

[0195] 在一种可能实现方式中，该筛选模块604用于：

[0196] 获取该任一待选句子对中的待翻译语种句子的机器翻译结果，该机器翻译结果为该目标语种；

[0197] 根据该机器翻译结果和该待选句子对中的目标语种句子，获取该任一待选句子对中的待翻译语种句子和目标语种句子的内容相似度。

[0198] 在一种可能实现方式中，该筛选模块604用于：

[0199] 从该待翻译语种句子和该目标语种句子中，筛选内容相似度大于第二目标阈值的待翻译语种句子和目标语种句子，得到候选句子对；

[0200] 从该候选句子对中确定内容相似度最大的句子对，将该内容相似度最大的句子对作为该语料句子对。

[0201] 在一种可能实现方式中，该装置还包括：

[0202] 显示模块，用于在语料生成界面中显示该语料句子对，该语料生成界面中设置有

第一编辑区域和第二编辑区域，该第一编辑区域用于对该语料句子对中的待翻译语种句子进行编辑，该第二编辑区域用于对该语料句子对中的目标语种句子进行编辑。

[0203] 在一种可能实现方式中，该显示模块用于：

[0204] 在该语料生成界面中对该语料句子对中的待翻译语种句子分行显示，不同待翻译语种句子位于不同行；

[0205] 在该语料生成界面中对该语料句子对中的目标语种句子分行显示，不同目标语种句子位于不同行。

[0206] 在一种可能实现方式中，该生成模块605用于执行下述任一项：

[0207] 响应于在该第一编辑区域内的编辑事件，获取编辑后的待翻译语种句子，基于该编辑后的待翻译语种句子和该语料句子对中的目标语种句子，生成该翻译语料；或，

[0208] 响应于在该第二编辑区域内的编辑事件，获取编辑后的目标语种句子，基于该编辑后的目标语种句子和该语料句子对中的待翻译语种句子，生成该翻译语料；或，

[0209] 响应于在该第一编辑区域和该第二编辑区域内的编辑事件，获取编辑后的待翻译语种句子和编辑后的目标语种句子，基于该编辑后的待翻译语种句子和该编辑后的目标语种句子，生成该翻译语料。

[0210] 在一种可能实现方式中，该装置还包括下述至少一项：

[0211] 第一输入模块，用于响应于第一粘贴事件，在该第一编辑区域内输入该语料句子对中的待翻译语种句子；

[0212] 第二输入模块，用于响应于第二粘贴事件，在该第二编辑区域内输入该语料句子对中的目标语种句子。

[0213] 在一种可能实现方式中，该装置还包括：

[0214] 识别模块，用于对该语料生成界面进行光学字符识别，得到该语料句子对中待翻译语种句子的文本坐标和目标语种句子的文本坐标；

[0215] 该获取模块601还用于基于该语料句子对中待翻译语种句子的文本坐标和目标语种句子的文本坐标，获取该语料句子对中待翻译语种句子的内容和目标语种句子的内容；

[0216] 生成模块605还用于基于该语料句子对中待翻译语种句子的内容和目标语种句子的内容，在该语料生成界面的该第一编辑区域内生成该语料句子对中的待翻译语种句子，在该语料生成界面的该第二编辑区域内生成该语料句子对中的目标语种句子。

[0217] 在一种可能实现方式中，该确定模块602用于：

[0218] 根据该第一内容文本的标题和正文中部分字符，生成该第一内容文本的摘要；

[0219] 获取该第一内容文本的摘要所对应目标语种的机器翻译结果；

[0220] 根据该机器翻译结果搜索该目标语种的搜索结果，从该搜索结果中选取待选内容文本；

[0221] 从该待选内容文本中，选取与该第一内容文本的内容相似度大于该第一目标阈值的内容文本作为该第二内容文本。

[0222] 在一种可能实现方式中，该确定模块602用于：

[0223] 获取该第一内容文本的标题所对应目标语种的机器翻译结果；

[0224] 对于任一待选内容文本，根据该第一内容文本的标题所对应目标语种的机器翻译结果、该第一内容文本的摘要所对应目标语种的机器翻译结果、该任一待选内容文本的标

