(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 111160049 B
(45)授权公告日 2023.06.06

(21)申请号 201911244875.5

(22)申请日 2019.12.06

(65)同一申请的已公布的文献号
申请公布号 CN 111160049 A

(43)申请公布日 2020.05.15

(73)专利权人 华为技术有限公司
地址 518129 广东省深圳市龙岗区坂田华为总部办公楼

(72)发明人 李良友 王龙跃 刘群 陈晓

(74)专利代理机构 北京龙双利达知识产权代理有限公司 11329
专利代理师 周乔 王君

(51)Int.Cl.
G06F 40/58(2020.01)
G06F 40/289(2020.01)
G06N 3/0455(2023.01)
G06N 3/0464(2023.01)
G06N 3/044(2023.01)
G06N 3/048(2023.01)
G06N 3/08(2023.01)

(56)对比文件
CN 109992629 A,2019.07.09
CN 110020440 A,2019.07.16
US 2019205761 A1,2019.07.04
Kai Song.Code-Switching for Enhancing NMT with Pre-Specified Translation.《paralarXiv: 1904.09107v4》.2019,第1-11页.
Yong Cheng.Agreement-based Joint Training for Bidirectional Attention-based Neural Machine Translation.《arXiv: 1512.04650v2》.2016,第1-7页.
毛曦;颜闻;马维军;殷红梅.注意力机制的英语地名机器翻译技术.测绘科学.2019,(06),全文.

审查员 叶珊

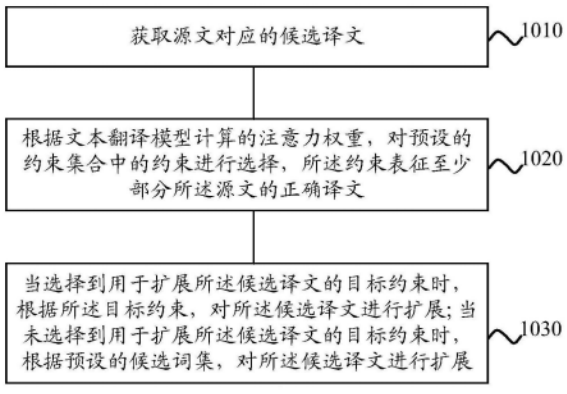
权利要求书2页 说明书20页 附图9页

(54)发明名称

文本翻译方法、装置、机器翻译系统和存储介质

(57)摘要

本申请公开了人工智能领域中的一种文本翻译方法和装置，该方法包括：获取候选译文；根据文本翻译模型计算的注意力权重，对预设的约束集合中的约束进行选择；当选择到用于扩展所述候选译文的目标约束时，根据所述目标约束，对所述候选译文进行扩展，或者，当未选择到用于扩展所述候选译文的目标约束时，根据预设的候选词集，对所述候选译文进行扩展。本申请在对候选译文进行扩展时，可以对预设的约束集合进行选择或者过滤，可以避免每次进行候选译文扩展时都使用全部约束，可以加快候选译文的扩展，从而提高翻译速度。



CN 111160049 B

1.一种文本翻译方法,其特征在于,包括:

获取源文对应的候选译文;

根据文本翻译模型计算的注意力权重,对预设的约束集合中的约束进行选择,所述约束表征至少部分所述源文的正确译文;

当选择到用于扩展所述候选译文的目标约束时,根据所述目标约束,对所述候选译文进行扩展;或者,

当未选择到用于扩展所述候选译文的目标约束时,根据预设的候选词集,对所述候选译文进行扩展,所述候选词集包括多个目标语言的词,所述目标语言为所述候选译文所属的语言。

2.根据权利要求1所述的方法,其特征在于,所根据文本翻译模型计算的注意力权重,对预设的约束集合中的约束进行选择,包括:

根据所述预设的约束集合中的每个约束对应的源端位置,从所述文本翻译模型获取分别对应于所述每个约束的注意力权重,所述源端位置为所述每个约束对应的词语在所述源文中的位置;

根据所述对应于所述每个约束的注意力权重,对所述预设的约束集合中的约束进行选择。

3.根据权利要求2所述的方法,其特征在于,所述根据所述对应于所述每个约束的注意力权重,对所述预设的约束集合中的约束进行选择,包括:

对所述对应于所述每个约束的注意力权重进行处理,得到所述每个约束的启发信号,所述启发信号用于指示在扩展所述候选译文时是否使用与所述启发信号对应的约束;

根据所述对应于所述每个约束的启发信号,对所述预设的约束集合中的约束进行选择。

4.根据权利要求1所述的方法,其特征在于,所根据文本翻译模型计算的注意力权重,对预设的约束集合中的约束进行选择,包括:

从所述注意力权重中选择满足预设要求的目标注意力权重;

根据所述目标注意力权重对应的源端位置,以及所述预设的约束集合中每个约束对应的源端位置,对预设的约束集合中的约束进行选择,所述目标注意力权重对应的源端位置为所述目标注意力权重对应的词语在所述源文中的位置,所述每个约束对应的源端位置为所述每个约束对应的词语在所述源文中的位置。

5.根据权利要求1所述的方法,其特征在于,所述根据文本翻译模型计算的注意力权重,对预设的约束集合中的约束进行选择,包括:

根据文本翻译模型计算的注意力权重,以及所述候选译文的状态,对预设的约束集合中的约束进行选择,所述候选译文的状态包括在约束中和不在约束中,其中,在所述候选译文是使用目标短语的部分词语扩展得到的情况下,所述候选译文的状态为在约束中,所述目标短语为所述预设的约束集合中的约束对应的目标端短语;

所述目标约束满足以下条件中的至少一个:

所述目标约束对应的注意力权重满足预设要求;

所述候选译文的状态为在约束中。

6.一种文本翻译装置,其特征在于,包括:

存储器，用于存储程序；

处理器，用于执行所述存储器存储的程序，当所述存储器存储的程序被所述处理器执行时，所述处理器用于：

获取源文对应的候选译文；

根据文本翻译模型计算的注意力权重，对预设的约束集合中的约束进行选择，所述约束表征至少部分所述源文的正确译文；

当选择到用于扩展所述候选译文的目标约束时，根据所述目标约束，对所述候选译文进行扩展；或者，

当未选择到用于扩展所述候选译文的目标约束时，根据预设的候选词集，对所述候选译文进行扩展，所述候选词集包括多个目标语言的词，所述目标语言为所述候选译文所属的语言。

7.根据权利要求6所述的装置，其特征在于，所述处理器具体用于：

根据所述预设的约束集合中的每个约束对应的源端位置，从所述文本翻译模型获取分别对应于所述每个约束的注意力权重，所述源端位置为所述每个约束对应的词语在所述源文中的位置；

根据所述对应于所述每个约束的注意力权重，对所述预设的约束集合中的约束进行选择。

8.根据权利要求7所述的装置，其特征在于，所述处理器具体用于：

对所述对应于所述每个约束的注意力权重进行处理，得到所述每个约束的启发信号，所述启发信号用于指示在扩展所述候选译文时是否使用与所述启发信号对应的约束；

根据所述对应于所述每个约束的启发信号，对所述预设的约束集合中的约束进行选择。

9.根据权利要求6所述的装置，其特征在于，所述处理器具体用于：

从所述注意力权重中选择满足预设要求的目标注意力权重；

根据所述目标注意力权重对应的源端位置，以及所述预设的约束集合中每个约束对应的源端位置，对预设的约束集合中的约束进行选择，所述目标注意力权重对应的源端位置为所述目标注意力权重对应的词语在所述源文中的位置，所述每个约束对应的源端位置为所述每个约束对应的词语在所述源文中的位置。

10.根据权利要求6所述的装置，其特征在于，所述处理器具体用于：

根据文本翻译模型计算的注意力权重，以及所述候选译文的状态，对预设的约束集合中的约束进行选择，所述候选译文的状态包括在约束中和不在约束中，其中，在所述候选译文是使用目标短语的部分词语扩展得到的情况下，所述候选译文的状态为在约束中，所述目标短语为所述预设的约束集合中的约束对应的目标端短语；

所述目标约束满足以下条件中的至少一个：

所述目标约束对应的注意力权重满足预设要求；

所述候选译文的状态为在约束中。

11.一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程序代码，所述程序代码包括用于执行如权利要求1至5中任一项所述的方法的部分或全部步骤的指令。

文本翻译方法、装置、机器翻译系统和存储介质

技术领域

[0001] 本申请涉及机器翻译技术领域，并且更具体地，涉及文本翻译方法、装置、机器翻译系统和存储介质。

背景技术

[0002] 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分支，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。

[0003] 随着人工智能技术的不断发展，让人机之间能够通过自然语言进行交互的自然语言人机交互系统变的越来越重要。人机之间能够通过自然语言进行交互，就需要系统能够识别出人类自然语言的具体含义。通常，系统通过采用对自然语言的句子进行关键信息提取来识别句子的具体含义。

[0004] 近年来，神经机器翻译取得了快速发展，并超越了传统的统计机器翻译，成为主流的机器翻译技术。很多公司，例如谷歌、百度、微软等，已经将神经机器翻译应用在了他们的翻译产品中。为了使得输入的源语言语句中的某些短语或词能被正确翻译，现在的神经机器翻译支持对神经机器翻译的翻译结果进行人工干预。一种方式是将已知的正确翻译作为约束加入到神经机器翻译中，并保证约束的目标端一定会出现在最后输出的译文中。

[0005] 因此，如何高效、准确地使用这些约束成为亟待解决的问题。

发明内容

[0006] 本申请提供一种文本翻译方法、装置、机器翻译系统和存储介质，在进行机器翻译时，能够高效、准确地使用约束，从而提高翻译速度。

[0007] 第一方面，本申请提供了一种文本翻译方法，该方法包括：获取源文对应的候选译文；根据文本翻译模型计算的注意力权重，对预设的约束集合中的约束进行选择，所述约束表征至少部分所述源文的正确译文；当选择到用于扩展所述候选译文的目标约束时，根据所述目标约束，对所述候选译文进行扩展；或者，当未选择到用于扩展所述候选译文的目标约束时，根据预设的候选词集，对所述候选译文进行扩展，所述候选词集包括多个目标语言的词，所述目标语言为所述候选译文所属的语言。

