(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 114417898 A
(43)申请公布日 2022.04.29

(21)申请号 202210058761.7

(22)申请日 2022.01.18

(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 李尔楠 熊明钧 孟凡东 周杰

(74)专利代理机构 广州三环专利商标代理有限公司 44202
代理人 杜维

(51)Int.Cl.
*G06F 40/58*(2020.01)
*G06F 40/279*(2020.01)
*G06N 3/04*(2006.01)
*G06N 3/08*(2006.01)

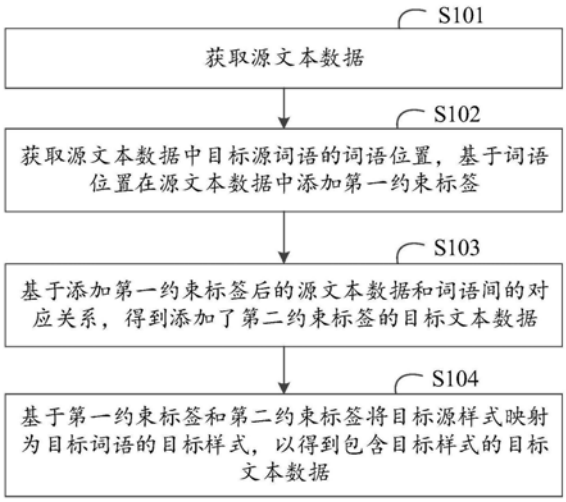
权利要求书3页 说明书20页 附图5页

(54)发明名称

# 数据处理方法、装置、设备及可读存储介质

(57)摘要

本申请实施例公开了一种数据处理方法、装置、设备及可读存储介质，涉及人工智能领域，其中，方法包括：获取富样式的源文本数据，源文本数据为富样式文本数据；获取源文本数据中目标源词语的词语位置，基于词语位置在源文本数据中添加第一约束标签，目标源词语的样式为目标源样式；基于添加第一约束标签后的源文本数据和词语间的对应关系，得到添加了第二约束标签的目标文本数据，第二约束标签是在目标文本数据中目标词语的词语位置添加的；基于第一约束标签和第二约束标签将目标源样式映射为第二约束标签对应的目标词语的目标样式，以得到包含目标样式的目标文本数据。采用本申请实施例，可以使得文本翻译更完整，提高数据处理准确性。




CN 114417898 A

1.一种数据处理方法,其特征在于,包括:

获取源文本数据,所述源文本数据为富样式文本数据;

获取所述源文本数据中目标源词语的词语位置,基于所述目标源词语的词语位置在所述源文本数据中添加第一约束标签,所述目标源词语的样式为目标源样式;

基于添加第一约束标签后的源文本数据和词语间的对应关系,得到添加了第二约束标签的目标文本数据,所述第二约束标签是在所述目标文本数据中目标词语的词语位置添加的,所述目标词语为与所述目标源词语对应的词语,所述目标文本数据是对所述源文本数据进行翻译得到的;

基于所述第一约束标签和所述第二约束标签将所述目标源样式映射为所述目标词语的目标样式,以得到包含所述目标样式的目标文本数据。

2.根据权利要求1所述的方法,其特征在于,所述获取源文本数据之前,所述方法还包括:

获取第一样本数据和第二样本数据,所述第二样本数据是对所述第一样本数据进行翻译得到的,所述第一样本数据为富样式文本数据;

对所述第一样本数据和所述第二样本数据进行对齐处理,以确定所述第一样本数据中的样本词语与所述第二样本数据中的样本词语之间的样本对应关系;

获取所述第一样本数据中第一样本词语的样本词语位置,基于所述样本词语位置在所述第一样本数据中添加第一样本约束标签,所述第一样本词语的样式为第一样式;

获取参考样本,基于添加第一样本约束标签后的第一样本数据、所述样本对应关系以及所述参考样本,训练得到目标处理模型;

所述基于添加第一约束标签后的源文本数据和词语间的对应关系,得到添加了第二约束标签的目标文本数据,包括:

采用所述目标处理模型基于添加第一约束标签后的源文本数据和所述词语间的对应关系,得到添加了第二约束标签的目标文本数据。

3.根据权利要求2所述的方法,其特征在于,所述对所述第一样本数据和所述第二样本数据进行对齐处理,以确定所述第一样本数据中的样本词语与所述第二样本数据中的样本词语之间的样本对应关系,包括:

对所述第一样本数据和第二样本数据进行分词处理,得到N个第一单词和M个第二单词之间的第一对应关系,第一单词为所述第一样本数据中的单词,第二单词为所述第二样本数据中的单词,N、M均为正整数;

基于所述N个第一单词和所述M个第二单词之间的第一对应关系,确定i个第一词语与j个第二词语之间的第二对应关系,第一词语为所述N个第一单词中的至少一个单词组成的词语,第二词语为所述M个第二单词中的至少一个单词组成的词语,i、j均为正整数;

基于所述第二对应关系确定所述第一样本数据和所述第二样本数据是否对齐,若所述第一样本数据和所述第二样本数据对齐,则将对齐关系作为所述源文本数据中的样本词语与所述目标文本数据中的样本词语之间的样本对应关系。

4.根据权利要求3所述的方法,其特征在于,所述基于所述第二对应关系确定所述第一样本数据和所述第二样本数据是否对齐,包括:

采用对齐一致性原则确定第一目标词语和第二目标词语是否对齐,所述第一目标词语

为所述i个第一词语中的任意一个，所述第二目标词语为所述j个第二词语中的任意一个，所述对齐一致性原则用于指示所述第一样本数据中的多个连续词语组成的词语与所述第二样本数据中的多个连续词语组成的词语是否对应；

若所述第一目标词语和所述第二目标词语对齐，则确定所述第一样本数据和所述第二样本数据对齐；

若所述第一目标词语和所述第二目标词语未对齐，则确定所述第一样本数据和所述第二样本数据未对齐。

5.根据权利要求4所述的方法，其特征在于，所述方法还包括：

若所述第一目标词语和所述第二目标词语对齐，则获取所述第一样本数据中的第一关键词语，以及第二样本数据中的第二关键词语；

对所述第一关键词语和所述第二关键词语进行匹配；

若所述第一关键词语和所述第二关键词语匹配，则执行确定所述第一样本数据和所述第二样本数据对齐的步骤。

6.根据权利要求4所述的方法，其特征在于，所述方法还包括：

若所述第一目标词语和所述第二目标词语对齐，则检测所述样本数据中是否存在对空词；

若所述样本数据中存在对空词，则分别将所述对空词加入第一邻居词语和第二邻居词语中，得到第一组合词语和第二组合词语，所述第一邻居词语和所述第二邻居词语为所述样本数据中与所述对空词相邻的两个词语，所述样本数据为所述第一样本数据和所述第二样本数据中的任意一个；

从所述第一组合词语或所述第二组合词语中确定目标组合词语；

基于所述目标组合词语对所述第一样本数据中的样本词语和所述第二样本数据中的样本词语之间的样本对应关系进行调整。

7.根据权利要求1所述的方法，其特征在于，所述方法还包括：

基于所述目标词语的目标样式对所述目标文本数据进行渲染，以得到渲染后的包含目标样式的目标文本数据；

调用关联的用户终端输出所述渲染后的包含目标样式的目标文本数据。

8.一种数据处理装置，其特征在于，包括：

文本获取模块，用于获取源文本数据所述源文本数据为富样式文本数据；

第一添加模块，用于获取所述源文本数据中目标源词语的词语位置，基于所述词语位置在所述源文本数据中添加第一约束标签，所述目标源词语的样式为目标源样式；

第二添加模块，用于基于添加第一约束标签后的源文本数据和词语间的对应关系，得到添加了第二约束标签的目标文本数据，所述目标文本数据是对所述源文本数据进行翻译得到的，所述第二约束标签是在所述目标文本数据中目标词语的词语位置添加的，所述目标词语为与所述目标源词语对应的词语；

数据映射模块，用于基于所述第一约束标签和所述第二约束标签将所述目标源样式映射为所述目标词语的目标样式，以得到包含所述目标样式的目标文本数据。

9.一种计算机设备，其特征在于，包括：处理器、存储器以及网络接口；

所述处理器与所述存储器、所述网络接口相连，其中，所述网络接口用于提供数据通信

功能，所述存储器用于存储程序代码，所述处理器用于调用所述程序代码，以使得所述计算机设备执行权利要求1-7任一项所述的方法。

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存储有计算机程序，所述计算机程序适于由处理器加载并执行，以使得具有所述处理器的计算机设备执行权利要求1-7任一项所述的方法。

# 数据处理方法、装置、设备及可读存储介质

## 技术领域

[0001] 本申请涉及人工智能中的自然语言处理技术领域，尤其涉及一种数据处理方法、装置、设备及可读存储介质。

## 背景技术

[0002] 文本翻译已经被广泛应用于各个领域中，并在一定程度上取得了很大的进展。然而对于富样式文本，即文本中的某些句子或者词语包含多种格式，例如包含不同的颜色、文字加粗、下划线等格式。在对这类富样式文本进行翻译时，目前只能实现对文本中的文字内容进行翻译，不能实现对文本中格式的完整保留，从而降低了文本翻译的完整性，导致数据处理准确性较低。

## 发明内容

[0003] 本申请实施例提供一种数据处理方法、装置、设备及可读存储介质，可以使得文本翻译更完整，提高数据处理准确性。

[0004] 第一方面，本申请提供一种数据处理方法，包括：

[0005] 获取源文本数据，源文本数据为富样式文本数据；

[0006] 获取源文本数据中目标源词语的词语位置，基于该目标源词语的词语位置在源文本数据中添加第一约束标签，该目标源词语的样式为目标源样式；

[0007] 基于添加第一约束标签后的源文本数据和词语间的对应关系，得到添加了第二约束标签的目标文本数据，目标文本数据是对源文本数据进行翻译得到的，该第二约束标签是在该目标文本数据中目标词语的词语位置添加的，该目标词语为与该目标源词语对应的词语；

[0008] 基于该第一约束标签和该第二约束标签将该目标源样式映射为该第二约束标签对应的目标词语的目标样式，以得到包含该目标样式的目标文本数据。

[0009] 第二方面，本申请提供一种数据处理装置，包括：

[0010] 文本获取模块，用于获取源文本数据，源文本数据为富样式文本数据；

[0011] 第一添加模块，用于获取源文本数据中目标源词语的词语位置，基于该目标源词语的词语位置在源文本数据中添加第一约束标签，该目标源词语的样式为目标源样式；

[0012] 第二添加模块，用于基于添加第一约束标签后的源文本数据和词语间的对应关系，得到添加了第二约束标签的目标文本数据，目标文本数据是对源文本数据进行翻译得到的，该第二约束标签是在该目标文本数据中目标词语的词语位置添加的，该目标词语为与该目标源词语对应的词语；

[0013] 数据映射模块，用于基于该第一约束标签和该第二约束标签将该目标源样式映射为该目标词语的目标样式，以得到包含该目标样式的目标文本数据。

[0014] 第三方面，本申请提供了一种计算机设备，包括：处理器、存储器、网络接口；

[0015] 上述处理器与存储器、网络接口相连，其中，网络接口用于提供数据通信功能，上

述存储器用于存储计算机程序，上述处理器用于调用上述计算机程序，以使包含该处理器的计算机设备执行上述数据处理方法。

[0016] 第四方面，本申请提供了一种计算机可读存储介质，该计算机可读存储介质中存储有计算机程序，该计算机程序适于由处理器加载并执行，以使得具有该处理器的计算机设备执行上述数据处理方法。

[0017] 第五方面，本申请提供了一种计算机程序产品或计算机程序，该计算机程序产品或计算机程序包括计算机指令，该计算机指令存储在计算机可读存储介质中。计算机设备的处理器从计算机可读存储介质读取该计算机指令，处理器执行该计算机指令，使得该计算机设备执行本申请第一方面中的各种可选方式中提供的数据处理方法。