题和摘要,获取该任一待选内容文本与该第一内容文本的标题相似度和摘要相似度;

[0225] 对该标题相似度和摘要相似度进行加权求和,得到该任一待选内容文本与该第一内容文本的内容相似度;

[0226] 当该任一待选内容文本与该第一内容文本的内容相似度大于该第一目标阈值时,将该述任一待选内容文本作为该第二内容文本。

[0227] 在一种可能实现方式中,该获取模块601用于:

[0228] 根据起始统一资源定位符URL,迭代爬取该起始URL对应的起始页面上的至少一个URL,该起始页面上的内容文本为该待翻译语种;

[0229] 对于当前爬取到的URL,对该当前爬取到的URL进行解析,得到该当前爬取到的URL对应的内容文本作为该第一内容文本。

[0230] 在一种可能实现方式中,该显示模块还用于:

[0231] 在显示该语料句子对时,显示该语料句子对中的待翻译语种句子的机器翻译结果,该机器翻译结果为该目标语种。

[0232] 在一种可能实现方式中,该装置还包括:

[0233] 训练模块,用于基于该翻译语料进行训练,得到翻译模型,该翻译模型用于将该待翻译语种的内容文本翻译为该目标语种的内容文本。

[0234] 在一种可能实现方式中,该装置还包括:

[0235] 需要说明的是:上述实施例提供的翻译语料生成装置在语料生成时,仅以上述各功能模块的划分进行举例说明,实际应用中,可以根据需要而将上述功能分配由不同的功能模块完成,即将设备的内部结构划分成不同的功能模块,以完成以上描述的全部或者部分功能。另外,上述实施例提供的翻译语料生成装置与翻译语料生成方法实施例属于同一构思,其具体实现过程详见方法实施例,这里不再赘述。

[0236] 上述实施例中的计算机设备可以为终端。

[0237] 图7是本申请实施例提供的一种终端700的结构示意图。该终端700可以是: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700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。

[0238] 通常,终端700包括有:一个或多个处理器701和一个或多个存储器702。

[0239] 处理器701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701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701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701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701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。

[0240] 存储器702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702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702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701所执行以实现本申请中方法实施例提供的翻译语料生成方法。

[0241] 在一些实施例中，终端700还可选包括有：外围设备接口703和至少一个外围设备。处理器701、存储器702和外围设备接口703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703相连。具体地，外围设备包括：射频电路704、显示屏705、摄像头组件706、音频电路707、定位组件708和电源709中的至少一种。

[0242] 外围设备接口703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701和存储器702。在一些实施例中，处理器701、存储器702和外围设备接口703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701、存储器702和外围设备接口703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。

[0243] 射频电路704用于接收和发射RF(Radio Frequency，射频)信号，也称电磁信号。射频电路704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704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704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704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：城域网、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704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。

[0244] 显示屏705用于显示UI(UserInterface，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705是触摸显示屏时，显示屏705还具有采集在显示屏705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701进行处理。此时，显示屏705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705可以为一个，设置终端700的前面板；在另一些实施例中，显示屏705可以为至少两个，分别设置在终端700的不同表面或呈折叠设计；在再一些实施例中，显示屏705可以是柔性显示屏，设置在终端700的弯曲表面上或折叠面上。甚至，显示屏705还可以设置成非矩形的不规则图形，也即异形屏。显示屏705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。

[0245] 摄像头组件706用于采集图像或视频。可选地，摄像头组件706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706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。

[0246] 音频电路707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701进行处理，或者输入至射频电路704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700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701或射频电路704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707还可以包括耳机插孔。

[0247] 定位组件708用于定位终端700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708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、俄罗斯的格雷纳斯系统或欧盟的伽利略系统的定位组件。

[0248] 电源709用于为终端700中的各个组件进行供电。电源709可以是交流电、直流电、一次性电池或可充电电池。当电源709包括可充电电池时，该可充电电池可以支持有线充电或无线充电。该可充电电池还可以用于支持快充技术。