[0008] 可选地，当选择到用于扩展所述候选译文的目标约束时，根据所述目标约束和所述预设的候选词集，对所述候选译文进行扩展。

[0009] 上述候选译文是对源文进行翻译的中间结果或最终结果。例如，在将“我毕业于合工大”翻译成“I graduated from Hefei University of Technology”的过程中，“I”、“I graduated”、“I graduated from”等均是对“我毕业于合工大”进行翻译的中间结果，当“I”作为候选译文时，对“I”进行扩展得到的新的候选译文中可以包括“I graduated”，进一步

地，可以对得到的候选译文“I graduated”继续进行扩展，对“I graduated”进行扩展得到的新的候选译文中可以包括“I graduated from”；而得到候选译文“I graduated from Hefei University of Technology”后则不会在进行扩展，此时候选译文为对源文进行翻译的最终结果。

[0010] 应理解，本申请中，上述的候选译文可以为一个或者多个，对一个或者多个候选译文中的每一个候选译文进行扩展时，可以得到一个或者多个新的候选译文。

[0011] 本申请的文本翻译模型可以是基于神经网络的神经网络翻译模型，该神经网络翻译模型包括与注意力机制相关的部分，该与注意力机制相关的部分可以在翻译过程中计算得到相应的注意力权重。可选地，该神经网络模型可以包括编码器、解码器和与注意力机制相关的部分，其中，编码器用于读取源文，并为源文包括的每一个源语言词生成一个数值化的表示，解码器用于生成源文的译文，即目标语言的句子，与注意力机制相关的部分用于根据编码器的输出和解码器的状态，为解码器在不同时刻生成目标词时动态提供注意力权重。

[0012] 注意力权重可以为用于表征在当前时刻源文中的每个源语言词与解码器状态的相关程度。例如，在某个时刻，对应于源文的4个源语言词的注意力权重分别为0.5、0.3、0.1、0.1，则可以表明4个源语言词与解码器状态的相关程度分别为0.5、0.3、0.1、0.1，第一个源语言词与解码器状态的相关程度最高，则解码器当前正在生成第一个源语言词对应的目标词的可能性最大。

[0013] 候选词集包括多个目标语言的词，目标语言为候选译文所属的语言。候选词集可以为预设的候选词库，文本翻译模型在每个时刻都会对候选词库中的候选词进行打分，以便于后续根据各个候选词的分值，确定用于扩展候选译文的候选词。例如，可以使用分值超过预设阈值的候选词对候选译文进行扩展。

[0014] 本申请中的约束可以表征至少部分源文的正确翻译或者正确译文。可选地，该至少部分源文可以是源文包括的源语言词或者源语言短语。可选地，约束可以包括源端位置信息和与该源端位置信息对应的目标词，其中，源端指源文输入端，目标词为该源端位置信息指示的源端位置的源词的正确译文。可选地，约束的形式可以是：[源语言词在源文中的位置]:目标词，例如，[4]:Hefei University of Technology。可选地，约束的形式可以是：源语言词:目标词等形式，例如，合工大:Hefei University of Technology。

[0015] 上述源文和上述候选译文均可以属于自然语言，自然语言通常是指一种自然地随文化演化的语言。可选地，上述源文属于第一自然语言，上述候选译文属于第二自然语言，第一语言和第二语言为不同种类的自然语言。上述源文属于第一自然语言可以是指上述源文是采用第一自然语言表达的一段文字，上述候选译文属于第二自然语言，可以是指上述候选译文是采用第二自然语言表达的一段文字。上述源文和上述候选译文可以属于任意两种不同种类的自然语言。

[0016] 应理解，本申请的文本翻译方法的执行主体可以是文本翻译装置或者机器翻译系统。

[0017] 上述技术方案中，在对候选译文进行扩展时，根据文本翻译模型计算的注意力权重，对预设的约束集合中的约束进行选择或者过滤。当选择到目标约束时，才使用目标约束对候选译文进行扩展，当未选择到目标约束时，则不使用约束预设的约束集合中的约束对

候选译文进行扩展，即仅根据预设的候选词集对候选译文进行扩展。这样可以避免每次进行候选译文扩展时都使用全部约束，可以加快候选译文的扩展；并且由于注意力权重可以表征在当前时刻源文中的每个源语言词与解码器状态的相关程度，因此根据注意力权重对预设的约束进行选择，可以忽略与当前解码器的相关程度较低的约束，降低对候选译文质量的影响。因此，上述技术方案可以在保证候选译文的质量的同时加快候选译文的扩展，从而提高翻译速度。

[0018] 在一种可能的实现方式中，所根据文本翻译模型计算的注意力权重，对预设的约束集合中的约束进行选择，包括：根据所述预设的约束集合中的每个约束对应的源端位置，从所述文本翻译模型获取分别对应于所述每个约束的注意力权重，所述源端位置为所述每个约束对应的词语在所述源文中的位置；根据所述对应于所述每个约束的注意力权重，对所述预设的约束集合中的约束进行选择。

[0019] 应理解，这里所说的约束对应的词语为约束对应的源语言词。

[0020] 约束对应的源端位置可以通过约束中的源端位置信息确定。

[0021] 在一种可能的实现方式中，所述根据所述对应于所述每个约束的注意力权重，对所述预设的约束集合中的约束进行选择，包括：对所述对应于所述每个约束的注意力权重进行处理，得到所述每个约束的启发信号，所述启发信号用于指示在扩展所述候选译文时是否使用与所述启发信号对应的约束；根据所述对应于所述每个约束的启发信号，对所述预设的约束集合中的约束进行选择。

[0022] 在一种可能的实现方式中，所根据文本翻译模型计算的注意力权重，对预设的约束集合中的约束进行选择，包括：从所述注意力权重中选择满足预设要求的目标注意力权重；根据所述目标注意力权重对应的源端位置，以及所述预设的约束集合中每个约束对应的源端位置，对预设的约束集合中的约束进行选择，所述目标注意力权重对应的源端位置为所述目标注意力权重对应的词语在所述源文中的位置，所述每个约束对应的源端位置为所述每个约束对应的词语在所述源文中的位置。

[0023] 应理解，这里所说的目标注意力权重对应的词语为目标注意力权重对应的源语言词，同理，约束对应的词语为约束对应的源语言词。

[0024] 在一种可能的实现方式中，所述根据文本翻译模型计算的注意力权重，对预设的约束集合中的约束进行选择，包括：根据文本翻译模型计算的注意力权重，以及所述候选译文的状态，对预设的约束集合中的约束进行选择，所述候选译文的状态包括在约束中和不在约束中，其中，在所述候选译文是使用目标短语的部分词语扩展得到的情况下，所述候选译文的状态为在约束中，所述目标短语为所述预设的约束集合中的约束对应的目标端短语；所述目标约束满足以下条件中的至少一个：所述目标约束对应的注意力权重满足预设要求；所述候选译文的状态为在约束中。

[0025] 一些情况下，约束可能包括多个限定词，例如，与“合工大”对应的约束[4]:Hefei University of Technology包括Hefei、University、of和Technology 4个限定词。在逐词对候选译文进行扩展的场景下，存在当前时刻的待扩展的候选译文中可能已经使用了某个约束的部分限定词，此时候选译文的状态可以称为在约束中，例如，当前时刻的候选译文为“I graduated from Hefei”，已经使用了约束[4]:Hefei University of Technology中的“Hefei”。考虑到上述情况，上述技术方案可以将文本翻译模型计算的注意力权重和当前候

选译文的状态结合起来，对预设的约束集合中的约束进行选择，使得选择结果更加准确。

[0026] 可选地，当候选译文在约束中时，可以仅根据目标约束对候选译文进行扩展。

[0027] 可选地，当候选译文在约束中时，可以根据预设的候选词集和目标约束对候选译文进行扩展。

[0028] 第二方面，本申请提供了一种文本翻译装置，该装置包括存储器，用于存储程序；处理器，用于执行所述存储器存储的程序，当所述存储器存储的程序被所述处理器执行时，所述处理器用于：获取源文对应的候选译文；根据文本翻译模型计算的注意力权重，对预设的约束集合中的约束进行选择，所述约束表征至少部分所述源文的正确译文；当选择到用于扩展所述候选译文的目标约束时，根据所述目标约束，对所述候选译文进行扩展；或者，当未选择到用于扩展所述候选译文的目标约束时，根据预设的候选词集，对所述候选译文进行扩展，所述候选词集包括多个目标语言的词，所述目标语言为所述候选译文所属的语言。

[0029] 上述技术方案中，在对候选译文进行扩展时，根据文本翻译模型计算的注意力权重，对预设的约束集合中的约束进行选择或者过滤。当选择到目标约束时，才使用目标约束对候选译文进行扩展，当未选择到目标约束时，则不使用约束预设的约束集合中的约束对候选译文进行扩展，即仅根据预设的候选词集对候选译文进行扩展。这样可以避免每次进行候选译文扩展时都使用全部约束，可以加快候选译文的扩展；并且由于注意力权重可以表征在当前时刻源文中的每个源语言词与解码器状态的相关程度，因此根据注意力权重对预设的约束进行选择，可以忽略与当前解码器的相关程度较低的约束，降低对候选译文质量的影响。因此，上述技术方案可以在保证候选译文的质量的同时加快候选译文的扩展，从而提高翻译速度。

[0030] 在一种可能的实现方式中，所述处理器具体用于：根据所述预设的约束集合中的每个约束对应的源端位置，从所述文本翻译模型获取分别对应于所述每个约束的注意力权重，所述源端位置为所述每个约束对应的词语在所述源文中的位置；根据所述对应于所述每个约束的注意力权重，对所述预设的约束集合中的约束进行选择。

[0031] 应理解，这里所说的约束对应的词语为约束对应的源语言词。

[0032] 约束对应的源端位置可以通过约束中的源端位置信息确定。

[0033] 在一种可能的实现方式中，所述处理器具体用于：对所述对应于所述每个约束的注意力权重进行处理，得到所述每个约束的启发信号，所述启发信号用于指示在扩展所述候选译文时是否使用与所述启发信号对应的约束；根据所述对应于所述每个约束的启发信号，对所述预设的约束集合中的约束进行选择。