[0018] 本申请实施例中，通过获取富样式的源文本数据，以及获取源文本数据中具有目标源样式的目标源词语的词语位置，可以基于目标源词语的词语位置在源文本数据中添加第一约束标签；可以基于添加第一约束标签后的源文本数据和词语间的对应关系，得到添加了第二约束标签的目标文本数据，从而基于第一约束标签和第二约束标签将目标源样式映射为第二约束标签对应的目标词语的目标样式，以得到包含目标样式的目标文本数据。由于确定了源文本数据中具有源样式的目标源词语的位置信息，可以基于该目标源词语在源文本数据中的位置信息为源文本数据添加约束标签，从而基于添加约束标签后的源文本数据和词语间的对应关系得到添加了约束标签的目标文本数据，基于约束标签将源文本数据中的源样式映射为目标文本数据中的目标样式，可以使得翻译后的目标文本数据中包含目标样式，从而实现在文本翻译的同时保留文本中的样式，保证文本数据翻译的完整性，提高数据处理的准确性。

## 附图说明

[0019] 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。

[0020] 图1是本申请实施例提供的一种数据处理系统的架构示意图；

[0021] 图2是本申请实施例提供的一种数据处理方法的应用场景示意图；

[0022] 图3是本申请实施例提供的一种数据处理方法的流程示意图；

[0023] 图4是本申请实施例提供的另一种数据处理方法的流程示意图；

[0024] 图5是本申请实施例提供的一种词语对应关系的示意图；

[0025] 图6是本申请实施例提供的另一种词语对应关系的示意图；

[0026] 图7是本申请实施例提供的一种对空词划分的示意图；

[0027] 图8是本申请实施例提供的一种模型结构示意图；

[0028] 图9是本申请实施例提供的一种数据处理装置的组成结构示意图；

[0029] 图10是本申请实施例提供的一种计算机设备的组成结构示意图。

## 具体实施方式

[0030] 下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行清楚、完

整地描述，显然，所描述的实施例仅是本申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本申请保护的范围。

[0031] 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。人工智能软件技术主要包括计算机视觉技术、语音处理技术、自然语言处理技术以及机器学习/深度学习等几大方向。

[0032] 自然语言处理(Nature Language processing,NLP)是计算机科学领域与人工智能领域中的一个重要方向。它研究能实现人与计算机之间用自然语言进行有效通信的各种理论和方法。自然语言处理是一门融语言学、计算机科学、数学于一体的科学。因此，这一领域的研究将涉及自然语言，即人们日常使用的语言，所以它与语言学的研究有着密切的联系。自然语言处理技术通常包括文本处理、语义理解、机器翻译、机器人问答、知识图谱等技术。

[0033] 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、式教学习等技术。

[0034] 本申请实施例中涉及到用户信息相关的数据均为用户授权后的数据。本申请涉及人工智能领域的自然语言处理技术和机器学习技术。可选地，例如，可以利用自然语言处理技术对源文本数据进行翻译得到目标文本数据。进一步地，本申请也可以利用机器学习技术对源文本数据进行翻译，得到目标文本数据；还可以利用机器学习技术基于添加第一约束标签后的源文本数据和词语间的对应关系，得到添加了第二约束标签的目标文本数据，并将目标源样式映射为目标词语的目标样式，以得到包含目标样式的目标文本数据，等等。本申请技术方案可以用于对富样式的源文本数据进行翻译，得到富样式的目标文本数据的场景中，例如可以应用于双语教学中对富样式的文本数据进行翻译、语言测试中对富样式的文本数据进行翻译、对医疗领域中包含专业术语的富样式文本数据进行翻译，或者其他需要对富样式文本数据进行翻译的场景中。由于确定了源文本数据中具有源样式的目标源词语的位置信息，可以基于该目标源词语在源文本数据中的位置信息为源文本数据添加约束标签，从而基于添加约束标签后的源文本数据和词语间的对应关系得到添加了约束标签的目标文本数据，基于约束标签将源文本数据中的源样式映射为目标文本数据中的目标样式，可以使得翻译后的目标文本数据中包含目标样式，从而实现在文本翻译的同时保留文本中的样式，保证文本数据翻译的完整性，提高数据处理的准确性。

[0035] 其中，富样式文本数据是指文本数据中的某些句子或者词语包含多种格式，例如包含不同的颜色、文字加粗、下划线等格式。

[0036] 请参见图1，图1是本申请实施例提供的一种数据处理系统的架构示意图，如图1所示，计算机设备101可以与用户终端进行数据交互，用户终端的数量可以为一个或者多个(至少两个)。例如，当用户终端的数量为多个时，用户终端可以包括图1中的用户终端102a、

用户终端102b及用户终端102c等。其中，以用户终端102a为例，计算机设备101可以从用户终端102a中获取源文本数据。进一步地，计算机设备101可以获取源文本数据中目标源词语的词语位置，基于词语位置在源文本数据中添加第一约束标签，目标源词语的样式为目标源样式。进一步地，计算机设备101可以基于添加第一约束标签后的源文本数据和词语间的对应关系，得到添加了第二约束标签的目标文本数据，基于第一约束标签和第二约束标签将目标源样式映射为目标词语的目标样式，以得到包含目标样式的目标文本数据。可选地，计算机设备101可以将包含目标样式的目标文本数据发送至用户终端102a，以使用户终端102a对包含目标样式的目标文本数据进行渲染展示。

[0037] 由于确定了源文本数据中具有源样式的目标源词语的位置信息，可以基于该目标源词语在源文本数据中的位置信息为源文本数据添加约束标签，从而基于添加约束标签后的源文本数据和词语间的对应关系，得到添加了约束标签的目标文本数据，基于约束标签将源文本数据中的源样式映射为目标文本数据中的目标样式，可以使得翻译后的目标文本数据中包含目标样式，从而实现在文本翻译的同时保留文本中的样式，保证文本数据翻译的完整性，提高数据处理的准确性。

[0038] 可以理解的是，本申请实施例中所提及的计算机设备包括但不限于用户终端或服务器。换句话说，计算机设备可以是服务器或用户终端，也可以是服务器和用户终端组成的系统。其中，以上所提及的用户终端可以是一种电子设备，包括但不限于手机、平板电脑、台式电脑、笔记本电脑、掌上电脑、车载设备、智能语音交互设备、增强现实/虚拟现实(Augmented Reality/Virtual Reality，AR/VR)设备、头盔显示器、可穿戴设备、智能音箱、智能家电、飞行器、数码相机、摄像头及其他具备网络接入能力的移动互联网设备(mobile internet device，MID)等。其中，以上所提及的服务器可以是独立的物理服务器，也可以是多个物理服务器构成的服务器集群或者分布式系统，还可以是提供云服务、云数据库、云计算、云函数、云存储、网络服务、云通信、中间件服务、域名服务、安全服务、车路协同、内容分发网络(Content Delivery Network，CDN)、以及大数据和人工智能平台等基础云计算服务的云服务器。

[0039] 进一步地，请参见图2，图2是本申请实施例提供的一种数据处理方法的应用场景示意图。如图2所示，用户终端21可以发送源文本数据22至计算机设备23，其中，源文本数据22为富样式文本数据。例如，源文本数据为“明天去游乐园玩”，其中源文本数据中的词语“游乐园”为富样式词语，例如样式包括字体加粗和下划线。进一步地，计算机设备23可以获取源文本数据中的目标源词语的词语位置，例如“游乐园”在源文本数据中的词语位置，基于该词语位置在源文本数据中添加第一约束标签，添加第一约束标签后的源文本数据如24所示，其中，目标源词语“游乐园”的样式为目标源样式，例如目标源样式为字体加粗和下划线。进一步地，计算机设备23可以基于添加第一约束标签后的源文本数据和词语间的对应关系，得到添加了第二约束标签的目标文本数据，添加第二约束标签的目标文本数据如25所示，目标文本数据是对源文本数据22进行翻译得到的，基于第一约束标签和第二约束标签将目标源样式映射为目标词语的目标样式，以得到包含目标样式的目标文本数据26，其中，目标词语为与目标源词语对应的词语。可选地，计算机设备还可以将包含目标样式的目标文本数据26发送至用户终端21，以使用户终端21对包含目标样式的目标文本数据26进行渲染展示。

[0040] 进一步地，请参见图3，图3是本申请实施例提供的一种数据处理方法的流程示意图；如图3所示，该数据处理方法可以应用于计算机设备，该数据处理方法包括但不限于以下步骤：

[0041] S101，获取源文本数据。

[0042] 本申请实施例中，计算机设备可以从用户终端获取源文本数据，也可以从存储有源文本数据的文本数据库中获取源文本数据，还可以从其他路径获取源文本数据。其中，源文本数据为富样式文本数据。源文本数据可以包括但不限于教学领域的文本数据、医学领域的文本数据或者其他领域的文本数据。源文本数据可以包括但不限于中文、英文、韩文、德文或者其他语言数据。可以理解的是，在本申请的具体实施方式中，若源文本数据涉及到用户信息等相关的数据，则计算机设备获取到的源文本数据为用户授权后的数据，当本申请以上实施例运用到具体产品或技术中时，需要获得用户许可或者同意，且相关数据的收集、使用和处理需要遵守相关国家和地区的相关法律法规和标准。

[0043] S102，获取源文本数据中目标源词语的词语位置，基于词语位置在源文本数据中添加第一约束标签。

[0044] 本申请实施例中，计算机设备可以获取源文本数据中目标源词语的词语位置，基于词语位置在源文本数据中添加第一约束标签。其中，目标源词语的样式为目标源样式。目标源词语的词语位置可以是指目标源词语中包含的单词在源文本数据中的位置，例如可以为单词顺序号。由于确定了目标源词语的词语位置，因此可以基于词语位置在源文本数据中添加第一约束标签，例如可以在目标源词语的前后位置添加第一约束标签，即在目标源词语与前一个单词之间的位置添加第一约束标签，以及在目标源词语与后一个单词之间的位置添加第一约束标签，第一约束标签用于指示目标源词语。

[0045] 如图2所示，例如源文本数据为22，源文本数据中的目标源词语为“游乐园”，计算机设备在源文本数据中添加第一约束标签后得到的文本数据为24，其中“<b1>”、“<e1>”表示第一约束标签。

[0046] 可选地，若源文本数据中存在至少两个目标源词语时，计算机设备可以分别获取每个目标源词语的词语位置，基于每个目标源词语的词语位置在源文本数据中添加第一约束标签，从而实现对源文本数据中的每个具有样式的目标源词语进行展示。

[0047] 可选地，计算机设备可以获取源文本数据所在的语义场景，基于该语义场景确定源文本数据中每个目标源词语的优先级，获取优先级大于等级阈值的目标源词语在源文本数据中的词语位置，基于该优先级大于等级阈值的目标源词语的词语位置在源文本数据中添加第一约束标签，从而实现优先展示优先级大于等级阈值的目标源词语。也就是说，基于源文本数据所在的语义场景，计算机设备可以确定在该语义场景下多个具有样式的源词语的优先级，从而根据优先级等级展示目标源词语，例如，计算机设备可以根据优先级和目标时间间隔依次对每个具有样式的源词语进行展示，数据展示方式更灵活，可以实现突出文本数据中的重点，提升用户体验。

[0048] S103，基于添加第一约束标签后的源文本数据和词语间的对应关系，得到添加了第二约束标签的目标文本数据。

[0049] 本申请实施例中，目标文本数据是对源文本数据进行翻译得到的，可以理解的是，这里对源文本数据进行翻译是指对源文本数据中的文字进行翻译。目标词语为与目标源词

语对应的词语。

[0050] 可选地，计算机设备可以基于目标处理模型对添加第一约束标签后的源文本数据进行处理，得到添加第二约束标签的目标文本数据，也就是说，计算机设备将添加第一约束标签后的源文本数据输入到目标处理模型，通过目标处理模型对数据进行处理，在对源文本数据进行翻译生成目标文本数据的同时可以在目标文本数据中的目标词语的词语位置添加第二约束标签，从而可以通过目标处理模型输出添加第二约束标签的目标文本数据，添加第二约束标签的目标文本数据可以如图2中的25所示，其中“<b1>”、“<e1>”表示第二约束标签。