[0249] 在一些实施例中，终端700还包括有一个或多个传感器710。该一个或多个传感器710包括但不限于：加速度传感器711、陀螺仪传感器712、压力传感器713、指纹传感器714、光学传感器715以及接近传感器716。

[0250] 加速度传感器711可以检测以终端700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711可以用于检测重力加速度在三个坐标轴上的分量。处理器701可以根据加速度传感器711采集的重力加速度信号，控制显示屏705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711还可以用于游戏或者用户的运动数据的采集。

[0251] 陀螺仪传感器712可以检测终端700的机体方向及转动角度，陀螺仪传感器712可以与加速度传感器711协同采集用户对终端700的3D动作。处理器701根据陀螺仪传感器712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。

[0252] 压力传感器713可以设置在终端700的侧边框和/或显示屏705的下层。当压力传感器713设置在终端700的侧边框时，可以检测用户对终端700的握持信号，由处理器701根据压力传感器713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713设置在显示屏705的下层时，由处理器701根据用户对显示屏705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。

[0253] 指纹传感器714用于采集用户的指纹，由处理器701根据指纹传感器714采集到的指纹识别用户的身份，或者，由指纹传感器714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701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714可以被设置终端700的正面、背面或侧面。当终端700上设置有物理按键或厂商Logo时，指纹传感器714可以

与物理按键或厂商Logo集成在一起。

[0254] 光学传感器715用于采集环境光强度。在一个实施例中，处理器701可以根据光学传感器715采集的环境光强度，控制显示屏705的显示亮度。具体地，当环境光强度较高时，调高显示屏705的显示亮度；当环境光强度较低时，调低显示屏705的显示亮度。在另一个实施例中，处理器701还可以根据光学传感器715采集的环境光强度，动态调整摄像头组件706的拍摄参数。

[0255] 接近传感器716，也称距离传感器，通常设置在终端700的前面板。接近传感器716用于采集用户与终端700的正面之间的距离。在一个实施例中，当接近传感器716检测到用户与终端700的正面之间的距离逐渐变小时，由处理器701控制显示屏705从亮屏状态切换为息屏状态；当接近传感器716检测到用户与终端700的正面之间的距离逐渐变大时，由处理器701控制显示屏705从息屏状态切换为亮屏状态。

[0256] 本领域技术人员可以理解，图7中示出的结构并不构成对终端700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。

[0257] 上述实施例中的计算机设备可以为服务器。

[0258] 图8是本申请实施例提供的一种服务器800的结构示意图，该服务器800可因配置或性能不同而产生比较大的差异，可以包括一个或多个处理器(Central Processing Units，CPU)801和一个或多个存储器802，其中，该存储器802中存储有至少一条程序代码，该至少一条程序代码由该处理器801加载并执行以实现上述各个方法实施例提供的方法。当然，该服务器还可以具有有线或无线网络接口、键盘以及输入输出接口等部件，以便进行输入输出，该服务器还可以包括其他用于实现设备功能的部件，在此不做赘述。

[0259] 在示例性实施例中，还提供了一种存储有至少一条程序代码的计算机可读存储介质，例如存储有至少一条程序代码的存储器，上述至少一条程序代码由处理器加载并执行，以实现上述实施例中的翻译语料生成方法。例如，该计算机可读存储介质可以是只读内存(Read-Only Memory，ROM)、随机存取存储器(Random Access Memory，RAM)、只读光盘(Compact Disc Read-Only Memory，CD-ROM)、磁带、软盘和光数据存储设备等。

[0260] 本领域普通技术人员可以理解实现上述实施例的全部或部分步骤可以通过硬件来完成，也可以通过程序指令相关的硬件完成，该程序可以存储于一种计算机可读存储介质中，上述提到的存储介质可以是只读存储器，磁盘或光盘等。

[0261] 以上仅为本申请的可选实施例，并不用以限制本申请，凡在本申请的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本申请的保护范围之内。