[0034] 在一种可能的实现方式中，所述处理器具体用于：从所述注意力权重中选择满足预设要求的目标注意力权重；根据所述目标注意力权重对应的源端位置，以及所述预设的约束集合中每个约束对应的源端位置，对预设的约束集合中的约束进行选择，所述目标注意力权重对应的源端位置为所述目标注意力权重对应的词语在所述源文中的位置，所述每个约束对应的源端位置为所述每个约束对应的词语在所述源文中的位置。

[0035] 应理解，这里所说的目标注意力权重对应的词语为目标注意力权重对应的源语言词，同理，约束对应的词语为约束对应的源语言词。

[0036] 在一种可能的实现方式中，所述处理器具体用于：根据文本翻译模型计算的注意

力权重，以及所述候选译文的状态，对预设的约束集合中的约束进行选择，所述候选译文的状态包括在约束中和不在约束中，其中，在所述候选译文是使用目标短语的部分词语扩展得到的情况下，所述候选译文的状态为在约束中，所述目标短语为所述预设的约束集合中的约束对应的目标端短语；所述目标约束满足以下条件中的至少一个：所述目标约束对应的注意力权重满足预设要求；所述候选译文的状态为在约束中。

[0037] 一些情况下，约束可能包括多个限定词，例如，与“合工大”对应的约束[4]:Hefei University of Technology包括Hefei、University、of和Technology 4个限定词。在逐词对候选译文进行扩展的场景下，存在当前时刻的待扩展的候选译文中可能已经使用了某个约束的部分限定词，此时候选译文的状态可以称为在约束中，例如，当前时刻的候选译文为“I graduated from Hefei”，已经使用了约束[4]:Hefei University of Technology中的“Hefei”。考虑到上述情况，上述技术方案可以将文本翻译模型计算的注意力权重和当前候选译文的状态结合起来，对预设的约束集合中的约束进行选择，使得选择结果更加准确。

[0038] 第三方面，本申请提供了一种文本翻译装置，该装置包括存储器，用于存储程序；处理器，用于执行所述存储器存储的程序，当所述存储器存储的程序被所述处理器执行时，所述文本翻译装置执行第一方面或第一方面任意一种可能的实现方式中的方法。

[0039] 可选地，所述装置还包括数据接口，所述处理器通过所述数据接口读取所述存储器上存储的程序。

[0040] 第四方面，本申请提供了一种机器翻译系统，该机器翻译系统包括第二方面或者第二方面的任意一种可能的实现方式中的文本翻译装置，其中，文本翻译装置用于执行第一方面或第一方面任意一种可能的实现方式中的方法。

[0041] 上述文本翻译装置可以是一种电子设备(或者是位于电子设备中的模块)，该电子设备具体可以是移动终端(例如，智能手机)，电脑，个人数字助理，可穿戴设备，车载设备，物联网设备或者其他能够进行自然语言处理的设备。

[0042] 第五方面，本申请提供一种计算机可读介质，该计算机可读介质存储用于设备执行的程序代码，该程序代码包括用于执行第一方面或第一方面任意一种可能的实现方式中的方法。

[0043] 第六方面，本申请提供一种包含指令的计算机程序产品，当该计算机程序产品在计算机上运行时，使得计算机执行上述第一方面或第一方面任意一种可能的实现方式中的方法。

[0044] 第七方面，本申请提供一种芯片，所述芯片包括处理器与数据接口，所述处理器通过所述数据接口读取存储器上存储的指令，执行第一方面或第一方面任意一种可能的实现方式中的方法。

[0045] 可选地，作为一种实现方式，所述芯片还可以包括存储器，所述存储器中存储有指令，所述处理器用于执行所述存储器上存储的指令，当所述指令被执行时，所述处理器用于执行第一方面中的方法。

[0046] 第八方面，本申请提供一种电子设备，该电子设备包括上述第二方面或者第二方面的任意一种可能的实现方式中的文本翻译装置，或者上述第三方面中的文本翻译装置，或者上述第四方面中的机器翻译系统。

附图说明

[0047] 图1是本申请实施例提供的一种自然语言处理的应用场景示意图。

[0048] 图2是本申请实施例提供的另一种自然语言处理的应用场景示意图。

[0049] 图3是本申请实施例提供的自然语言处理的相关设备的示意图。

[0050] 图4是本申请实施例提供的一种系统架构的示意图。

[0051] 图5是本申请实施例提供的一种RNN模型的示意图。

[0052] 图6是本申请实施例提供的一种芯片的硬件结构的示意图。

[0053] 图7是本申请实施例提供的一种基于神经网络的文本翻译模型的示意图。

[0054] 图8是本申请实施例提供的一种文本翻译过程的流程图。

[0055] 图9是本申请实施例提供的另一种文本翻译过程的示意图。

[0056] 图10是本申请实施例提供的文本翻译方法的示意性流程图。

[0057] 图11是本申请实施例提供的一种选取约束的方法的示意性流程图。

[0058] 图12是本申请实施例提供的文本翻译装置的示意性结构图。

[0059] 图13是本申请另一实施例提供的文本翻译装置的示意性结构图。

[0060] 图14是本申请实施例提供的机器翻译系统的示意性图。

具体实施方式

[0061] 下面将结合附图,对本申请中的技术方案进行描述。

[0062] 为了更好地理解本申请实施例的方案,下面先结合图1至图3对本申请实施例可能的应用场景进行简单的介绍。本申请实施例的技术方案可以应用于各种场景,只要在该场景中需要进行受限解码的序列生成任务。例如,机器翻译场景、自动生成文本摘要的过程等。下面以机器翻译场景为例,对本申请的技术方案进行描述。

[0063] 图1示出了一种自然语言处理系统,该自然语言处理系统包括用户设备以及数据处理设备。其中,用户设备包括手机、个人电脑或者信息处理中心等智能终端。用户设备为自然语言数据处理的发起端,作为语言问答或者查询等请求的发起方,通常用户通过用户设备发起请求。

[0064] 上述数据处理设备可以是云服务器、网络服务器、应用服务器以及管理服务器等具有数据处理功能的设备或服务器。数据处理设备通过交互接口接收来自智能终端的查询语句/语音/文本等问句,再通过存储数据的存储器以及数据处理的处理器环节进行机器学习,深度学习,搜索,推理,决策等方式的语言数据处理。数据处理设备中的存储器可以是一个统称,包括本地存储以及存储历史数据的数据库,数据库可以再数据处理设备上,也可以在其它网络服务器上。

[0065] 在图1所示的自然语言处理系统中,用户设备可以接收用户的指令,以请求对源文(例如,该源文可以是用户输入的一段中文)进行机器翻译得到机器译文(例如,该机器译文可以是机器翻译得到的英文),然后向数据处理设备发送源文,从而使得数据处理设备对源文进行翻译得到机器译文。

[0066] 在图1中,数据处理设备可以执行本申请实施例的文本翻译方法。

[0067] 图2示出了另一种自然语言处理系统,在图2中,用户设备直接作为数据处理设备,该用户设备能够直接接收来自用户的输入并直接由用户设备本身的硬件进行处理,具体过

程与图1相似,可参考上面的描述,在此不再赘述。

[0068] 在图2所示的自然语言处理系统中,用户设备可以接收用户的指令,由用户设备自身对源文进行机器翻译得到机器译文。

[0069] 在图2中,用户设备自身就可以执行本申请实施例的文本翻译方法。

[0070] 图3是本申请实施例提供的自然语言处理的相关设备的示意图。

[0071] 上述图1和图2中的用户设备具体可以是图3中的本地设备301或者本地设备302,图1中的数据处理设备具体可以是图3中的执行设备210,其中,数据存储系统250可以存储执行设备210的待处理数据,数据存储系统250可以集成在执行设备210上,也可以设置在云上或其它网络服务器上。

[0072] 图1和图2中的处理器可以通过神经网络模型或者其它模型(例如,基于支持向量机的模型)进行数据训练/机器学习/深度学习,并利用数据最终训练或者学习得到的模型对源文进行翻译,从而得到机器译文。

[0073] 图4示出了本申请实施例提供的一种系统架构100。在图4中,数据采集设备160用于采集训练数据,本申请实施例中训练数据包括训练源文及训练机器译文(训练源文经过机器翻译系统翻译得到的译文)。

[0074] 在采集到训练数据之后,数据采集设备160将这些训练数据存入数据库130,训练设备120基于数据库130中维护的训练数据训练得到目标模型/规则101。

[0075] 下面对训练设备120基于训练数据得到目标模型/规则101进行描述,训练设备120对输入的训练源文进行处理,将输出的机器译文与训练机器译文进行对比,直到训练设备120输出的机器译文与训练机器译文的差值小于一定的阈值,从而完成目标模型/规则101的训练。

[0076] 上述目标模型/规则101能够用于实现本申请实施例的文本翻译方法,即将源文通过相关预处理(可以采用预处理模块113和/或预处理模块114进行处理)后输入该目标模型/规则101,即可得到机器译文。本申请实施例中的目标模型/规则101具体可以为神经网络。需要说明的是,在实际的应用中,所述数据库130中维护的训练数据不一定都来自于数据采集设备160的采集,也有可能是从其他设备接收得到的。另外需要说明的是,训练设备120也不一定完全基于数据库130维护的训练数据进行目标模型/规则101的训练,也有可能从云端或其他地方获取训练数据进行模型训练,上述描述不应该作为对本申请实施例的限定。

[0077] 根据训练设备120训练得到的目标模型/规则101可以应用于不同的系统或设备中,如应用于图4所示的执行设备110,所述执行设备110可以是终端,如手机终端,平板电脑,笔记本电脑,增强现实(augmented reality,AR)/虚拟现实(virtual reality,VR),车载终端等,还可以是服务器或者云端等。在图4中,执行设备110配置输入/输出(input/output,I/O)接口112,用于与外部设备进行数据交互,用户可以通过客户设备140向I/O接口112输入数据,所述输入数据在本申请实施例中可以包括:客户设备输入的源文。

[0078] 预处理模块113和预处理模块114用于根据I/O接口112接收到的输入数据(如源文)进行预处理(具体可以是对源文进行处理,得到词向量),在本申请实施例中,也可以没有预处理模块113和预处理模块114(也可以只有其中的一个预处理模块),而直接采用计算模块111对输入数据进行处理。