[0051] 具体地，目标处理模型可以对添加第一约束标签后的源文本数据进行翻译得到目标文本数据的同时，在目标文本数据中添加第二约束标签，从而得到添加了第二约束标签目标文本数据。换句话说，第二约束标签也可以认为是通过第一约束标签“翻译”得到的，即对源文本数据进行翻译时，不仅对源文本数据进行了翻译还可以对第一约束标签进行翻译，从而得到添加了第二约束标签的目标文本数据。

[0052] 可选地，目标处理模型在目标文本数据中添加第二约束标签的方法可以包括：获取源文本数据中的词语与目标文本数据中的词语间的对应关系，基于词语间的对应关系确定目标文本数据中与目标源词语对应的目标词语，从而基于目标词语在目标文本数据中的词语位置在目标文本数据中添加第二约束标签。也就是说，目标处理模型具有对源文本数据进行翻译，以及确定源文本数据中的词语和该源文本数据对应的翻译文本数据中的词语间的对应关系的能力。具体实现中，计算机设备可以预先获取大量样本数据对目标处理模型进行训练，使得目标处理模型具有确定源文本数据中的词语和该源文本数据对应的翻译文本数据中的词语之间的对应关系的能力。其中，训练目标处理模型的方法可以参考图4对应的实施例，此处不做过多描述。通过使用目标处理模型对文本数据进行处理，可以提高数据处理的效率。由于预先对目标处理模型进行训练，因此使用训练后的模型对文本数据进行处理，可以提高数据处理的准确性。

[0053] 在一种可能的实现方式中，计算机设备也可以对源文本数据进行翻译得到目标文本数据，对源文本数据和目标文本数据进行对齐处理，从而确定源文本数据中的词语与目标文本数据中的词语之间的对应关系。通过对源文本数据和目标文本数据进行对齐处理，后续可以基于该对应关系确定源文本数据中的词语对应目标文本数据中的哪个词语，进而确定在目标文本数据中的哪些位置添加第二约束标签。可选地，计算机设备可以对源文本数据和目标文本数据进行分词处理，得到源文本数据中的每个单词与目标文本数据中的每个单词之间的词对应关系，再基于单词之间的词对应关系确定源文本数据中的每个词语与目标文本数据中的词语之间的词语对应关系，从而确定源文本数据和目标文本数据是否对齐，若源文本数据和目标文本数据对齐，则将对齐关系确定为源文本数据中的词语和目标文本数据中的词语之间的对应关系。可以理解的是，若文本数据为中文，则文本数据中的单词可以包括文本数据中的单字词、词语、成语或者标点符号，等等；若文本数据为英文，则单词可以包括文本数据中的每个英文单词或者标点符号，等等。

[0054] 可选地，计算机设备可以采用对齐一致性原则对源文本数据中的词语和目标文本数据中的词语进行判断，确定源文本数据中的词语和目标文本数据中的词语是否对齐，若源文本数据中的词语和目标文本数据中的词语对齐，则计算机设备还可以获取源文本数据

中的关键词语和目标文本数据中的关键词语，基于源文本数据中的关键词语和目标文本数据中的关键词语之间的匹配情况确定源文本数据中的关键词语和目标文本数据中的关键词语是否对齐。进一步可选地，若源文本数据中的关键词语和目标文本数据中的关键词语对齐，则计算机设备还可以检测源文本数据或者目标文本数据中是否存在对空词，若均不存在对空词，则确定源文本数据和目标文本数据对齐。若存在对空词，则分别将对空词加入与对空词相邻的前邻居词语和后邻居词语中，确定对空词属于前邻居词语或者后邻居词语，基于对空词所属的邻居词语对源文本数据中的词语和目标文本数据中的词语之间的对应关系进行调整，使得源文本数据中的词语和目标文本数据中的词语对齐，进而将该对齐关系确定为源文本数据中的词语和目标文本数据中的词语之间的对应关系。

[0055] 可选地，本申请实施例中还可以采用其他方式确定源文本数据和目标文本数据是否对齐，计算机设备在确定源文本数据和目标文本数据对齐之后，可以获取源文本数据中的关键词语和目标文本数据中的关键词语进行匹配，基于关键词语匹配关系确定源文本数据中的词语和目标文本数据中的词语之间的对应关系。可选地，计算机设备还可以在确定源文本数据和目标文本数据对齐之后，检测源文本数据或者目标文本数据中是否存在对空词，并在源文本数据或者目标文本数据中存在对空词的情况下，对源文本数据或者目标文本数据中的对空词进行划分，从而将对空词划分至对应的组合词语中，基于得到的组合词语对源文本数据中的词语和目标文本数据中的词语之间的对应关系进行调整。

[0056] 可以理解的是，若源文本数据和目标文本数据未对齐，且计算机设备可以不进行后续处理，即计算机设备未确定源文本数据中的词语与目标文本数据中的词语之间的对应关系，则无需进行后续处理；或者，计算机设备可以重新获取源文本数据对应的目标文本数据，基于重新获取的目标文本数据与源文本数据进行对齐处理，以确定源文本数据中的词语与目标文本数据中的词语之间的对应关系。

[0057] S104，基于第一约束标签和第二约束标签将目标源样式映射为目标词语的目标样式，以得到包含目标样式的目标文本数据。

[0058] 本申请实施例中，计算机设备可以基于第一约束标签和第二约束标签将目标源样式映射为目标词语的目标样式，以得到包含目标样式的目标文本数据。具体地，由于计算机设备可以确定源文本数据中的词语和目标文本数据中的词语间的对应关系，因此可以确定源文本数据中的目标源词语与目标文本数据中对应的词语，即目标词语，则可以确定在目标文本数据中目标词语的词语位置添加第二约束标签，换句话说，第二约束标签指示的词语即为目标词语，且该目标词语与目标源词语具有对应关系。也就是说，由于基于目标源词语和词语间的对应关系，计算机设备可以确定目标文本数据中与目标源词语对应的目标词语，因此计算机设备可以获取目标词语在目标文本数据中的位置信息，基于目标词语在目标文本数据中的位置信息在目标文本数据中添加第二约束标签，并基于第一约束标签和第二约束将目标源词语的目标源样式映射为目标词语的目标样式，以得到包含目标样式的目标文本数据。

[0059] 如图2所示，计算机设备基于添加第一约束标签后的源文本数据24和词语间的对应关系，例如词语间的对应关系包括源文本数据22中的目标源词语“游乐园”与目标词语为“amusement park”对应，则计算机设备在目标文本数据中目标词语的词语位置添加第二约束标签后得到的文本数据为25，计算机设备将第一约束标签对应的目标源样式(加粗和下

划线)映射为第二约束标签对应的目标样式(加粗和下划线)后,得到包含目标样式的目标文本数据26。

[0060] 可选地,计算机设备还可以获取用户生理特征数据或者用户历史文本展示数据,基于用户生理特征数据或者用户历史文本展示数据确定样式映射关系,基于该样式映射关系将目标源样式映射为目标词语的目标样式。其中,用户生理特征数据和用户历史文本展示数据均为经过用户授权后的数据。用户生理特征数据可以包括但不限于用户的喜好或者用户是否具有视觉障碍,如红绿色盲,等等。用户历史文本展示数据可以是指用户在历史时间段内的用户常用样式。样式映射关系可以包括两种相同的样式之间的映射关系或者两种不同的样式之间的映射关系,若计算机设备基于用户生理特征数据或者用户历史文本展示数据确定的样式与源文本数据中的目标源词语的样式相同,则样式映射关系表示两种相同的样式之间的映射关系,即映射前后的样式相同,例如目标源样式包括红色和下划线,映射后得到的目标样式也包括红色和下划线。若计算机设备基于用户生理特征数据或者用户历史文本展示数据确定的样式与源文本数据中的目标源词语的样式不同,则样式映射关系表示两种不同的样式之间的映射关系,即映射前后的样式不同,例如目标源样式包括红色和下划线,映射后得到的目标样式也包括紫色和加粗,等等。

[0061] 由于可以根据用户生理特征数据和用户历史文本展示数据对目标文本数据中的样式进行针对化展示,因此可以实现个性化展示翻译后的目标文本数据。例如,若用户具有视觉障碍,如红绿色盲,或者用户倾向于某一种样式,则该种数据展示方法较为灵活,可以实现个性化展示,进而提升用户体验。

[0062] 可选地,计算机设备还可以获取用户选择指令,基于用户选择指令确定样式映射关系,基于该样式映射关系将目标源样式映射为目标词语的目标样式。用户选择指令包括用户选择的样式。具体实现中,计算机设备可以通过用户终端输出至少一种样式,用户可以从至少一种样式中选择想要的参考样式,当用户选择参考样式时,触发用户选择指令,则计算机设备响应于用户选择指令,基于用户选择的参考样式确定样式映射关系,基于该样式映射关系将目标源样式映射为目标词语的目标样式。由于可以根据用户选择对目标文本数据中的样式进行针对性展示,因此数据展示方法更为灵活,可以实现个性化展示。

[0063] 可选地,一种可能的实现方式中,目标处理模型也可以获取富样式的源文本数据;获取源文本数据中目标源词语的词语位置,基于词语位置在源文本数据中添加第一约束标签;基于添加第一约束标签后的源文本数据和词语间的对应关系,在目标文本数据中目标词语的词语位置添加第二约束标签,基于第一约束标签和第二约束标签将目标源样式映射为目标词语的目标样式,以得到包含目标样式的目标文本数据。也就是说,该种实现方式中,计算机设备可以将富样式的源文本数据输入目标处理模型,基于目标处理模型对富样式的源文本数据添加第一约束标签,最终输出包含目标样式的目标文本数据。

[0064] 可选地,计算机设备还可以基于目标词语的目标样式对目标文本数据进行渲染,以得到渲染后的包含目标样式的目标文本数据;调用关联的用户终端输出渲染后的包含目标样式的目标文本数据。若计算机设备为用户终端,则计算机设备可以直接输出渲染后的包含目标样式的目标文本数据;若计算机设备为服务器,则计算机设备可以调用与计算机设备关联的用户终端输出渲染后的包含目标样式的目标文本数据。可选地,计算机设备还可以调用关联的用户终端输出源文本数据。可选地,计算机设备还可以将渲染后的包含目

标样式的目标文本数据发送至其他用户终端，以使其他用户终端输出包含目标样式的目标文本数据。通过对目标文本数据进行渲染后输出，可以直观地查看到目标文本数据中与源文本数据中的源样式所对应的样式，从而实现将包含样式的源文本数据和目标文本数据进行输出，便于进行对应查看，用户无需根据源文本数据中包含样式的词语在目标文本数据中进行语义查找，可以提高数据查看效率，进而提升用户体验。

[0065] 本申请实施例中，通过获取富样式的源文本数据，以及获取源文本数据中具有目标源样式的目标源词语的词语位置，可以基于目标源词语的词语位置在源文本数据中添加第一约束标签；可以基于添加第一约束标签后的源文本数据和词语间的对应关系，在对源文本数据进行翻译得到的目标文本数据中目标词语的词语位置添加第二约束标签，从而基于第一约束标签和第二约束标签将目标源样式映射为第二约束标签对应的目标词语的目标样式，以得到包含目标样式的目标文本数据。由于确定了源文本数据中具有源样式的目标源词语的位置信息，可以基于该目标源词语在源文本数据中的位置信息为源文本数据添加约束标签，从而基于添加约束标签后的源文本数据和词语间的对应关系在目标文本数据中对应的目标词语的词语位置添加约束标签，基于约束标签将源文本数据中的源样式映射为目标文本数据中的目标样式，可以使得翻译后的目标文本数据中包含目标样式，从而实现在文本翻译的同时保留文本中的样式，保证文本数据翻译的完整性，提高数据处理的准确性。