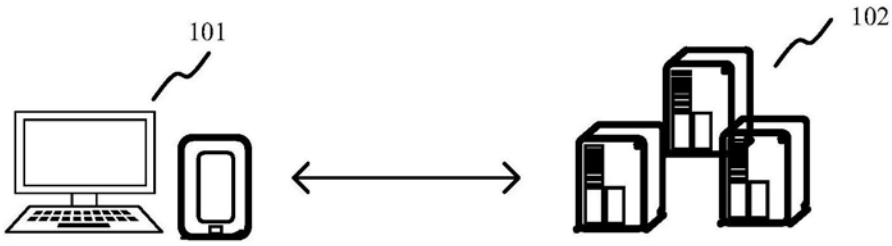


图1

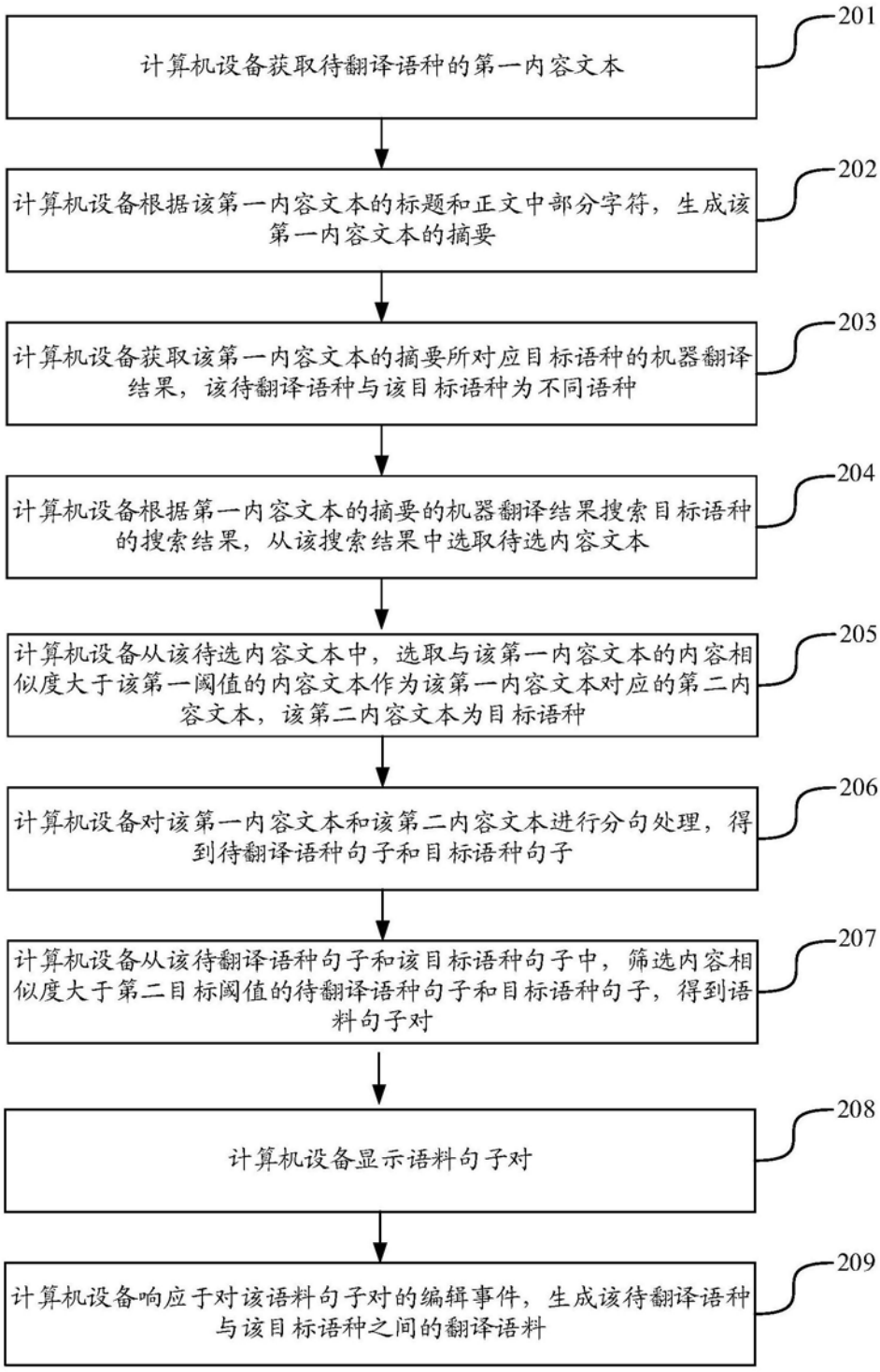

201
计算机设备获取待翻译语种的第一内容文本
202
计算机设备根据该第一内容文本的标题和正文中部分字符，生成该第一内容文本的摘要
203
计算机设备获取该第一内容文本的摘要所对应目标语种的机器翻译结果，该待翻译语种与该目标语种为不同语种
204
计算机设备根据第一内容文本的摘要的机器翻译结果搜索目标语种的搜索结果，从该搜索结果中选取待选内容文本
205
计算机设备从该待选内容文本中，选取与该第一内容文本的内容相似度大于该第一阈值的内容文本作为该第一内容文本对应的第二内容文本，该第二内容文本为目标语种
206
计算机设备对该第一内容文本和该第二内容文本进行分句处理，得到待翻译语种句子和目标语种句子
207
计算机设备从该待翻译语种句子和该目标语种句子中，筛选内容相似度大于第二目标阈值的待翻译语种句子和目标语种句子，得到语料句子对
208
计算机设备显示语料句子对
209
计算机设备响应于对该语料句子对的编辑事件，生成该待翻译语种与该目标语种之间的翻译语料