[0079] 在执行设备110对输入数据进行预处理，或者在执行设备110的计算模块111执行计算等相关的处理过程中，执行设备110可以调用数据存储系统150中的数据、代码等以用于相应的处理，也可以将相应处理得到的数据、指令等存入数据存储系统150中。

[0080] 最后，I/O接口112将处理结果，例如，机器译文反馈给客户设备140。

[0081] 值得说明的是，训练设备120可以针对不同的下游系统，生成该下游系统对应的目标模型/规则101，该相应的目标模型/规则101即可以用于实现上述目标或完成上述任务，从而为用户提供所需的结果。

[0082] 在图4中所示情况下，用户可以手动给定输入数据(例如，输入一段文字)，该手动给定可以通过I/O接口112提供的界面进行操作。另一种情况下，客户设备140可以自动地向I/O接口112发送输入数据(例如，输入一段文字)，如果要求客户设备140自动发送输入数据需要获得用户的授权，则用户可以在客户设备140中设置相应权限。用户可以在客户设备140查看执行设备110输出的结果，具体的呈现形式可以是显示、声音、动作等具体方式(例如，输出结果可以是机器译文)。客户设备140也可以作为数据采集端，采集如图所示输入I/O接口112的输入数据及输出I/O接口112的输出结果作为新的样本数据，并存入数据库130。当然，也可以不经过客户设备140进行采集，而是由I/O接口112直接将如图所示输入I/O接口112的输入数据及输出I/O接口112的输出结果，作为新的样本数据存入数据库130。

[0083] 值得注意的是，图4仅是本申请实施例提供的一种系统架构的示意图，图中所示设备、器件、模块等之间的位置关系不构成任何限制。例如，在图4中，数据存储系统150相对执行设备110是外部存储器，在其它情况下，也可以将数据存储系统150置于执行设备110中。

[0084] 如图4所示，根据训练设备120训练得到目标模型/规则101，该目标模型/规则101可以是本申请实施例中的神经机器翻译模型，具体的，本申请实施例提供的神经网络可以是卷积神经网络(convolutional neural network，CNN)，深度卷积神经网络(deep convolutional neural network，DCNN)，循环神经网络(recurrent neural network，RNN)等等。

[0085] 由于RNN是一种非常常见的神经网络，下面结合图5重点对RNN的结构进行详细的介绍。

[0086] 图5是本申请实施例提供的一种RNN模型的结构示意图。其中，每个圆圈可以看作是一个单元，而且每个单元做的事情也是一样的，因此可以折叠成左半图的样子。用一句话解释RNN，就是一个单元结构的重复使用。

[0087] RNN是一个序列到序列的模型，假设x_{t-1}，x_t，x_{t+1}是一个输入：“我是中国”，那么o_{t-1}，o_t就应该对应“是”，“中国”这两个，预测下一个词最有可能是什么?就是o_{t+1}应该是“人”的概率比较大。

[0088] 因此，我们可以做这样的定义：

[0089] x_t：表示t时刻的输入，o_t：表示t时刻的输出，s_t：表示t时刻的记忆。因为当前时刻的输出是由记忆和当前时刻的输出决定的，就像你现在大四，你的知识是由大四学到的知识(当前输入)和大三以及大三以前学到的东西的(记忆)的结合，RNN在这点上也类似，神经网络最擅长做的就是通过一系列参数把很多内容整合到一起，然后学习这个参数，因此就定义了RNN的基础：

[0090] $s_t = f(U*x_t + W*s_{t-1})$

[0091] f()函数是神经网络中的激活函数,但为什么要加上它呢?举个例子,假如在大学学了非常好的解题方法,那初中那时候的解题方法还要用吗?显然是不用了的。RNN的想法也一样,既然能记忆了,那当然是只记重要的信息,其他不重要的,就肯定会忘记。但是在神经网络中什么最适合过滤信息呀?肯定是激活函数,因此在这里就套用一个激活函数,来做一个非线性映射,来过滤信息,这个激活函数可能为tanh或ReLU,也可为其他。

[0092] 假设大四快毕业了,要参加考研,请问参加考研是不是先记住你学过的内容然后去考研,还是直接带几本书去参加考研呢?很显然嘛,那RNN的想法就是预测的时候带着当前时刻的记忆s_t去预测。假如你要预测“我是中国“的下一个词出现的概率,这里已经很显然了,运用softmax来预测每个词出现的概率再合适不过了,但预测不能直接带用一个矩阵来预测,所有预测的时候还要带一个权重矩阵V,用公式表示为:

[0093] $ot=softmax(V*s_t)$

[0094] 其中,o_t就表示时刻t的输出。

[0095] 需要说明的是,如图5所示的RNN仅作为一种循环神经网络的示例,在具体的应用中,循环神经网络还可以以其他网络模型的形式存在。

[0096] 图6为本申请实施例提供的一种芯片的硬件结构的示意图。该芯片包括神经网络处理器(neural processing unit,NPU)50。该芯片可以被设置在如图4所示的执行设备110中,用以完成计算模块111的计算工作。该芯片也可以被设置在如图4所示的训练设备120中,用以完成训练设备120的训练工作并输出目标模型/规则101。如图5所示的循环神经网络中的算法可在如图6所示的芯片中得以实现。

[0097] 本申请实施例的文本翻译方法的具体可以在NPU 50中的运算电路503和/或向量计算单元507中执行,从而得到机器译文。

[0098] 下面对NPU 50中的各个模块和单元进行简单的介绍。

[0099] NPU 50作为协处理器可以挂载到主CPU(host CPU)上,由主CPU分配任务。NPU50的核心部分为运算电路503,在NUP 50工作时,NPU 50中的控制器504可以控制运算电路503提取存储器(权重存储器或输入存储器)中的数据并进行运算。

[0100] 在一些实现中,运算电路503内部包括多个处理单元(process engine,PE)。在一些实现中,运算电路503是二维脉动阵列。运算电路503还可以是一维脉动阵列或者能够执行例如乘法和加法这样的数学运算的其它电子线路。在一些实现中,运算电路503是通用的矩阵处理器。

[0101] 举例来说,假设有输入矩阵A,权重矩阵B,输出矩阵C。运算电路从权重存储器502中取矩阵B相应的数据,并缓存在运算电路中每一个PE上。运算电路从输入存储器501中取矩阵A数据与矩阵B进行矩阵运算,得到的矩阵的部分结果或最终结果,保存在累加器(accumulator)508中。

[0102] 向量计算单元507可以对运算电路的输出做进一步处理,如向量乘,向量加,指数运算,对数运算,大小比较等等。例如,向量计算单元507可以用于神经网络中非卷积/非全连接层(fully connected layers,FC)层的网络计算,如池化(pooling),批归一化(batch normalization),局部响应归一化(local response normalization)等。

[0103] 在一些实现种,向量计算单元507能将经处理的输出的向量存储到统一缓存器506。例如,向量计算单元507可以将非线性函数应用到运算电路503的输出,例如累加值的

向量,用以生成激活值。在一些实现中,向量计算单元507生成归一化的值、合并值,或二者均有。在一些实现中,处理过的输出的向量能够用作到运算电路503的激活输入,例如用于在神经网络中的后续层中的使用。

[0104] 统一存储器506用于存放输入数据以及输出数据。

[0105] 权重数据直接通过存储单元访问控制器505(direct memory access controller,DMAC)将外部存储器中的输入数据搬运到输入存储器501和/或统一存储器506、将外部存储器中的权重数据存入权重存储器502,以及将统一存储器506中的数据存入外部存储器。

[0106] 总线接口单元(bus interface unit,BIU)510,用于通过总线实现主CPU、DMAC和取指存储器509之间进行交互。

[0107] 与控制器504连接的取指存储器(instruction fetch buffer)509,用于存储控制器504使用的指令。

[0108] 控制器504,用于调用指存储器509中缓存的指令,实现控制该运算加速器的工作过程。

[0109] 一般地,统一存储器506,输入存储器501,权重存储器502以及取指存储器509均可以为片上(on-chip)存储器。NPU的外部存储器可以为该NPU外部的存储器,该外部存储器可以为双倍数据率同步动态随机存储器(double data rate synchronous dynamic random access memory,DDR SDRAM)、高带宽存储器(high bandwidth memory,HBM)或其他可读可写的存储器。

[0110] 图7是一种基于神经网络的文本翻译模型的示意图。如图7所示,该文本翻译模型包括编码器710、解码器720和与注意力机制相关的部分730。其中,编码器710用于读取的源文(例如,I am a student),并为源文包括的每一个源语言词(例如,I、am、a、student)生成一个数值化的表示;经过解码器720,源文被翻译成译文,即目标语言的句子(例如,Je suis étudiant);与注意力机制相关的部分730用于根据编码器的输出和解码器的状态,为解码器在不同时刻生成目标词时动态提供注意力权重,其中解码器状态可以通过解码器的中间输出表征。

[0111] 其中,注意力权重可以为用于表征在当前时刻源文中的每个源语言词与解码器状态的相关程度。例如,在某个时刻,对应于源文的4个源语言词的注意力权重分别为0.5、0.3、0.1、0.1,则可以表明4个源语言词与解码器状态的相关程度分别为0.5、0.3、0.1、0.1,第一个源语言词与解码器状态的相关程度最高,则解码器当前正在生成第一个源语言词对应的目标词的可能性最大。

[0112] 为了使得输入的源文中的某些源语言词能被正确翻译,现在的神经机器翻译支持对神经机器翻译的翻译结果进行人工干预。一种方式是将已知的正确翻译作为约束加入到神经机器翻译中,并保证约束包括的目标词一定会出现在最后输出的译文中。

[0113] 图8是一种文本翻译过程的流程图。图8所示的文本翻译模型可以为图7中所示的文本翻译模型,在该流程中,可以将已知的正确翻译作为约束加入到神经机器翻译中。具体地,每一个时刻有K个候选译文作为输入,通过文本翻译模型对每个候选译文进行解码,得到预设的候选词集中每个候选词的分值;根据得到的候选词的分值和预设的约束集合,对候选译文进行扩展,得到新候选译文集;然后从新候选译文集合中选取一定数量作为下一