[0066] 可选的，请参见图4，图4是本申请实施例提供的另一种数据处理方法的流程示意图。该数据处理方法可以应用于计算机设备；如图4所示，该数据处理方法包括但不限于以下步骤：

[0067] S201，获取第一样本数据和第二样本数据。

[0068] 本申请实施例中，计算机设备获取第一样本数据和第二样本数据的方法可以参考步骤S101中获取源文本数据的方法，此处不做过多描述。或者，计算机设备可以从数据库中下载大量双语语料样本数据集，大量双语语料样本数据集中包括第一样本数据和第二样本数据。第一样本数据为富样式文本数据，第一样本数据可以包括但不限于教学领域的文本数据、医学领域的文本数据或者其他领域的文本数据。第二样本数据是第一样本数据的翻译文本数据，第一样本数据与第二样本数据为两种不同语言的文本数据。可以理解的是，若第一样本数据和第二样本数据涉及到用户信息等相关的数据时，第一样本数据和第二样本数据均为用户授权后的数据，即第一样本数据和第二样本数据的收集、使用和处理需要遵守相关国家和地区的相关法律法规和标准。

[0069] S202，对第一样本数据和第二样本数据进行对齐处理，以确定第一样本数据中的样本词语与第二样本数据中的样本词语之间的样本对应关系。

[0070] 本申请实施例中，计算机设备可以对第一样本数据和第二样本数据进行对齐处理，从而确定第一样本数据中的样本词语与第二样本数据中的样本词语之间的样本对应关系。通过对第一样本数据和第二样本数据进行对齐处理，后续可以基于该样本对应关系确定第一样本数据中的样本词语对应第二样本数据中的哪个样本词语，便于后续将第一样本数据中样本词语的第一样式映射为第二样本数据中样本词语的第二样式。

[0071] 可选地，计算机设备可以对第一样本数据和第二样本数据进行分词处理，确定第一样本数据和第二样本数据是否对齐。具体地，计算机设备可以对第一样本数据和第二样

本数据进行分词处理，得到N个第一单词和M个第二单词之间的第一对应关系，其中，第一单词为第一样本数据中的任意一个单词，第二单词为第二样本数据中的任意一个单词，N、M均为正整数，第一对应关系可以是指单词之间的对应关系；基于N个第一单词和M个第二单词之间的第一对应关系，确定i个第一词语与j个第二词语之间的第二对应关系，第一词语为N个第一单词中的至少一个单词组成的词语，第二词语为M个第二单词中的至少一个单词组成的词语，i为小于或等于N的正整数、j为小于或等于M的正整数，第二对应关系可以是指词语之间的对应关系；基于第二对应关系确定第一样本数据和第二样本数据是否对齐。进一步地，若第一样本数据和第二样本数据对齐，则将对齐关系作为源文本数据中的样本词语与目标文本数据中的样本词语之间的样本对应关系。

[0072] 可以理解的是，若第一样本数据为中文，则第一单词可以包括第一样本数据中的单字词、词语、成语或者标点符号，等等；若第二样本数据为中文，则第二单词可以包括第二样本数据中的单字词、词语、成语或者标点符号，等等；若第一样本数据为英文，则第一单词可以包括第一样本数据中的每个英文单词或者标点符号，等等；若第二样本数据为英文，则第二单词可以包括第二样本数据中的每个英文单词或者标点符号，等等。

[0073] 可选地，计算机设备可以采用词对齐工具包对第一样本数据和第二样本数据进行对齐处理，词对齐工具包可以包括但不限于fast_align(快速词对齐工具包)和giza++(一种词对齐工具)，通过采用词对齐工具包对第一样本数据和第二样本数据进行对齐处理，可以得到单词对齐矩阵，单词对齐矩阵用于通过矩阵的方式表示两组数据中的单词之间的第一对应关系。基于单词对齐矩阵可以确定第一样本数据中的单词和第二样本数据中的单词之间的第一对应关系。举例来说，第一样本数据为“Go to amusement park to play tomorrow”，第二样本数据为“明天去游乐园玩”，通过词对齐工具包进行处理后输出单词对齐矩阵[0-1 1-1 2-2 3-2 4-3 5-3 6-0]，则基于单词对齐矩阵可以确定“Go”对应“去”、“amusement”和“park”对应“游乐园”、“play”对应“玩”、“tomorrow”对应“明天”，从而根据第一单词和第二单词之间的第一对应关系确定第一词语和第二词语之间的第二对应关系。举例来说，由于“amusement”和“park”均对应“游乐园”，则第一词语“amusement park”对应第二词语“游乐园”。如图5所示，图5是本申请实施例提供的一种词语对应关系的示意图，其中，灰色圆圈51表示第一样本数据中的第一词语，灰色圆圈52表示第二样本数据中的第二词语，图5中可以表示第一词语与第二词语之间的第二对应关系。

[0074] 可选地，计算机设备在获取到第二对应关系之后，可以采用对齐一致性原则确定第一词语和第二词语是否对齐。具体地，计算机设备可以采用对齐一致性原则确定第一目标词语和第二目标词语是否对齐，其中，第一目标词语为i个第一词语中的任意一个，第二目标词语为j个第二词语中的任意一个，对齐一致性原则用于指示第一样本数据中的多个连续词语组成的词语与第二样本数据中的多个连续词语组成的词语是否对应；若第一目标词语和第二目标词语对齐，则确定第一样本数据和第二样本数据对齐；若第一目标词语和第二目标词语未对齐，则确定第一样本数据和第二样本数据未对齐。

[0075] 具体实现中，由于计算机设备获取到第一样本数据中的i个第一词语与第二样本数据中的j个第二词语之间的第二对应关系，计算机设备可以获取第一样本数据中的连续多个单词组成的第一词语，基于该词语对应关系确定第二样本数据中是否存在连续多个单词组成的第二词语与第一词语对应，若第二样本数据中存在连续多个单词组成的第二词语

与第一词语对应，则确定第一样本数据和第二样本数据对齐，如图5所示。若第一样本数据中存在不连续的多个单词组成的第一词语与第二样本数据中的连续多个单词组成的第二词语对应，则确定第一样本数据和第二样本数据未对齐，如图6所示，图6是本申请实施例提供的另一种词语对应关系的示意图，其中，灰色圆圈61表示第一样本数据中的第一词语，灰色圆圈62表示第二样本数据中的第二词语。由于第二样本数据中存在一个单词(白色虚线圆圈)对应到第一样本数据中的第一词语以外的单词(白色虚线圆圈)，则确定第一样本数据和第二样本数据整体上不是对齐一致性双语词语，因此确定第一样本数据和第二样本数据未对齐。

[0076] 可以理解的是，本申请实施例中是采用对齐一致性原则对i个第一词语中的任意一个词语以及j个第二词语中的任意一个词语进行判断，确定i个第一词语中的任意一个词语与j个第二词语中的任意一个词语是否对齐，当i个第一词语中的每个词语与j个第二词语中的词语均对齐，则确定第一样本数据和第二样本数据对齐。若i个第一词语中存在一个或多个词语与j个第二词语中的词语未对齐，则确定第一样本数据和第二样本数据未对齐。由于在训练模型的时候使用了大量样本数据进行训练，即第一样本数据和第二样本数据的数量均为多个，当存在某一个第一样本数据和该第一样本数据对应的翻译文本数据(即第二样本数据)未对齐，则计算机设备可以删除该第一样本数据和第二样本数据，即删除样本数据中的误差数据，使得模型训练结果更准确，提高模型训练效率。

[0077] 可选地，若第一目标词语和第二目标词语对齐，则计算机设备还可以从第一样本数据中获取关键词语以及从第二样本数据中获取关键词语进行匹配，从而确定第一样本数据和第二样本数据是否对齐。具体地，计算机设备可以获取第一样本数据中的第一关键词语，以及第二样本数据中的第二关键词语；对第一关键词语和第二关键词语进行匹配；若第一关键词语和第二关键词语匹配，则计算机设备可以确定第一样本数据和第二样本数据对齐。

[0078] 其中，第一关键词语可以用于指示第一样本数据的含义，第二关键词语可以用于指示第二样本数据的含义，例如第一样本数据为“明天去游乐园玩”，则第一关键词语可以包括“游乐园”、“明天”，等等。对第一关键词语和第二关键词语进行匹配可以是指对获取到的第一关键词语和第二关键词语的词语含义进行匹配，若第一关键词语与第二关键词语的词语含义相同，则表示第一关键词语和第二关键词语匹配；若第一关键词语与第二关键词语的词语含义不同，则表示第一关键词语和第二关键词语不匹配。

[0079] 举例来说，若第一样本数据为“Spend a lot of time and energy and resources”，第二样本数据为“花费很多钱，精力和资源”，第一关键词语可以包括“energy and resources”和“time”，第二关键词语可以包括“精力和资源”和“钱”，由于“energy and resources”和“精力和资源”匹配，而“time”和“钱”不匹配，则表示第一关键词语和第二关键词语不匹配，确定第一样本数据和第二样本数据未对齐，则计算机设备可以删除该第一样本数据和第二样本数据，即删除样本数据中的误差数据，使得模型训练结果更准确，提高模型训练效率。若第一关键词语和第二关键词语匹配，确定第一样本数据和第二样本数据对齐，则计算机设备可以将对齐关系作为第一样本数据中的样本词语与第二样本数据中的样本词语之间的样本对应关系。可选地，计算机设备可以采用关键词提取工具获取第一样本数据和第二样本数据中的关键词，关键词提取工具包括但不限于rake-nltk(关键词自动

抽取模块)和ckpe(一种快速从自然语言文本中提取和识别关键短语的工具)。

[0080] 可以理解的是,第一样本数据中可以包括多个第一关键词语,第二样本数据中可以包括多个第二关键词语,若第一样本数据中的每个第一关键词语均与第二样本数据中的第二关键词语匹配,则确定第一样本数据和第二样本数据对齐。若第一样本数据中存在一个或多个第一关键词语与第二样本数据中的第二关键词语均不匹配,则确定第一样本数据和第二样本数据未对齐。由于在采用对齐一致性原则确定样本数据是否对齐的基础上,进一步获取第一关键词语和第二关键词语来确定第一样本数据和第二样本数据是否对齐,可以进一步提高数据对齐的准确性,进而提高模型训练的准确性。

[0081] 可选地,若第一目标词语和第二目标词语对齐,则计算机设备还可以检测第一样本数据或者第二样本数据中是否存在对空词,基于对空词对第一样本数据和第二样本数据之间的样本对应关系进行调整。具体地,若样本数据中存在对空词,则分别将对空词加入第一邻居词语和第二邻居词语中,得到第一组合词语和第二组合词语,其中,第一邻居词语和第二邻居词语为样本数据中与对空词相邻的两个词语,样本数据为第一样本数据和第二样本数据中的任意一个;从第一组合词语或第二组合词语中确定目标组合词语;基于目标组合词语对第一样本数据中的样本词语和第二样本数据中的样本词语之间的样本对应关系进行调整。其中,对空词可以指示第一样本数据中的某一个单词与第二样本数据中的每个单词均不对应,或者,对空词可以指示第二样本数据中的某一个单词与第一样本数据中的每个单词均不对应。可选地,计算机设备执行检测第一样本数据或者第二样本数据中是否存在对空词的步骤可以在执行确定第一样本数据中的关键词语和第二样本数据中的关键词语是否对齐的步骤之后;或者,计算机设备执行确定第一样本数据中的关键词语和第二样本数据中的关键词语是否对齐的步骤可以在执行检测第一样本数据或者第二样本数据中是否存在对空词的步骤之后;或者,计算机设备也可以同时执行两个步骤,本申请实施例对此不作限定。