图2

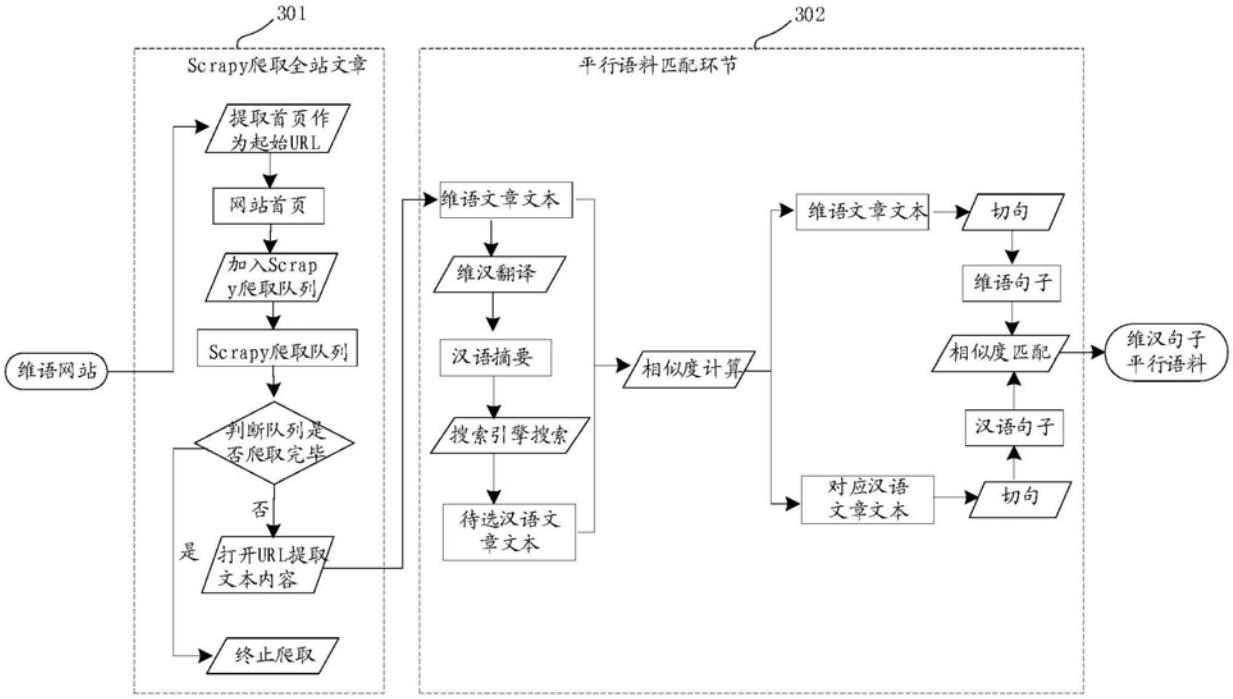

301
Scrapy爬取全站文章
提取首页作为起始URL
网站首页
加入Scrapy爬取队列
Scrapy爬取队列
判断队列是否爬取完毕
否
是
打开URL提取文本内容
终止爬取
维语网站
302
平行语料匹配环节
维语文章文本
维汉翻译
汉语摘要
搜索引擎搜索
待选汉语文章文本
相似度计算
维语文章文本
切句
维语句子
相似度匹配
维汉句子平行语料
汉语句子
对应汉语文章文本
切句

图3


维语内容：
机翻结果：阿克苏市行政服务中心党组书记曹辉说，我们结合群众反映比较强烈的热点问题，及时纠正，不动产一般登记法定30个工作日至10个工作日，企业登记3个工作日，食品销售业务审批事项由20个工作日压缩至1个工作日，
相似内容：阿克苏市行政服务中心党组书记曹慧说：“结合群众反映比较强烈的热点问题，我们及时进行了整改，把不动产一般登记由法定的30个工作日缩短为10个工作日，企业登记压缩到3个工作日，食品销售行业审批事项由20个工作日压缩至1个工作日
401
405
402
维语输入区
汉语输入区
阿克苏市行政服务中心党组书记曹辉说，我们结合群众反映比较强烈的热点问题，及时纠正，不动产一般登记法定30个工作日至10个工作日，企业登记3个工作日，食品销售业务审批事项由20个工作日压缩至1个工作日，
403
404

图4


维语内容：
机器结果：新华社北京10月21日电 中国选手张正21日在赛后庆祝。
当天,在第七届世界军人运动会军事五项男子全能障碍游泳比赛中,中国选手张正以23秒8的成绩获得第一名,并与世界纪录持平。
相似内容：新华社武汉10月21日电（记者侯文坤）中国选手张政在21日结束的军事五项男子个人全能障碍游泳比赛中，以23秒8的成绩排名第一，并追平了该项目的世界纪录。
10月21日，中国选手张政在比赛后庆祝。。
501
505
502
维语输入区
汉语输入区
新华社北京10月21日电 中国选手张正21日在赛后庆祝。
当天,在第七届世界军人运动会军事五项男子全能障碍游泳比赛中,中国选手张正以23秒8的成绩获得第一名,并与世界纪录持平。
503
504