步的输入;若满足解码终止条件,则输出翻译结果,即源文的译文,若不满足解码终止条件,则对选取的一定数量的候选译文进行进一步扩展。

[0114] 其中,上述约束集合包括一个或多个约束,约束可以表征至少部分源文的正确翻译或者正确译文。可选地,该至少部分源文可以是源文包括的源语言词或者源语言短语。约束集合可以是用户手动给定的。上述候选译文是对源文进行翻译的中间结果或最终结果。例如,将“我毕业于合工大”翻译成英文“I graduated from Hefei University of Technology”,约束集合中包括“哈工大”的正确翻译“Hefei University of Technology”,在翻译过程中,“I”、“I graduated”、“I graduated from”等均是对“我毕业于合工大”进行翻译的中间结果,当“I”作为候选译文时,对“I”进行扩展得到的新的候选译文中可以包括“I graduated”“I Hefei”等,进一步地,可以对得到的新候选译文继续进行扩展;而得到候选译文“I graduated from Hefei University of Technology”后则不会在进行扩展,此时候选译文为对源文进行翻译的最终结果。

[0115] 上述生成机器译文的方法使用约束集合中约束扩展候选译文,同时在选取阶段也会考虑新候选译文对约束集合的覆盖情况,从而保证约束的目标词一定会出现在最后输出的译文中。但是上述生成机器译文的方法中,会使用约束集合中的全部约束对每一个候选译文进行扩展,例如,在产生“I”的过程中,也把“Hefei”加入到新候选译文集,在产生“I graduated”的过程中,也会生成“I Hefei”、“Hefei University”等,这样显然会带来时间和空间的浪费。因此,如何高效、准确地使用这些约束成为亟待解决的问题。

[0116] 针对上述问题,本申请实施例提供了文本翻译的方法和装置,能够在进行机器翻译时,能够高效、准确地使用约束,从而提高翻译速度,降低对空间的浪费。

[0117] 下面结合附图对本申请实施例的文本翻译方法进行详细介绍。本申请实施例的文本翻译方法可以由图1中的数据处理设备、图2中的用户设备、图3中的执行设备210以及图4中的执行设备110等设备执行。

[0118] 图9是本申请实施例提供的文本翻译过程的示意图。如图9所示,与图8所示的文本翻译过程相比,增加了选取约束的步骤,用于对约束集合进行过滤。在图8所示的过程中,在所有时刻对所有候选译文均使用相对于源文的完整约束集合,而图9所示的过程中,在当前时刻可以选取与当前时刻待扩展的候选译文相关的约束,形成一个新约束集合,新约束集合是完整约束集合的一个子集,进而使用新约束集合对当前时刻的待扩展候选译文进行扩展。这样,图9所示的过程可以避免每次进行候选译文扩展时都使用全部约束,可以加快候选译文的扩展,从而提高翻译速度。

[0119] 图10是本申请实施例提供的文本翻译方法的示意性流程图。本申请实施例的文本翻译方法可以由图1中的数据处理设备、图2中的用户设备、图3中的执行设备210以及图4中的执行设备110等设备执行。

[0120] 在1010中,获取源文对应的候选译文。

[0121] 其中,所述候选译文是当前时刻待扩展的候选译文,可以采用目标语言描述。

[0122] 在1020中,根据文本翻译模型输出的注意力权重,对预设的约束集合中的约束进行选择。

[0123] 注意力权重可以为用于表征源文中的每个源语言词与解码器状态的相关程度。例如,在某个时刻,对应于源文的4个源语言词的注意力权重分别为0.5、0.3、0.1、0.1,则可以

表明4个源语言词与解码器状态的相关程度分别为0.5、0.3、0.1、0.1，第一个源语言词与解码器状态的相关程度最高，则解码器当前正在生成第一个源语言词对应的目标词的可能性最大。

[0124] 注意力权重为文本翻译模型当前时刻计算得到的注意力权重。也就是说，根据当前时刻文本翻译模型输出的注意力权重，对预设的约束集合中的约束进行选择。

[0125] 本申请中的约束可以表征至少部分源文的正确翻译或者正确译文。可选地，该至少部分源文可以是源文包括的源语言词或者源语言短语。

[0126] 在一些实施例中，约束可以包括源端信息和目标端信息，其中，源端可以为源文输入端，目标端可以为译文输出端。

[0127] 例如，约束的形式可以是：源语言词：目标词的形式，例如，合工大：Hefei University of Technology。

[0128] 又例如，约束包括的源端信息为源端位置信息，目标端信息为与该源端位置信息对应的目标词。例如，约束的形式可以是：[源语言词在源文中的位置]：目标词，例如，[4]：Hefei University of Technology。

[0129] 本申请中，根据文本翻译模型输出的注意力权重对预设的约束集合中的约束进行选择的方式有很多，本申请不作具体限定。

[0130] 在一些实施例中，根据所述预设的约束集合中的每个约束对应的源端位置，从所述文本翻译模型获取分别对应于所述每个约束的注意力权重，并根据获取到的注意力权重，对所述预设的约束集合中的约束进行选择。其中，约束对应的源端位置是指约束对应的源文中的位置。约束包括的源端位置信息可以指示约束表征的是源文中哪个位置的源词的正确译文。

[0131] 例如，约束集合为{[4]：Hefei University of Technology，[6]：Sushma Swaraj}，两个约束分别对应于源端位置4和源端位置6，则从文本翻译模型获取源端位置4和源端位置6对应的注意力权重，根据获取到的权重，对两个约束进行选择。

[0132] 可选地，从文本翻译模型获取到分别对应于约束集合中每个约束的注意力权重之后，可以对获取到的注意力权重进行处理，得到所述每个约束的启发信号，根据每个约束的启发信号，对预设的约束集合中的约束进行选择。其中，启发信号用于指示在扩展所述候选译文时是否使用与所述启发信号对应的约束。例如，可以将获取到的注意力权重分别与预设阈值比较，确定大于或者等于预设阈值的注意力权重所对应的约束的启发信号指示在扩展当前时刻的候选译文时使用该约束，确定小于预设阈值的注意力权重所对应的约束的启发信号指示在扩展当前时刻的候选译文时不使用该约束。又例如，当文本翻译模型在每个时刻对于同一源端位置计算多个注意力权重时，还可以对获取到的对应于同一源端位置的多个注意力权重进行处理，例如，对该多个注意力权重求和、求平均、取最大的前Q个或者其他更复杂的处理；根据处理结果进步确定每个约束的启发信号。

[0133] 在另一些实施例中，可以从文本翻译模型获取当前时刻每个源端位置对应的注意力权重，再从获取到的注意力权重中选择满足预设要求的目标注意力权重，进而根据目标注意力权重对应的源端位置，以及预设的约束集合中每个约束对应的源端位置，对预设的约束集合中的约束进行选择。这里所说的源端位置为源文中的位置，具体地，目标注意力权重对应的源端位置为目标注意力权重对应的词语在源文中的位置，约束对应的源端位置为

约束对应的词语在所述源文中的位置。

[0134] 例如，约束集合为{[4]:Hefei University of Technology,[6]:Sushma Swaraj}，从文本翻译模型获取的源端位置1到源端位置6对应的当前时刻的注意力权重分别为0.01、0.01、0.01、0.95、0.01、0.01，将6个注意力权重与预设阈值0.5比较，得到目标注意力权重包括源端位置4对应的注意力权重，从约束集合中选择源端位置为源端位置4的约束，即[4]:Hefei University of Technology，进而使用约束[4]:Hefei University of Technology对候选译文进行扩展。

[0135] 在一些实施例中，一些约束可能包括多个限定词，例如，与“合工大”对应的约束[4]:Hefei University of Technology包括Hefei、University、of和Technology 4个限定词。在逐词对候选译文进行扩展的场景下，存在当前时刻的待扩展的候选译文中可能已经使用了某个约束的部分限定词，此时候选译文的状态可以称为在约束中，例如，当前时刻的候选译文为“I graduated from Hefei”，已经使用了约束[4]:Hefei University of Technology中的“Hefei”。考虑到上述情况，本申请可以将文本翻译模型计算的注意力权重和当前候选译文的状态结合起来，对预设的约束集合中的约束进行选择。

[0136] 在一种可能的实现方式中，可以根据文本翻译模型计算的注意力权重，以及候选译文的状态，对预设的约束集合中的约束进行选择，用于扩展当前时刻的候选译文的目标约束满足以下条件中的至少一个：目标约束对应的注意力权重满足预设要求；候选译文在约束中。

[0137] 其中，当使用目标短语的部分词语扩展得到当前时刻待扩展的候选译文时，待扩展的候选译文在约束中，目标短语为预设的约束集合中的某个约束对应的短语。

[0138] 上述预设要求可以为大于预设阈值、使得对应的约束的启发信号指示在扩展当前时刻的候选译文时使用该约束等，本申请实施例不作具体限定。

[0139] 下面结合具体的例子，对本申请实施例提供的选取约束的方法进行描述。图11是本申请实施例提供的一种选取约束的方法的示意性流程图。图11中的N表示注意力权重的个数，M表示预设的约束集合中约束的个数。

[0140] 如图11所示，在1110中，计算预设的约束集合中每个约束的置信度。

[0141] 具体地，根据M个约束中的第k个约束包括的n个源端位置信息，从N个注意力权重$\{aw_1,...,aw_N\}$中提取出相应位置的注意力权重$\{aw_{i,j}|1\leqslant i\leqslant N,1\leqslant j\leqslant n\}$，i表示第i个注意力权重，j表示第j个源端位置，k大于等于1且小于等于M；根据公式$\{c_1,...,c_L\}=f(\{aw_{i,j}\})$计算得到L个置信度$\{c_1,...,c_L\}$。

[0142] 其中，函数f可以是简单函数。例如，函数f为求和函数、求平均函数等。函数f也可以是复杂函数。例如，函数f可以是一个神经网络等。

[0143] 在1120中，根据1110中得到的L个置信度，计算约束k的启发信号。

[0144] 具体地，可以根据公式$h_k=g(\{c_1,...,c_L\})$计算得到约束k的启发信号，其中h_k表示约束k的启发信号。启发信号h_k可以具有两个值，分别表示是否在当前候选译文扩展时使用约束k。例如，启发信号的两个值为1和0，当启发信号取值为1时，表示在当前候选译文扩展时使用约束k，当启发信号为0时，表示在当前候选译文扩展时不使用约束k。

[0145] 其中，函数g可以是简单函数。例如，求和、求平均等，然后再与一个预设阈值比较，如果大于预设阈值则返回1，否则返回0。函数g也可以是复杂函数。例如，一个输出1或0的神