[0082] 可选地,本申请实施例中还可以采用其他方式确定第一样本数据和第二样本数据是否对齐,计算机设备在确定第一样本数据和第二样本数据对齐之后,可以获取第一样本数据中的第一关键词语和第二样本数据中的第二关键词语进行匹配,基于关键词语匹配关系确定第一样本数据中的样本词语和第二样本数据中的样本词语之间的样本对应关系。可选地,计算机设备还可以在确定第一样本数据和第二样本数据对齐之后,检测第一样本数据或者第二样本数据中是否存在对空词,并在第一样本数据或者第二样本数据中存在对空词的情况下,对源第一样本数据或者第二样本数据中的对空词进行划分,从而将对空词划分至对应的组合词语中,基于得到的组合词语对第一样本数据中的样本词语和第二样本数据中的样本词语之间的样本对应关系进行调整。

[0083] 可选地,在确定第一样本数据和第二样本数据未对齐之后,计算机设备还可以接收修正终端发送的针对第二样本数据的修正请求,基于修正请求对第二样本数据进行修正,基于修正后的第二样本数据与第一样本数据进行对齐处理,以确定第一样本数据中的样本词语与第二样本数据中的样本词语之间的样本对应关系,修正终端可以对第二样本数据中的单词或者词语进行调整,修正请求中可以包括需要修正的单词或者词语。

[0084] 由于前述步骤通过对齐一致性原则确定第一目标词语和第二目标词语对齐,若检测到第一样本数据或者第二样本数据中存在对空词,则计算机设备可以将对空词划分至其

邻居词语中，从而实现将第一样本数据中的每个单词与第二样本数据中的每个单词对应。举例来说，第一样本数据为“On July 16local time,Zhang San Yong in city A”，第二样本数据为“当地时间7月16日，张三勇在A市”，通过对第一样本数据和第二样本数据进行对齐处理后得到单词对齐矩阵[0-5 1-2 2-4 3-0 4-1 5-6 6-7 7-7 8-7 9-9 10-10 11-10]，其中，单词对齐矩阵中的部分对齐关系为“6-7 7-7 8-7”，表示“Zhang San Yong”均对应“张三”，而第一样本数据中不存在与“勇”对应的单词，则表示“勇”为对空词，则计算机设备可以将“勇”分别加入第一邻居词语“张三”和第二邻居词语“在”中，得到第一组合词语“张三勇”和第二组合词语“勇在”。进一步地，计算机设备可以基于目标词典从第一组合词语或第二组合词语中确定目标组合词语，例如目标组合词语为“张三勇”，则计算机设备基于目标组合词语对第一样本数据中的样本词语和第二样本数据中的样本词语之间的样本对应关系进行调整，例如调整前的“Zhang San Yong”均对应“张三”，而第一样本数据中不存在与“勇”对应的单词，调整后的“Zhang San Yong”对应“张三勇”。可以理解的是，调整样本对应关系之后的样本文本数据中不存在对空词。目标词典可以是指预先设置的词典，目标词典中包括至少一个单字词、词语、成语、歇后语、姓名、地名，等等。

[0085] 如图7所示，图7是本申请实施例提供的一种对空词划分的示意图，可以看出，第一样本数据和第二样本数据对齐后，第二样本数据中存在对空词70，第一样本数据中不存在与对空词70对应的单词(图中的连接线可以表示单词之间的对应关系)，则计算机设备可以将对空词70分别加入到第一邻居词语73和第二邻居词语74中，从第一组合词语或第二组合词语中确定目标组合词语，例如目标组合词语包括第一组合词语73和对空词70，则划分后的对应关系如75所示。可以理解的是，图7是对第一样本数据或第二样本数据中存在的某一个对空词进行的处理，若存在多个对空词，则可以参考该方法对多个对空词进行处理，将多个对空词均划分至对应的邻居词语中，从而实现调整第一样本数据和第二样本数据之间的样本对应关系。

[0086] S203，获取对齐后的第一样本数据中第一样本词语的样本词语位置，基于样本词语位置在第一样本数据中添加第一样本约束标签。

[0087] 本申请实施例中，计算机设备可以获取对齐后的第一样本数据中第一样本词语的样本词语位置，基于样本词语位置在第一样本数据中添加第一样本约束标签，第一样本词语的样式为第一样式。样本词语位置可以是指第一样本词语中包含的样本单词在第一样本数据中的位置，例如可以为样本单词顺序号。由于确定了第一样本词语的样本词语位置，因此可以基于样本词语位置在第一样本数据中添加第一样本约束标签，例如可以在第一样本词语的前后位置添加第一样本约束标签，即在第一样本词语与前一个样本单词之间的位置添加第一样本约束标签，以及在第一样本词语与后一个样本单词之间的位置添加第一样本约束标签，第一样本约束标签用于指示第一样本词语。如图2所示，例如第一样本数据为22，第一样本数据中的第一样本词语为“游乐园”，计算机设备在第一样本数据中添加第一样本约束标签后得到的样本数据为24，其中“<b1>”、“<e1>”表示第一样本约束标签。

[0088] S204，获取参考样本，基于添加第一样本约束标签后的第一样本数据、样本对应关系以及参考样本，训练得到目标处理模型。

[0089] 本申请实施例中，计算机设备可以获取参考样本，基于添加第一样本约束标签后的第一样本数据、样本对应关系以及参考样本，训练得到目标处理模型。其中，参考样本可

以是指在第二样本数据中添加标记标签的样本，即目标处理模型的期望输出结果。具体地，计算机设备可以基于添加第一样本约束标签后的第一样本数据、样本对应关系以及第二样本数据，在第二样本数据中第二样本词语的词语位置添加第二样本约束标签，基于添加第二样本约束标签后的第二样本数据和参考样本确定初始处理模型的损失函数，基于损失函数训练初始处理模型，得到目标处理模型。

[0090] 其中，由于参考样本可以是指模型预期输出结果，添加第二样本约束标签后的第二样本数据可以是指模型实际输出结果，计算机设备可以基于添加第二约束标签后的第二样本数据(即模型实际输出结果)和参考样本(即模型预期输出结果)之间的重合度确定初始处理模型的损失函数。若添加第二约束标签后的第二样本数据和参考样本之间的重合度大于重合度阈值，则保存此时的初始处理模型，并将此时的初始处理模型确定为目标处理模型。若添加第二约束标签后的第二样本数据和参考样本之间的重合度小于或等于重合度阈值，则继续调整初始处理模型中的参数，并在重合度大于重合度阈值时，将此时的初始处理模型确定为目标处理模型。其中，模型的损失函数大于损失阈值，表示模型的准确度低于准确度阈值；模型的损失函数小于或等于损失阈值，表示模型的准确度高于准确度阈值。

[0091] 具体地，由于计算机设备对第一样本数据和第二样本数据进行了对齐处理，确定了第一样本数据中的样本词语和第二样本数据中的样本词语之间的样本对应关系，因此可以基于添加第一样本约束标签后的第一样本数据和样本对应关系、以及第二样本数据，确定在第二样本数据中的哪个位置添加第二样本约束标签，即在第二样本数据中第二样本词语的词语位置添加第二样本约束标签，换句话说，第二样本约束标签指示的词语即为第二样本词语，且该第二样本词语与第一样本词语具有对应关系。也就是说，由于基于第一样本词语和样本对应关系，计算机设备可以确定第二样本数据中与第一样本词语对应的第二样本词语，因此计算机设备可以获取第二词语在第二样本数据中的位置信息，基于第二样本词语在第二样本数据中的位置信息在第二样本数据中添加第二样本约束标签。

[0092] 可选地，若第一样本数据或第二样本数据中存在对空词，计算机设备对该对空词进行划分后，基于划分后得到的目标词语组合对第一样本数据中的样本词语和第二样本数据中的样本词语之间的样本对应关系进行调整后，计算机设备可以基于添加第一样本约束标签后的第一样本数据、调整后的样本对应关系以及参考样本，训练得到目标处理模型。

[0093] 如图2所示，计算机设备对第一样本数据和第二样本数据进行对齐处理后，确定第一样本数据22中的第一样本词语“游乐园”与第二样本数据中对应的第二样本词语为“amusement park”，计算机设备在第二样本数据中添加第二样本约束标签后得到的文本数据为25，其中“<b1>”、“<e1>”表示第二样本约束标签，计算机设备将第一样本词语的第一样式(加粗和下划线)映射为第二样本词语的第二样式(加粗和下划线)后，得到包含第二样式的第二样本数据26。

[0094] 本申请实施例中，通过使用大量样本数据对初始处理模型进行训练，并在训练过程中删除具有误差的样本数据，以及对样本数据中的样本对应关系进行调整，可以提高数据处理的准确性，从而提高模型训练的准确性，进而在使用训练后的模型对文本数据进行处理时，可以提高数据处理的准确性。由于在模型训练的过程中对误差数据进行删除，因此可以减少计算量，提高数据处理的效率。

[0095] 可选地，本申请实施例中涉及的目标处理模型的模型结构可以包括但不限于

Transformer模型结构、循环神经网络(Recurrent Neural Networks,RNN)、以及Transformer-mixaan模型结构。其中,Transformer-mixaan模型为Transformer模型的变种,传统的Transformer模型的decoder部分均由self-attention组成,如图8中8a所示,图8是本申请实施例提供的一种模型结构示意图,图8中,8a表示Transformer模型的解码器部分结构示意图,Transformer模型可以包括两个解码器,第一解码器中分别包括自注意力层、第一归一化层、注意力编码解码层、第二归一化层、前馈神经网络层。第二解码器也分别包括自注意力层、第一归一化层、注意力编码解码层、第二归一化层、前馈神经网络层。Transformer模型中还可以包括编码器,具体地,计算机设备将训练数据(如包含第一样本约束标签的第一样本数据、第二样本数据,等等)输入到Transformer模型之后,可以由编码器对训练数据进行编码处理,得到编码特征;由第一解码器的自注意力层对训练数据进行处理,得到训练数据中句子的语义特征;基于第一归一化层对句子的语义特征进行归一化处理,得到归一化后的语义特征;基于注意力编码解码层对归一化后的语义特征和编码特征进行处理,得到组合特征;基于第二归一化层对组合特征进行归一化处理,得到归一化后的组合特征;基于前馈神经网络层对归一化后的组合特征进行映射,得到第一解码器输出结果;再基于第二解码器对第一解码器输出结果和编码特征进行解码处理,得到模型输出结果。

[0096] 其中,自注意力层可以用于捕获句子中单词之间的语义特征;第一归一化层和第二归一化层可以用于对数据进行归一化,从而加快训练速度,提高训练的稳定性。注意力编码解码层可以用于帮助当前节点获取到当前需要关注的重点内容,即捕获当前句子中的重点信息。前馈神经网络层可以用于通过简单的非线性处理单元,使得模型获得非线性处理的能力,进而实现将第二归一化层输出的多个特征向量进行静态非线性化映射。可以理解的是,图8中的两种模型中均可以包括多个解码器,图8中只是示出了每个模型中的其中两个解码器。

[0097] 图8中8b表示Transformer-mixaan模型的解码器部分结构示意图,第一解码器中分别包括平均注意力层、第一归一化层、注意力编码解码层、第二归一化层、前馈神经网络层。第二解码器包括自注意力层、第一归一化层、注意力编码解码层、第二归一化层、前馈神经网络层。在Transformer模型中,第一解码器和第二解码器中的自注意力层均可以为self-attention、第一归一化层可以为add&normalize、注意力编码解码层可以为encoder-decoder attention、第二归一化层可以为add&normalize、前馈神经网络层可以为feed forward。在Transformer-mixaan模型中,第一解码器中的平均注意力层可以为average-attention、第一归一化层可以为add&normalize、注意力编码解码层可以为encoder-decoder attention、第二归一化层可以为add&normalize、前馈神经网络层可以为feed forward。可以看出,Transformer模型的两个解码器的自注意力层均是由self-attention组成,而Transformer-mixaan模型的两个解码器分别是由self-attention和average-attention交替的方式组成,即第一解码器的平均注意力层是由average-attention组成,第二解码器的自注意力层是由self-attention组成。由于self-attention是通过动态计算当前注意力权重,得到句子间的语义特征,而average-attention是基于固定的平均值计算累加历史权重,通过总结之前的历史信息,因此Transformer-mixaan模型的多样性更好,进而使用该模型对文本数据进行处理时,模型处理效果更好。