图5

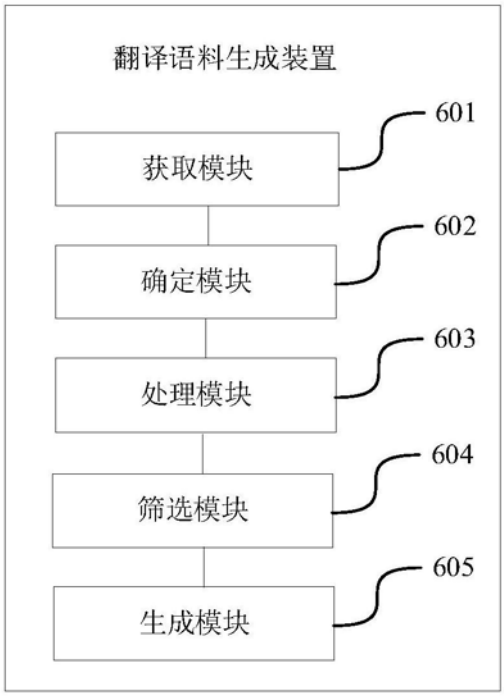

翻译语料生成装置
获取模块
601
确定模块
602
处理模块
603
筛选模块
604
生成模块
605

图6

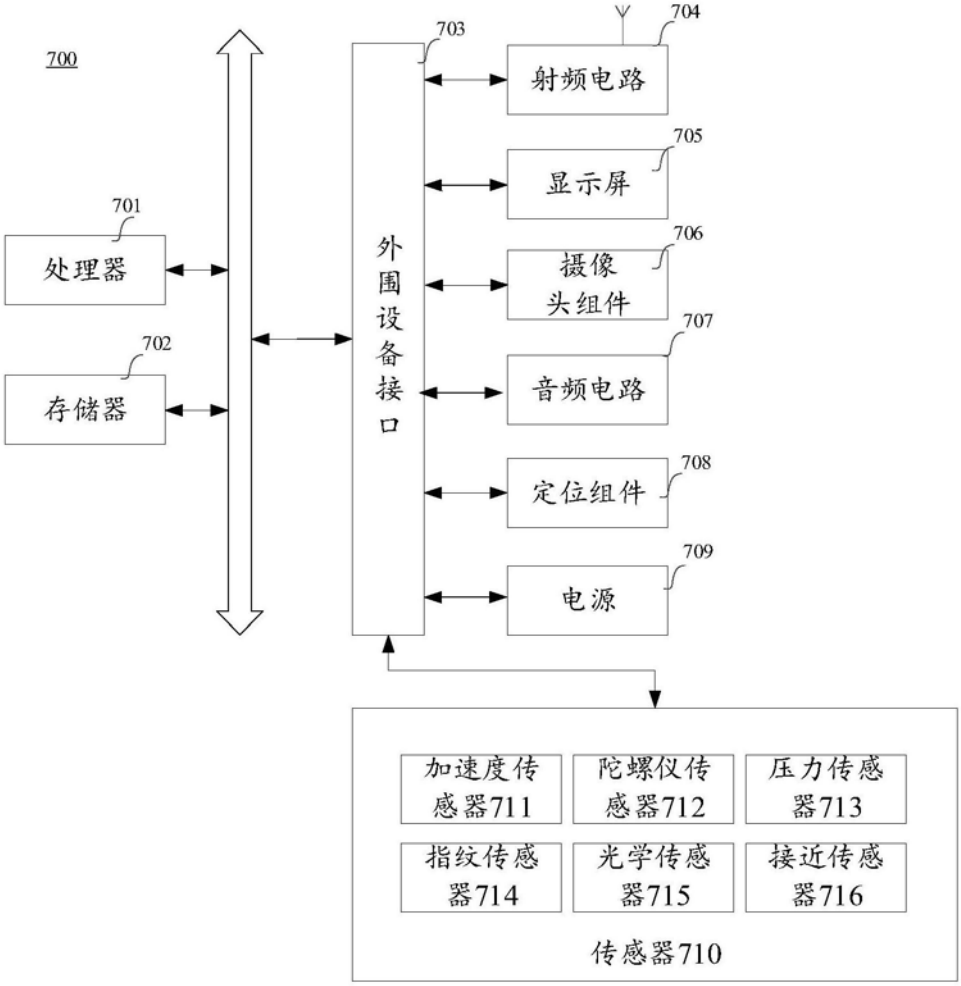

700
701
处理器
702
存储器
703
外围设备接口
704
射频电路
705
显示屏
706
摄像头组件
707
音频电路
708
定位组件
709
电源
加速度传感器711
陀螺仪传感器712
压力传感器713
指纹传感器714
光学传感器715
接近传感器716
传感器710

图7

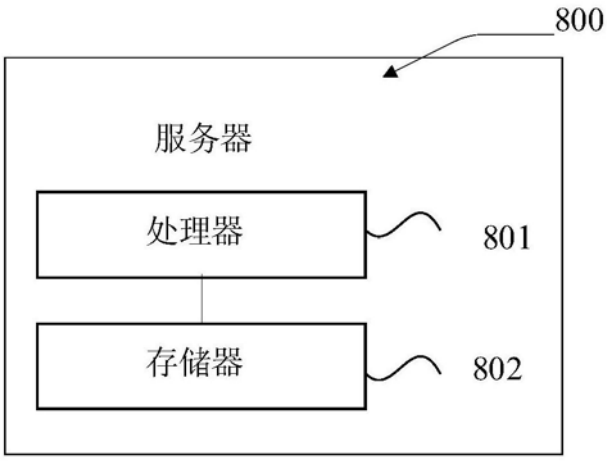

800
服务器
处理器
801
存储器
802

图8