经网络等。

[0146] 在1130中，根据1120中得到的每个约束的启发信号，以及候选译文的状态，对预设的约束集合进行过滤或选择。

[0147] 具体地，假设用s_k表示候选译文的状态，当$s_k=0$表示当前候选译文处在第k个约束中，$s_k=1$表示当前候选译文不在第k个约束中，则新约束集合可以表示为：

[0148] $\{k|s_k=0或h_k=1\}$

[0149] 即，如果第k个约束满足下面两个条件中的任何一个，第k个约束为用于扩展当前时刻的候选译文的目标约束，加入到新约束集合中：

[0150] 条件1：当前候选译文处在此约束中；

[0151] 条件2：此约束的启发信号为1。

[0152] 本申请的技术方案中，在对候选译文进行扩展之前，对预设的约束集合中的约束进行选择或者过滤，可以忽略与当前解码器状态的相关程度较低的约束，从而在不影响对候选译文质量的同时加快翻译速度。

[0153] 在1030中，当选择到用于扩展所述候选译文的目标约束时，根据所述目标约束对所述候选译文进行扩展。或者当未选择到用于扩展所述候选译文的目标约束时，根据预设的候选词集，对所述候选译文进行扩展。

[0154] 候选词集包括多个目标语言的词，目标语言为候选译文所属的语言。候选词集可以为预设的候选词库，文本翻译模型在每个时刻都会对候选词库中的候选词进行打分，以便于后续根据各个候选词的分值，确定用于扩展候选译文的候选词。例如，可以使用分值超过预设阈值的候选词对候选译文进行扩展。

[0155] 可选地，当选择到用于扩展所述候选译文的目标约束时，可以仅根据所述目标约束，对所述候选译文进行扩展，也可以根据所述目标约束和所述预设的候选词集，对所述候选译文进行扩展，本申请实施例不作具体限定。

[0156] 例如，当前时刻的候选译文为“I graduated from”，约束集合为{[4]:Hefei University of Technology,[6]:Sushma Swaraj}，根据注意力权重以及候选译文的状态，选择到用于扩展候选译文的目标约束[4]:Hefei University of Technology，则根据预设的候选词集和目标约束[4]:Hefei University of Technology对候选译文“I graduated from”进行扩展，而不使用约束[6]:Sushma Swaraj对候选译文“I graduated from”进行扩展。

[0157] 又例如，当前时刻的候选译文为“I graduated”，约束集合为{[4]:Hefei University of Technology,[6]:Sushma Swaraj}，根据注意力权重以及候选译文的状态，未选择到用于扩展候选译文的目标约束，则仅根据预设的候选词集对候选译文“I graduated”进行扩展。

[0158] 应理解，上述源文、上述候选译文和上述译文均可以属于自然语言，自然语言通常是指一种自然地随文化演化的语言。可选地，上述源文属于第一自然语言，上述候选译和上述译文属于第二自然语言，第一语言和第二语言为不同种类的自然语言。上述源文属于第一自然语言可以是指上述源文是采用第一自然语言表达的一段文字，上述候选译文和上述译文属于第二自然语言，可以是指上述候选译文和上述译文是采用第二自然语言表达的一段文字。上述源文和上述候选译文可以属于任意两种不同种类的自然语言。

[0159] 下面结合具体的例子,对本申请的文本翻译方法进行更详细地描述。

[0160] 示例1

[0161] 源文是“我毕业于合工大”,包含“我”、“毕业”、“于”、“合工大”4个源语言词。

[0162] 约束集合为“{[4]:Hefei University of Technology}”,只包含有一个约束,其对应于第4个源语言词“合工大”的正确翻译。

[0163] 一个正确的译文“I graduated from Hefei University of Technology”的生成过程可能如下。

[0164] 1)在生成前三个源语言词对应的目标词时,约束的启发信号均为0,表示不使用约束扩展。

[0165] 例如,对输入的候选译文“I graduated”进行扩展,获取注意力权重,4个源端位置对应的注意力权重分别为[0.01,0.01,0.97,0.01],得到约束[4]的置信度为0.01,低于预先设定的阈值0.5,产生启发信号0,根据预设的候选词集对“I graduated”进行扩展,选取后得到新候选译文“I graduated from”。

[0166] 2)对输入的候选译文“I graduated from”进行扩展。

[0167] 获取注意力权重,4个源端位置对应的注意力权重分别为[0.01,0.01,0.01,0.97],得到约束[4]的置信度为0.97,高于预先设定的阈值0.5,产生启发信号1,根据预设的候选词集和约束[4]中的第一个限定词“Hefei”对候选译文“I graduated from”进行扩展,选取得到“I graduated from Hefei”。

[0168] 3)对输入的候选译文“I graduated from Hefei”进行扩展。

[0169] 获取注意力权重,4个源端位置对应的注意力权重分别为[0.25,0.25,0.25,0.25],得到约束[4]的置信度为0.25,低于预先设定的阈值0.5,产生启发信号0。但是因为当前候选译文处在约束中,因此,依然根据约束[4]中的第二个限定词“University”对候选译文“I graduated from Hefei”进行扩展,选取后得到“I graduated from Hefei University”。

[0170] 4)继续扩展,因为候选译文“I graduated from Hefei University”处在约束中,所以依次使用约束[4]中的第三个限定词“of”和第四个限定词“Technology”对候选译文进行扩展,最终得到“I graduated from Hefei University of Technology”。

[0171] 5)解码终止,输出翻译结果“I graduated from Hefei University of Technology”。

[0172] 示例2

[0173] 源文是“我毕业于合工大”,包含“我”、“毕业”、“于”、“合工大”4个源语言词。

[0174] 约束集合为“{[4]:Hefei University of Technology}”,只包含有一个约束,其对应于第4个源语言词“合工大”的正确翻译。

[0175] 一个正确的译文“I graduated from Hefei University of Technology”的生成过程可能如下。

[0176] 1)在生成前三个目标词时,约束的启发信号均为0,表示不使用约束扩展。

[0177] 例如,对输入的候选译文“I graduated”进行扩展,获取注意力权重,4个源端位置对应的注意力权重分别为[0.01,0.01,0.97,0.01],得到约束[4]的置信度为0.01,低于预先设定的阈值0.5,产生启发信号0,根据预设的候选词集对“I graduated”进行扩展,选取

后得到新候选译文“I graduated from”。

[0178] 2)对输入的候选译文“I graduated from”进行扩展。

[0179] 获取注意力权重,4个源端位置对应的注意力权重分别为[0.01,0.01,0.01,0.97],得到约束[4]的置信度为0.97,高于预先设定的阈值0.5,产生启发信号1,根据预设的候选词集和约束[4]的全部限定词对候选译文“I graduated from”进行扩展,选取得到“I graduated from Hefei University of Technology”。

[0180] 3)解码终止,输出翻译结果“I graduated from Hefei University of Technology”。

[0181] 本申请的技术方案中,在对候选译文进行扩展时,根据选择或者过滤后得到的目标约束对候选译文进行扩展,这样可以避免每次进行候选译文扩展时都使用全部约束,并且在未选择到目标约束时,仅根据预设的候选词集对候选译文进行扩展,可以避免在不需要使用约束的时候仍然使用约束对候选译文进行扩展。这样,本申请的技术方案可以加快候选译文的扩展,从而进一步提高翻译速度。

[0182] 上文结合附图对本申请方法实施例进行了详细的描述,下面结合图12至图14对本申请装置实施例进行描述,应理解,图12至图14中描述的各装置能够执行本申请实施例的文本翻译方法的各个步骤,下面在介绍本申请的装置实施例时适当省略重复的描述。

[0183] 图12是本申请实施例提供的文本翻译装置的示意性结构图。该文本翻译装置1200可以相当于是图1所示的数据处理设备或者图2所示的用户设备。文本翻译装置1200还可以相当于图3所示的执行设备210、图4所示的执行设备110。

[0184] 装置1200可以包括获取模块1210和处理模块1220。其中,装置1200包括的各个模块可以通过软件和/或硬件方式实现。

[0185] 可选地,所述获取模块1210可以是通信接口,或者,所述获取模块1210和所述处理模块1220可以为同一个模块。

[0186] 在本申请中,装置1200可以用于执行图10所描述的方法中的步骤。

[0187] 例如:

[0188] 获取模块1210,用于获取源文对应的候选译文;

[0189] 处理模块1220,用于根据文本翻译模型计算的注意力权重,对预设的约束集合中的约束进行选择,所述约束表征至少部分所述源文的正确译文;

[0190] 处理模块1220,还用于当选择到用于扩展所述候选译文的目标约束时,根据所述目标约束,对所述候选译文进行扩展;或者,当未选择到用于扩展所述候选译文的目标约束时,根据预设的候选词集,对所述候选译文进行扩展。

[0191] 可选地,处理模块1220,具体用于根据所述预设的约束集合中的每个约束对应的源端位置,从所述文本翻译模型获取分别对应于所述每个约束的注意力权重,所述源端位置为所述每个约束对应的词语在所述源文中的位置;根据所述对应于所述每个约束的注意力权重,对所述预设的约束集合中的约束进行选择。

[0192] 可选地,处理模块1220,具体用于对所述对应于所述每个约束的注意力权重进行处理,得到所述每个约束的启发信号,所述启发信号用于指示在扩展所述候选译文时是否使用与所述启发信号对应的约束;根据所述对应于所述每个约束的启发信号,对所述预设的约束集合中的约束进行选择。

[0193] 可选地，处理模块1220，具体用于从所述注意力权重中选择满足预设要求的目标注意力权重；根据所述目标注意力权重对应的源端位置，以及所述预设的约束集合中每个约束对应的源端位置，对预设的约束集合中的约束进行选择，所述目标注意力权重对应的源端位置为所述目标注意力权重对应的词语在所述源文中的位置，所述每个约束对应的源端位置为所述每个约束对应的词语在所述源文中的位置。

[0194] 可选地，处理模块1220，具体用于根据文本翻译模型计算的注意力权重，以及所述候选译文的状态，对预设的约束集合中的约束进行选择，所述候选译文的状态包括在约束中和不在约束中，其中，在所述候选译文是使用目标短语的部分词语扩展得到的情况下，所述候选译文的状态为在约束中，所述目标短语为所述预设的约束集合中的约束对应的目标端短语；