[0098] 可选地，本申请实施例可以主要包括构建训练语料、模型训练以及模型使用三个部分。其中，在构建训练语料时，计算机设备可以获取常规的双语训练语料，通过对双语训练语料进行对齐处理，可以确定双语训练语料中的词语之间的对应关系，从而得到具有对应关系的双语训练语料。进一步地，在训练模型时，通过使用具有对应关系的双语训练语料训练目标处理模型，使目标处理模型在对源文本数据进行翻译的同时也可以生成出与源文本数据中的第一约束标签对应的第二约束标签，得到添加了第二约束标签的目标文本数据，也就是说，在训练模型时使得目标处理模型具有对文本数据进行翻译的同时且添加第二约束标签的能力。进一步地，在模型使用时，通过获取富样式的文本数据，对富样式的文本数据添加第一约束标签后输入到目标处理模型，可以基于目标处理模型输出添加第二约束标签后的翻译文本数据。最后，通过将第一约束标签对应的词语的样式映射为第二约束标签对应的词语的样式，从而得到富样式的翻译文本数据。

[0099] 通过上述方法训练目标处理模型后，本申请实施例中可以使用该目标处理模型。可选地，计算机设备可以先获取源文本数据，如图2中22所示，其中，源文本数据中“游乐园”的源样式为加粗和下划线。进一步地，计算机设备可以获取源文本数据在中目标源词语的词语位置，如目标源词语为“游乐园”，基于目标源词语的词语位置在源文本数据中添加第一约束标签，如“明天去<b1>游乐园<e1>玩”，第一约束标签为“<b1><e1>”。进一步地，计算机设备将添加第一约束标签的源文本数据输入目标处理模型，基于目标处理模型输出添加第二约束标签的目标文本数据，添加第二约束标签的目标文本数据如“Go to<b1> amusement park<e1>to play tomorrow”，第二约束标签为“<b1><e1>”。进一步地，计算机设备可以基于第一约束标签和第二约束标签将目标源词语的目标源样式映射为目标词语的目标样式，如图2中26所示”。通过对源文本数据中具有源样式的目标源词语添加约束标签，使得得到的目标文本数据中包括对应的约束标签，进而实现将源文本数据中的源样式映射到目标文本数据中，保留文本翻译过程中的样式信息，使得文本翻译更完整，提高数据处理准确性。

[0100] 本申请实施例中，由于确定了源文本数据中具有源样式的目标源词语的位置信息，可以基于该目标源词语在源文本数据中的位置信息为源文本数据添加约束标签，从而基于添加约束标签后的源文本数据和词语间的对应关系得到添加了约束标签的目标文本数据，基于约束标签将源文本数据中的源样式映射为目标文本数据中的目标样式，可以使得翻译后的目标文本数据中包含目标样式，从而实现在文本翻译的同时保留文本中的样式，保证文本数据翻译的完整性，提高数据处理的准确性。进一步地，通过获取大量样本数据对模型进行训练，并在训练过程中删除具有误差的样本数据，以及对样本数据中的样本对应关系进行调整，可以提高数据处理的准确性，从而提高模型训练的准确性，进而在使用训练后的模型对文本数据进行处理时，可以提高数据处理的准确性。由于在模型训练的过程中对误差数据进行删除，因此可以减少计算量，提高数据处理的效率。

[0101] 上面介绍了本申请实施例的方法，下面介绍本申请实施例的装置。

[0102] 参见图9，图9是本申请实施例提供的一种数据处理装置的组成结构示意图，上述数据处理装置可以是运行于计算机设备中的一个计算机程序(包括程序代码)，例如该数据处理装置为一个应用软件；该数据处理装置可以用于执行本申请实施例提供的数据处理方法中的相应步骤。该数据处理装置90包括：

[0103] 文本获取模块91，用于获取源文本数据，源文本数据为富样式文本数据；

[0104] 第一添加模块92，用于获取源文本数据中目标源词语的词语位置，基于该目标源词语的词语位置在源文本数据中添加第一约束标签，该目标源词语的样式为目标源样式；

[0105] 第二添加模块93，用于基于添加第一约束标签后的源文本数据和词语间的对应关系，得到添加了第二约束标签的目标文本数据，目标文本数据是对源文本数据进行翻译得到的，该目标词语为与该目标源词语对应的词语；

[0106] 数据映射模块94，用于基于该第一约束标签和该第二约束标签将该目标源样式映射为该目标词语的目标样式，以得到包含该目标样式的目标文本数据。

[0107] 可选地，该数据处理装置90还包括模型训练模块95，该模型训练模块95包括：

[0108] 样本获取模块951，用于获取第一样本数据和第二样本数据，该第二样本数据是对该第一样本数据进行翻译得到的，该第一样本数据为富样式文本数据；

[0109] 样本对齐模块952，用于对该第一样本数据和该第二样本数据进行对齐处理，以确定该第一样本数据中的样本词语与该第二样本数据中的样本词语之间的样本对应关系；

[0110] 样本添加模块953，用于获取该第一样本数据中第一样本词语的样本词语位置，基于该样本词语位置在该第一样本数据中添加第一样本约束标签，该第一样本词语的样式为第一样式；

[0111] 样本映射模块954，用于获取参考样本标签，基于添加第一样本约束标签后的第一样本数据、该样本对应关系以及该参考样本，训练得到目标处理模型；

[0112] 该第二添加模块93，具体用于采用该目标处理模型基于添加第一约束标签后的源文本数据和词语间的对应关系，得到添加了第二约束标签的目标文本数据。

[0113] 可选地，该样本对齐模块952，具体用于：

[0114] 对该第一样本数据和第二样本数据进行分词处理，得到N个第一单词和M个第二单词之间的第一对应关系，第一单词为该第一样本数据中的单词，第二单词为该第二样本数据中的单词，N、M均为正整数；

[0115] 基于该N个第一单词和该M个第二单词之间的第一对应关系，确定i个第一词语与j个第二词语之间的第二对应关系，第一词语为该N个第一单词中的至少一个单词组成的词语，第二词语为该M个第二单词中的至少一个单词组成的词语，i、j均为正整数；

[0116] 基于该第二对应关系确定该第一样本数据和该第二样本数据是否对齐，若该第一样本数据和该第二样本数据对齐，则将对齐关系作为该源文本数据中的样本词语与该目标文本数据中的样本词语之间的样本对应关系。

[0117] 可选地，该样本对齐模块952，具体用于：

[0118] 采用对齐一致性原则确定第一目标词语和第二目标词语是否对齐，该第一目标词语为该i个第一词语中的任意一个，该第二目标词语为该j个第二词语中的任意一个，该对齐一致性原则用于指示该第一样本数据中的多个连续词语组成的词语与该第二样本数据中的多个连续词语组成的词语是否对应；

[0119] 若该第一目标词语和该第二目标词语对齐，则确定该第一样本数据和该第二样本数据对齐；

[0120] 若该第一目标词语和该第二目标词语未对齐，则确定该第一样本数据和该第二样本数据未对齐。

[0121] 可选地，该数据处理装置90还包括词语匹配模块96，用于：

[0122] 若该第一目标词语和该第二目标词语对齐，则获取该第一样本数据中的第一关键词语，以及第二样本数据中的第二关键词语；

[0123] 对该第一关键词语和该第二关键词语进行匹配；

[0124] 若该第一关键词语和该第二关键词语匹配，则确定该第一样本数据和该第二样本数据对齐。

[0125] 可选地，该数据处理装置90还包括单词划分模块97，用于：

[0126] 若该第一目标词语和该第二目标词语对齐，则检测该样本数据中是否存在对空词；

[0127] 若该样本数据中存在对空词，则分别将该对空词加入第一邻居词语和第二邻居词语中，得到第一组合词语和第二组合词语，该第一邻居词语和该第二邻居词语为该样本数据中与该对空词相邻的两个词语，该样本数据为该第一样本数据和该第二样本数据中的任意一个；

[0128] 从该第一组合词语或该第二组合词语中确定目标组合词语；

[0129] 基于该目标组合词语对该第一样本数据中的样本词语和该第二样本数据中的样本词语之间的样本对应关系进行调整。

[0130] 可选地，该数据处理装置90还包括数据渲染模块98，用于：

[0131] 基于该目标词语的目标样式对该目标文本数据进行渲染，以得到渲染后的包含目标样式的目标文本数据；

[0132] 调用关联的用户终端输出该渲染后的包含目标样式的目标文本数据。

[0133] 需要说明的是，图9对应的实施例中未提及的内容可参见方法实施例的描述，这里不再赘述。

[0134] 本申请实施例中，通过获取富样式的源文本数据，以及获取源文本数据中具有目标源样式的目标源词语的词语位置，可以基于目标源词语的词语位置在源文本数据中添加第一约束标签；可以基于添加第一约束标签后的源文本数据和词语间的对应关系，得到添加了第二约束标签的目标文本数据，从而基于第一约束标签和第二约束标签将目标源样式映射为第二约束标签对应的目标词语的目标样式，以得到包含目标样式的目标文本数据。由于确定了源文本数据中具有源样式的目标源词语的位置信息，可以基于该目标源词语在源文本数据中的位置信息为源文本数据添加约束标签，从而基于添加约束标签后的源文本数据和词语间的对应关系得到添加了约束标签的目标文本数据，基于约束标签将源文本数据中的源样式映射为目标文本数据中的目标样式，可以使得翻译后的目标文本数据中包含目标样式，从而实现在文本翻译的同时保留文本中的样式，保证文本数据翻译的完整性，提高数据处理的准确性。

[0135] 参见图10，图10是本申请实施例提供的一种计算机设备的组成结构示意图。如图10所示，上述计算机设备100可以包括：处理器1001，网络接口1004和存储器1005，此外，上述计算机设备100还可以包括：用户接口1003，和至少一个通信总线1002。其中，通信总线1002用于实现这些组件之间的连接通信。其中，用户接口1003可以包括显示屏(Display)、键盘(Keyboard)，可选用户接口1003还可以包括标准的有线接口、无线接口。网络接口1004可选的可以包括标准的有线接口、无线接口(如WI-FI接口)。存储器1005可以是高速RAM存

储器，也可以是非易失性的存储器(non-volatile memory)，例如至少一个磁盘存储器。存储器1005可选的还可以是至少一个位于远离前述处理器1001的存储装置。如图10所示，作为一种计算机可读存储介质的存储器1005中可以包括操作系统、网络通信模块、用户接口模块以及设备控制应用程序。

[0136] 在图10所示的计算机设备100中，网络接口1004可提供网络通讯功能；而用户接口1003主要用于为用户提供输入的接口；而处理器1001可以用于调用存储器1005中存储的设备控制应用程序，以执行以下操作：

[0137] 获取源文本数据，源文本数据为富样式文本数据；

[0138] 获取源文本数据中目标源词语的词语位置，基于该目标源词语的词语位置在源文本数据中添加第一约束标签，该目标源词语的样式为目标源样式；

[0139] 基于添加第一约束标签后的源文本数据和词语间的对应关系，得到添加了第二约束标签的目标文本数据，目标文本数据是对源文本数据进行翻译得到的，该目标词语为与该目标源词语对应的词语；

[0140] 基于该第一约束标签和该第二约束标签将该目标源样式映射为该第二约束标签对应的目标词语的目标样式，以得到包含该目标样式的目标文本数据。

[0141] 应当理解，本申请实施例中所描述的计算机设备100可执行前文图3和图4所对应实施例中对上述数据处理方法的描述，也可执行前文图9所对应实施例中对上述数据处理装置的描述，在此不再赘述。另外，对采用相同方法的有益效果描述，也不再进行赘述。