[0195] 所述目标约束满足以下条件中的至少一个：

[0196] 所述目标约束对应的注意力权重满足预设要求；

[0197] 所述候选译文的状态为在约束中。

[0198] 应理解，图12示出的文本翻译装置1200仅是示例，本申请实施例的装置还可包括其他模块或单元。

[0199] 获取模块1210可以由通信接口或者处理器实现。处理模块1220可以由处理器实现。获取模块1210和处理模块1220的具体功能和有益效果可以参见方法实施例的相关描述，在此就不再赘述。

[0200] 图13是本申请另一实施例提供的文本翻译装置的示意性结构图。其中，文本翻译装置1300可以相当于是图1所示的数据处理设备或者图2所示的用户设备。文本翻译装置1300还可以相当于图3所示的执行设备210、图4所示的执行设备110。

[0201] 如图13所示，文本翻译装置1300可以包括存储器1310和处理器1320。图13中仅示出了一个存储器和处理器。在实际的文本翻译装置产品中，可以存在一个或多个处理器和一个或多个存储器。存储器也可以称为存储介质或者存储设备等。存储器可以是独立于处理器设置，也可以是与处理器集成在一起，本申请实施例对此不做限制。

[0202] 存储器1310和处理器1320之间通过内部连接通路互相通信，传递控制和/或数据信号。

[0203] 具体地，存储器1310，用于存储程序；

[0204] 处理器1320，用于执行所述存储器1310存储的程序，当所述存储器1310存储的程序被所述处理器1320执行时，所述处理器1320用于：

[0205] 获取源文对应的候选译文；

[0206] 根据文本翻译模型计算的注意力权重，对预设的约束集合中的约束进行选择，所述约束表征至少部分所述源文的正确译文；

[0207] 当选择到用于扩展所述候选译文的目标约束时，根据所述目标约束，对所述候选译文进行扩展；或者，当未选择到用于扩展所述候选译文的目标约束时，根据预设的候选词集，对所述候选译文进行扩展。

[0208] 上述文本翻译装置1300还可以包含输入输出接口1330，文本翻译装置1300通过输入输出接口1330能够获取源文，具体地，通过输入输出接口1330可以从其他设备(例如，终端设备)获取源文，在获取到源文之后，通过处理器1320的处理能够最终得到机器译文。文

本翻译装置1300通过输入输出接口1330能够将机器译文传输给其他设备。

[0209] 应理解,图13示出的文本翻译装置1300仅是示例,本申请实施例的装置还可包括其他模块或单元。

[0210] 文本翻译装置1300的具体工作过程和有益效果可以参见方法实施例中的相关描述,在此不再赘述。

[0211] 图14是本申请实施例提供的机器翻译系统1400的示意性图。

[0212] 其中,机器翻译系统1400可以相当于是图1所示的数据处理设备或者图2所示的用户设备。机器翻译系统1400还可以相当于图3所示的执行设备210、图4所示的执行设备110。

[0213] 如图14所示,机器翻译系统1400可以包括存储器1410和处理器1420。图14中仅示出了一个存储器和处理器。在实际的机器翻译系统产品中,可以存在一个或多个处理器和一个或多个存储器。存储器也可以称为存储介质或者存储设备等。存储器可以是独立于处理器设置,也可以是与处理器集成在一起,本申请实施例对此不做限制。

[0214] 存储器1410和处理器1420之间通过内部连接通路互相通信,传递控制和/或数据信号。

[0215] 具体地,存储器1410,用于存储程序;

[0216] 处理器1420,用于执行所述存储器1410存储的程序,当所述存储器1410存储的程序被所述处理器1420执行时,所述处理器1420用于:

[0217] 获取源文对应的候选译文;

[0218] 根据文本翻译模型计算的注意力权重,对预设的约束集合中的约束进行选择,所述约束表征至少部分所述源文的正确译文;

[0219] 当选择到用于扩展所述候选译文的目标约束时,根据所述目标约束,对所述候选译文进行扩展;或者,当未选择到用于扩展所述候选译文的目标约束时,根据预设的候选词集,对所述候选译文进行扩展。

[0220] 上述机器翻译系统1400还可以包含输入输出接口1430,文本翻译装置1400通过输入输出接口1430能够获取源文,具体地,通过输入输出接口1430可以从其他设备(例如,终端设备)获取源文,在获取到源文之后,通过处理器1420的处理能够最终得到机器译文。机器翻译系统1400通过输入输出接口1430能够将机器译文传输给其他设备。

[0221] 应理解,图14示出的机器翻译系统1400仅是示例,本申请实施例的机器翻译系统还可包括其他模块或单元。

[0222] 机器翻译系统1400的具体工作过程和有益效果可以参见方法实施例中的相关描述,在此不再赘述。

[0223] 应理解,本申请实施例中的处理器可以为中央处理单元(central processing unit,CPU),该处理器还可以是其他通用处理器、数字信号处理器(digital signal processor,DSP)、专用集成电路(application specific integrated circuit,ASIC)、现成可编程门阵列(field programmable gate array,FPGA)或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等。通用处理器可以是微处理器或者该处理器也可以是任何常规的处理器等。

[0224] 还应理解,本申请实施例中的存储器可以是易失性存储器或非易失性存储器,或可包括易失性和非易失性存储器两者。其中,非易失性存储器可以是只读存储器(read-

only memory,ROM)、可编程只读存储器(programmable ROM,PROM)、可擦除可编程只读存储器(erasable PROM,EPROM)、电可擦除可编程只读存储器(electrically EPROM,EEPROM)或闪存。易失性存储器可以是随机存取存储器(random access memory,RAM),其用作外部高速缓存。通过示例性但不是限制性说明,许多形式的随机存取存储器(random access memory,RAM)可用,例如静态随机存取存储器(static RAM,SRAM)、动态随机存取存储器(DRAM)、同步动态随机存取存储器(synchronous DRAM,SDRAM)、双倍数据速率同步动态随机存取存储器(double data rate SDRAM,DDR SDRAM)、增强型同步动态随机存取存储器(enhanced SDRAM,ESDRAM)、同步连接动态随机存取存储器(synchlink DRAM,SLDRAM)和直接内存总线随机存取存储器(direct rambus RAM,DR RAM)。

[0225] 上述实施例,可以全部或部分地通过软件、硬件、固件或其他任意组合来实现。当使用软件实现时,上述实施例可以全部或部分地以计算机程序产品的形式实现。所述计算机程序产品包括一个或多个计算机指令或计算机程序。在计算机上加载或执行所述计算机指令或计算机程序时,全部或部分地产生按照本申请实施例所述的流程或功能。所述计算机可以为通用计算机、专用计算机、计算机网络、或者其他可编程装置。所述计算机指令可以存储在计算机可读存储介质中,或者从一个计算机可读存储介质向另一个计算机可读存储介质传输,例如,所述计算机指令可以从一个网站站点、计算机、服务器或数据中心通过有线(例如红外、无线、微波等)方式向另一个网站站点、计算机、服务器或数据中心进行传输。所述计算机可读存储介质可以是计算机能够存取的任何可用介质或者是包含一个或多个可用介质集合的服务器、数据中心等数据存储设备。所述可用介质可以是磁性介质(例如,软盘、硬盘、磁带)、光介质(例如,DVD)、或者半导体介质。半导体介质可以是固态硬盘。

[0226] 应理解,本文中术语“和/或”,仅仅是一种描述关联对象的关联关系,表示可以存在三种关系,例如,A和/或B,可以表示:单独存在A,同时存在A和B,单独存在B这三种情况,其中A,B可以是单数或者复数。另外,本文中字符“/”,一般表示前后关联对象是一种“或”的关系,但也可能表示的是一种“和/或”的关系,具体可参考前后文进行理解。

[0227] 本申请中,“至少一个”是指一个或者多个,“多个”是指两个或两个以上。“以下至少一项(个)”或其类似表达,是指的这些项中的任意组合,包括单项(个)或复数项(个)的任意组合。例如,a,b,或c中的至少一项(个),可以表示:a,b,c,a-b,a-c,b-c,或a-b-c,其中a,b,c可以是单个,也可以是多个。

[0228] 应理解,在本申请的各种实施例中,上述各过程的序号的大小并不意味着执行顺序的先后,各过程的执行顺序应以其功能和内在逻辑确定,而不应对本申请实施例的实施过程构成任何限定。

[0229] 本领域普通技术人员可以意识到,结合本文中所公开的实施例描述的各示例的单元及算法步骤,能够以电子硬件、或者计算机软件和电子硬件的结合来实现。这些功能究竟以硬件还是软件方式来执行,取决于技术方案的特定应用和设计约束条件。专业技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能,但是这种实现不应认为超出本申请的范围。

[0230] 所属领域的技术人员可以清楚地了解到,为描述的方便和简洁,上述描述的系统、装置和单元的具体工作过程,可以参考前述方法实施例中的对应过程,在此不再赘述。

[0231] 在本申请所提供的几个实施例中,应该理解到,所揭露的系统、装置和方法,可以

通过其它的方式实现。例如，以上所描述的装置实施例仅仅是示意性的，例如，所述单元的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式，例如多个单元或组件可以结合或者可以集成到另一个系统，或一些特征可以忽略，或不执行。另一点，所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口，装置或单元的间接耦合或通信连接，可以是电性，机械或其它的形式。

[0232] 所述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目的。

[0233] 另外，在本申请各个实施例中的各功能单元可以集成在一个处理单元中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个单元中。

[0234] 所述功能如果以软件功能单元的形式实现并作为独立的产品销售或使用时，可以存储在一个计算机可读取存储介质中。基于这样的理解，本申请的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质中，包括若干指令用以使得一台计算机设备(可以是个人计算机，服务器，或者网络设备等)执行本申请各个实施例所述方法的全部或部分步骤。而前述的存储介质包括:U盘、移动硬盘、只读存储器(read-only memory，ROM)、随机存取存储器(random access memory，RAM)、磁碟或者光盘等各种可以存储程序代码的介质。

[0235] 以上所述，仅为本申请的具体实施方式，但本申请的保护范围并不局限于此，任何熟悉本技术领域的技术人员在本申请揭露的技术范围内，可轻易想到变化或替换，都应涵盖在本申请的保护范围之内。因此，本申请的保护范围应以所述权利要求的保护范围为准。