[0142] 本申请实施例中，通过获取富样式的源文本数据，以及获取源文本数据中具有目标源样式的目标源词语的词语位置，可以基于目标源词语的词语位置在源文本数据中添加第一约束标签；可以基于添加第一约束标签后的源文本数据和词语间的对应关系，得到添加了第二约束标签的目标文本数据，从而基于第一约束标签和第二约束标签将目标源样式映射为第二约束标签对应的目标词语的目标样式，以得到包含目标样式的目标文本数据。由于确定了源文本数据中具有源样式的目标源词语的位置信息，可以基于该目标源词语在源文本数据中的位置信息为源文本数据添加约束标签，从而基于添加约束标签后的源文本数据和词语间的对应关系得到添加了约束标签的目标文本数据，基于约束标签将源文本数据中的源样式映射为目标文本数据中的目标样式，可以使得翻译后的目标文本数据中包含目标样式，从而实现在文本翻译的同时保留文本中的样式，保证文本数据翻译的完整性，提高数据处理的准确性。

[0143] 本申请实施例还提供一种计算机可读存储介质，该计算机可读存储介质存储有计算机程序，该计算机程序包括程序指令，该程序指令当被计算机执行时使该计算机执行如前述实施例的方法，该计算机可以为上述提到的计算机设备的一部分。例如为上述的处理器1001。作为示例，程序指令可被部署在一个计算机设备上执行，或者被部署位于一个地点的多个计算机设备上执行，又或者，在分布在多个地点且通过通信网络互连的多个计算机设备上执行，分布在多个地点且通过通信网络互连的多个计算机设备可以组成区块链网络。

[0144] 本领域普通技术人员可以理解实现上述实施例的方法中的全部或部分流程，是可以通过计算机程序来指令相关的硬件来完成，该的程序可存储于计算机可读取存储介质中，该程序在执行时，可包括如上述各方法的实施例的流程。其中，该的存储介质可为磁碟、

光盘、只读存储记忆体(Read-Only Memory,ROM)或随机存储记忆体(Random Access Memory,RAM)等。

[0145] 以上所揭露的仅为本申请较佳实施例而已,当然不能以此来限定本申请之权利范围,因此依本申请权利要求所作的等同变化,仍属本申请所涵盖的范围。

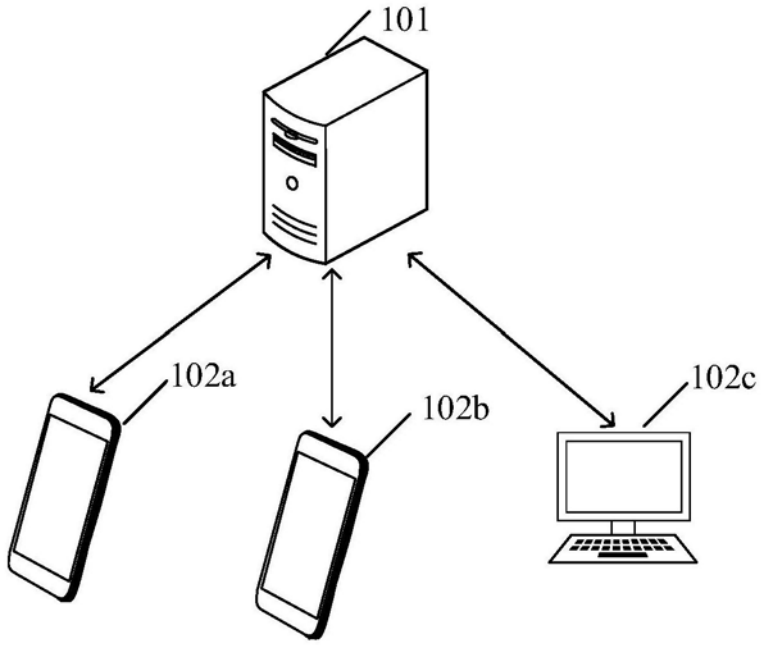

101
102a
102b
102c

图1

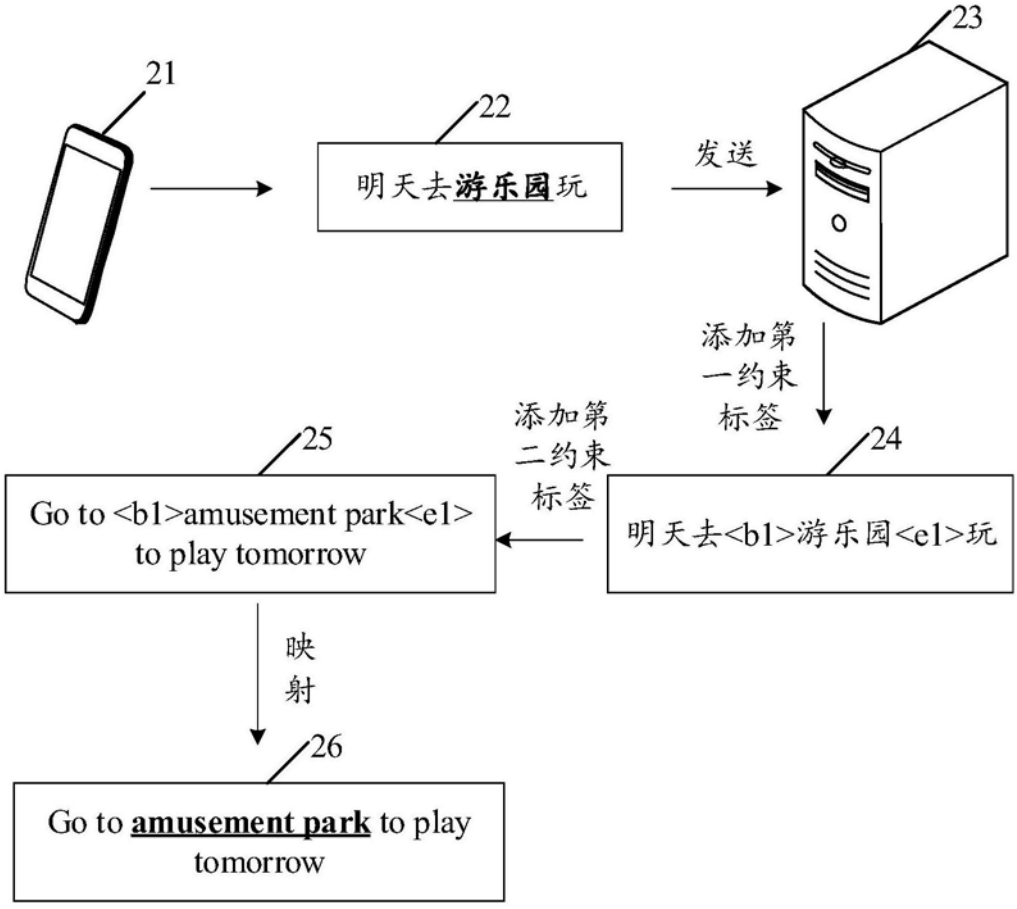

21
22
明天去游乐园玩
发送
23
添加第一约束标签
添加第二约束标签
24
明天去<b1>游乐园<e1>玩
25
Go to <b1>amusement park<e1> to play tomorrow
映射
26
Go to amusement park to play tomorrow

图2

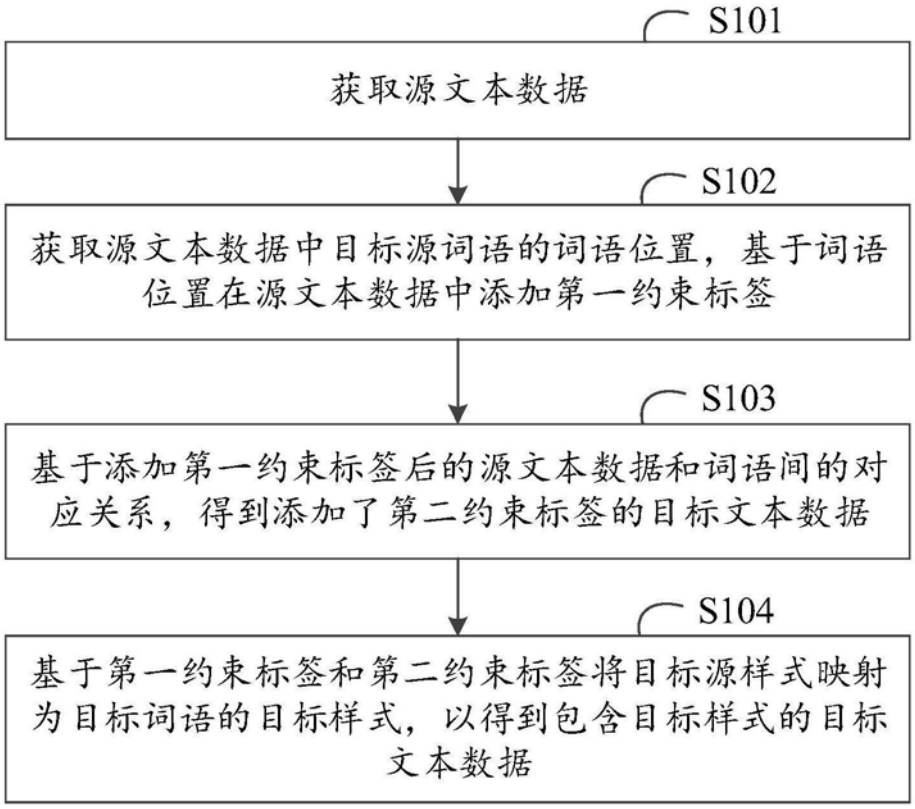

S101
获取源文本数据
S102
获取源文本数据中目标源词语的词语位置，基于词语位置在源文本数据中添加第一约束标签
S103
基于添加第一约束标签后的源文本数据和词语间的对应关系，得到添加了第二约束标签的目标文本数据
S104
基于第一约束标签和第二约束标签将目标源样式映射为目标词语的目标样式，以得到包含目标样式的目标文本数据

图3

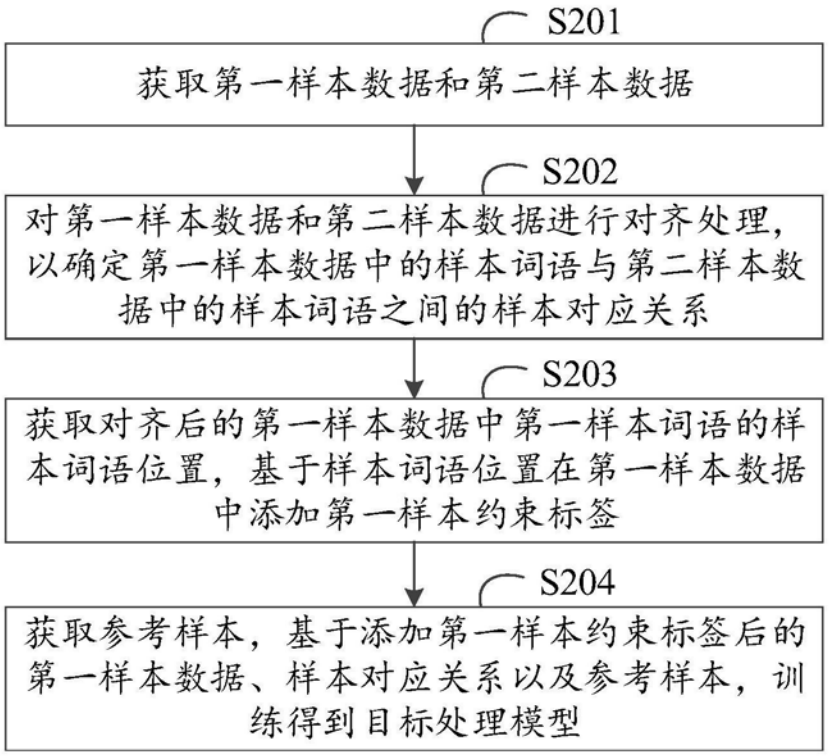

S201
获取第一样本数据和第二样本数据
S202
对第一样本数据和第二样本数据进行对齐处理，以确定第一样本数据中的样本词语与第二样本数据中的样本词语之间的样本对应关系
S203
获取对齐后的第一样本数据中第一样本词语的样本词语位置，基于样本词语位置在第一样本数据中添加第一样本约束标签
S204
获取参考样本，基于添加第一样本约束标签后的第一样本数据、样本对应关系以及参考样本，训练得到目标处理模型

图4

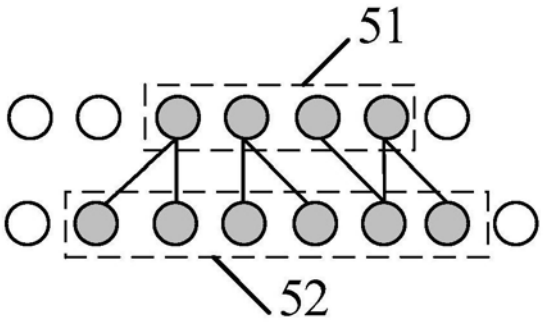



图5

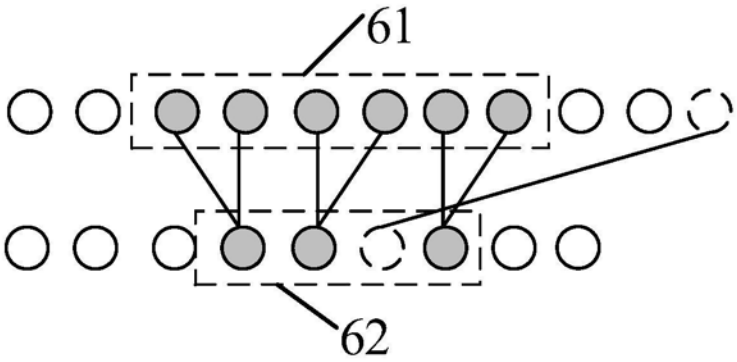



图6

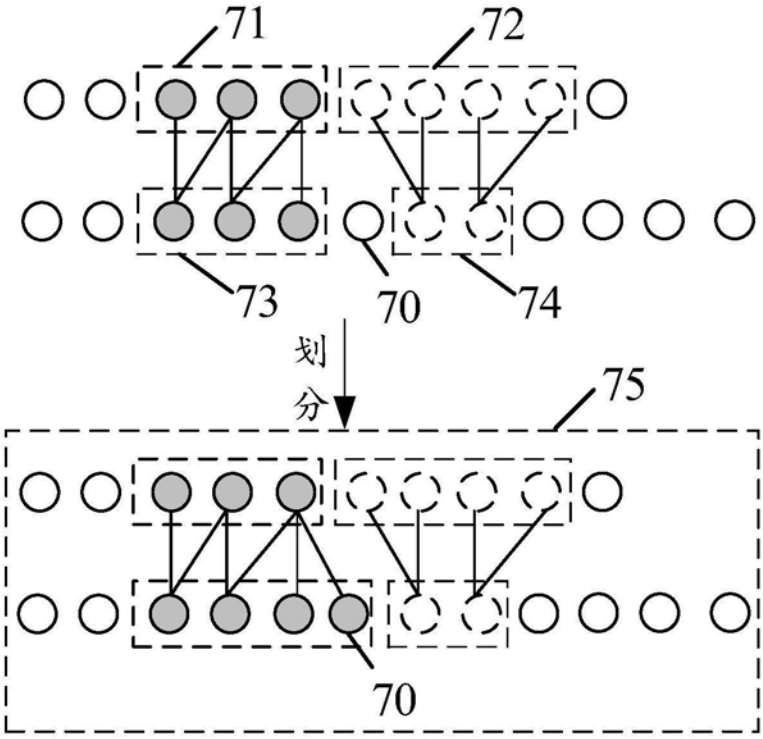



图7

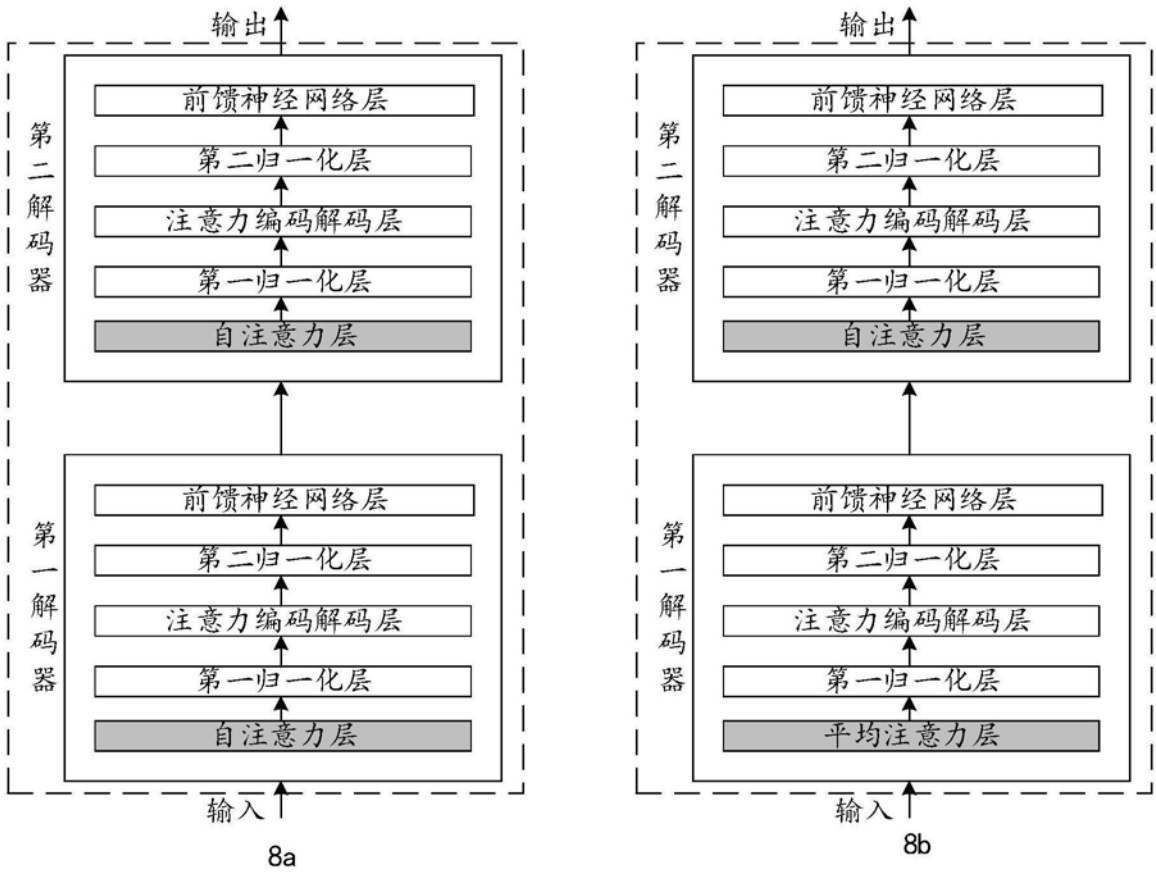

输出
第二解码器
前馈神经网络层
第二归一化层
注意力编码解码层
第一归一化层
自注意力层
第一解码器
前馈神经网络层
第二归一化层
注意力编码解码层
第一归一化层
自注意力层
输入
8a
输出
第二解码器
前馈神经网络层
第二归一化层
注意力编码解码层
第一归一化层
自注意力层
第一解码器
前馈神经网络层
第二归一化层
注意力编码解码层
第一归一化层
平均注意力层
输入
8b

图8

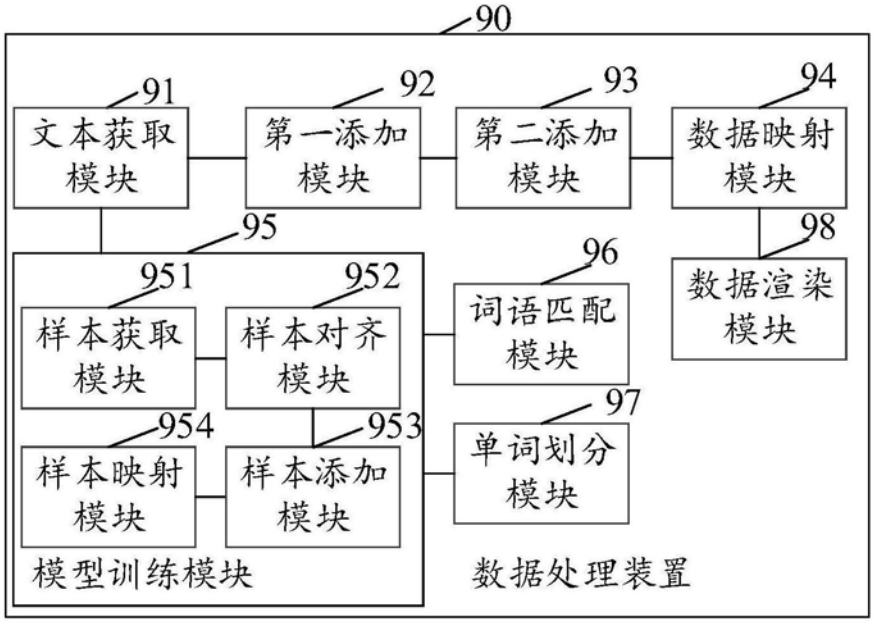

90
91
文本获取模块
92
第一添加模块
93
第二添加模块
94
数据映射模块
95
951
样本获取模块
952
样本对齐模块
954
样本映射模块
953
样本添加模块
模型训练模块
96
词语匹配模块
97
单词划分模块
98
数据渲染模块
数据处理装置

图9

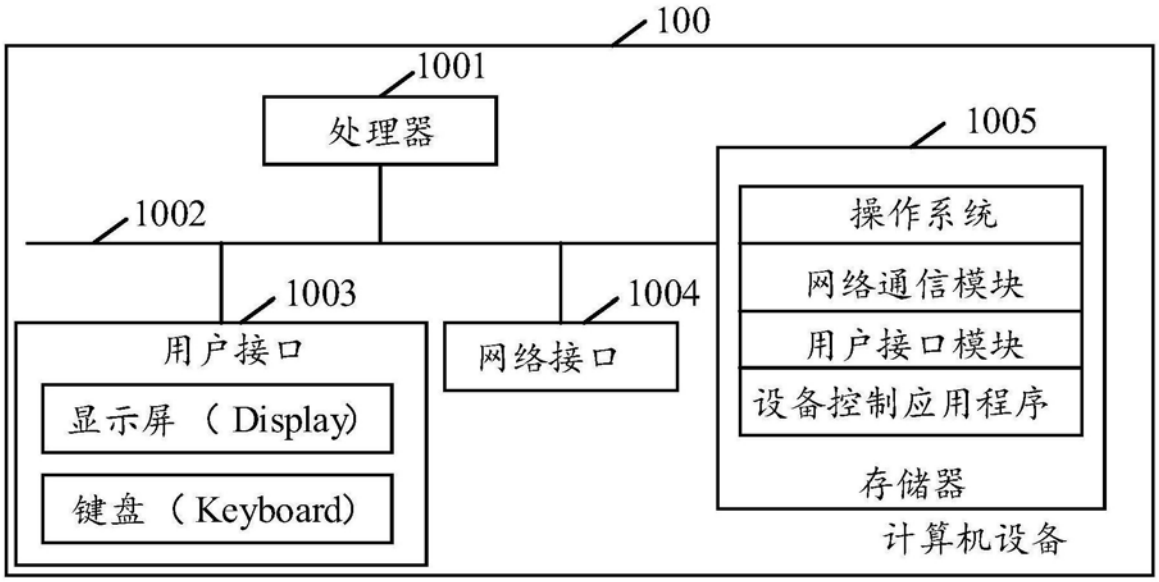

100
1001
处理器
1005
操作系统
1002
网络通信模块
用户接口模块
1003
1004
用户接口
网络接口
显示屏（Display）
键盘（Keyboard）
设备控制应用程序
存储器
计算机设备

图10