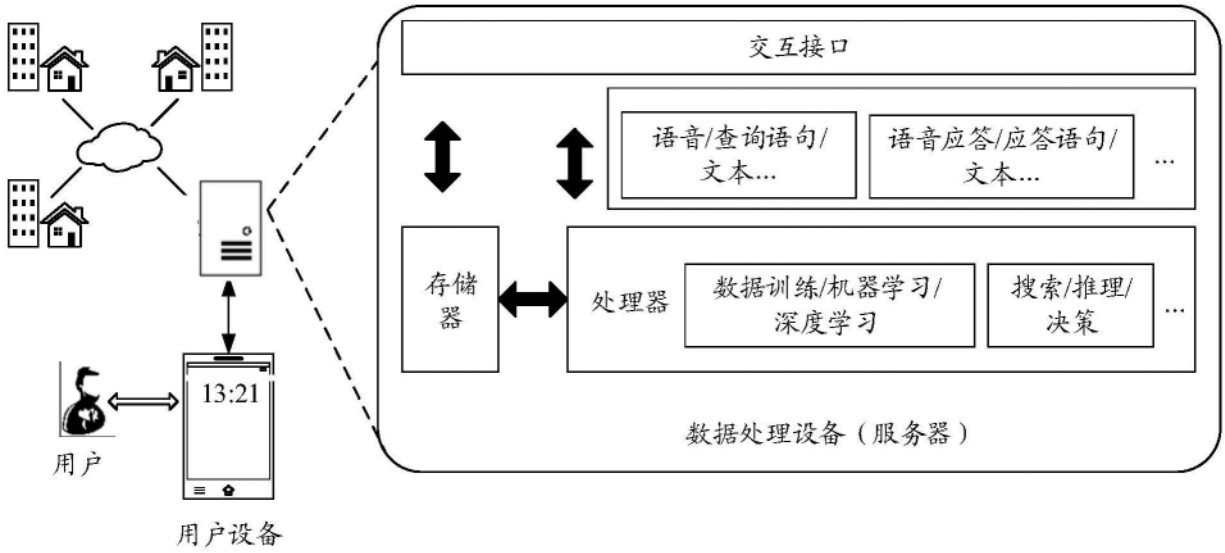

交互接口
语音/查询语句/文本...
语音应答/应答语句/文本...
...
存储器
处理器
数据训练/机器学习/深度学习
搜索/推理/决策
...
数据处理设备（服务器）
13:21
用户
用户设备

图1

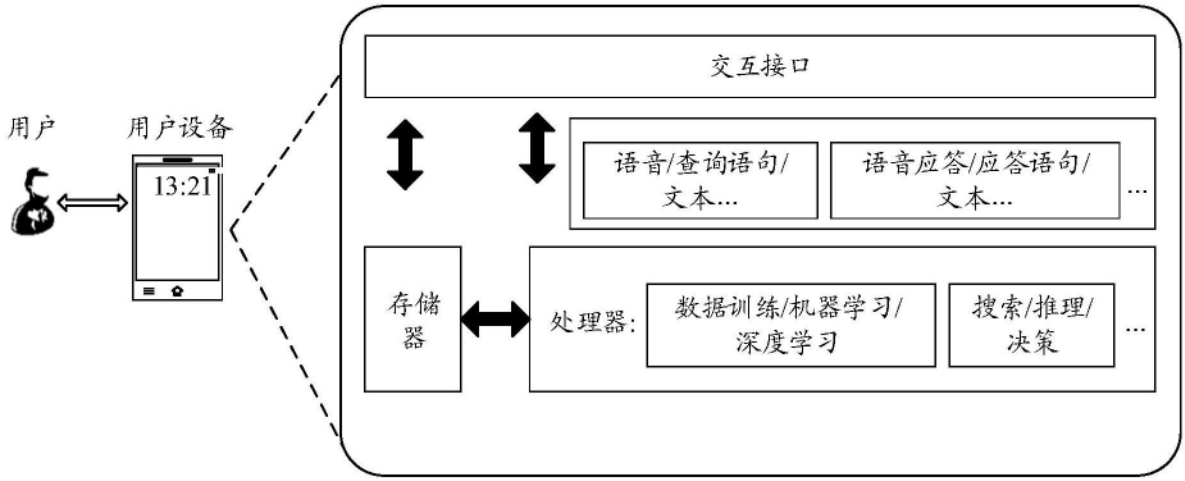

交互接口
用户
用户设备
13:21
语音/查询语句/文本...
语音应答/应答语句/文本...
...
存储器
处理器:
数据训练/机器学习/深度学习
搜索/推理/决策
...

图2

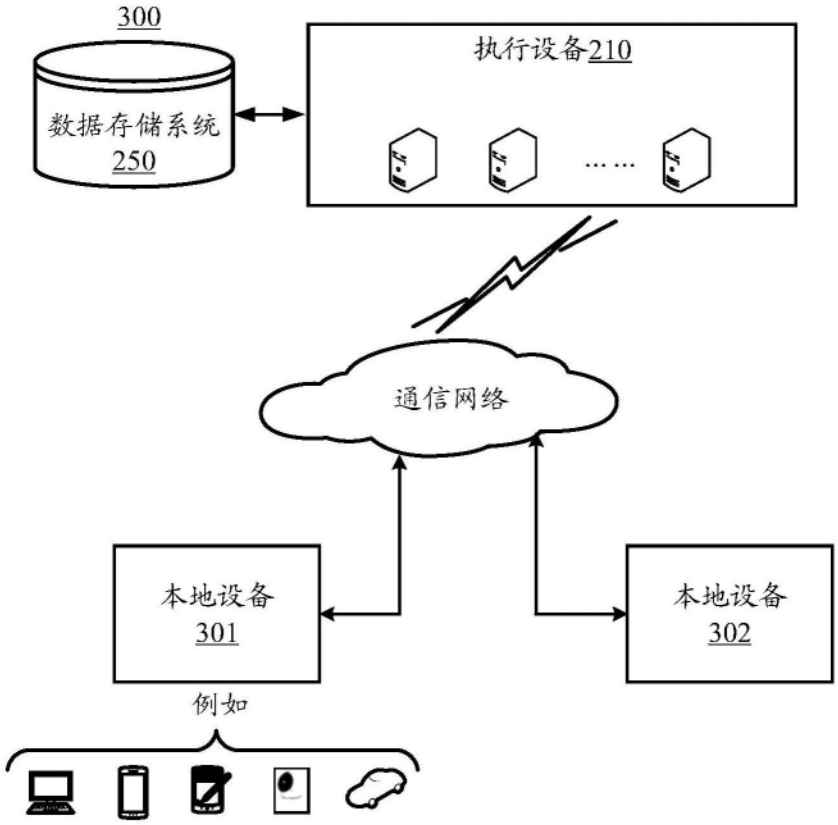


图3

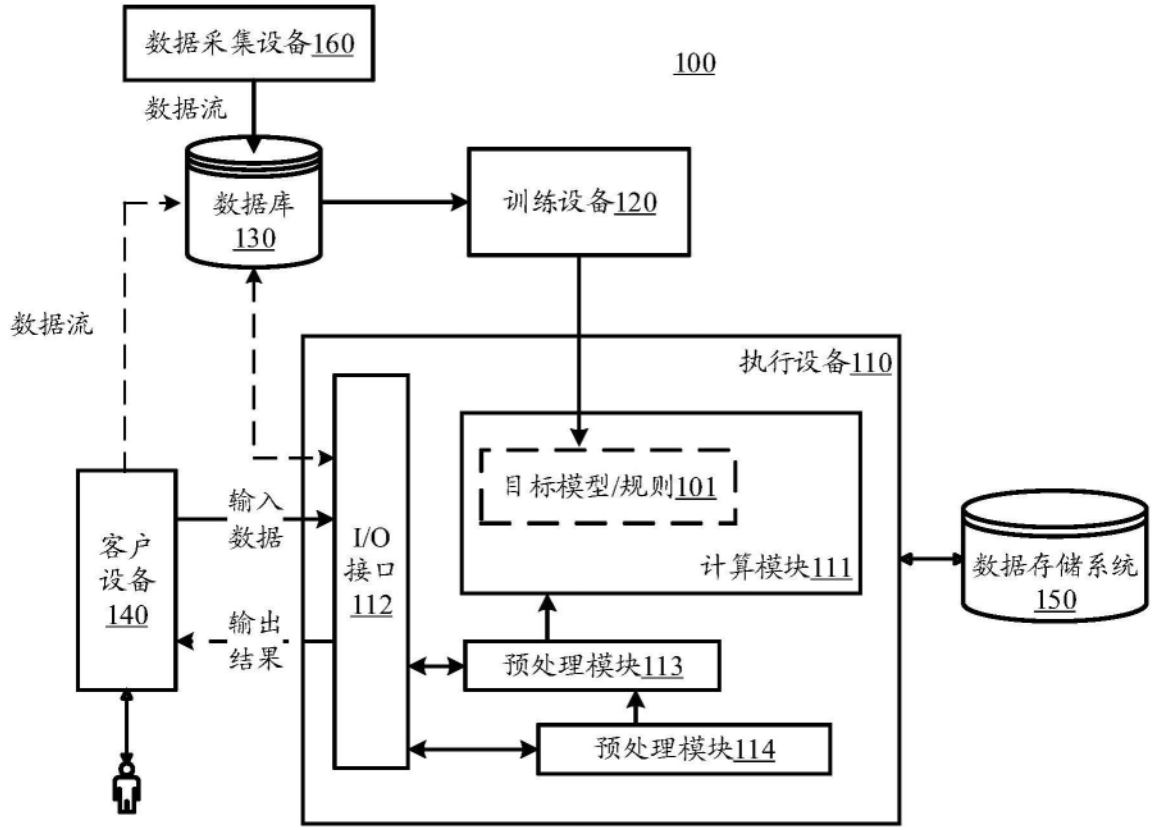

数据采集设备160
100
数据流
数据库
130
训练设备120
数据流
执行设备110
目标模型/规则101
计算模块111
数据存储系统
150
客户
设备
140
输入
数据
输出
结果
I/O
接口
112
预处理模块113
预处理模块114

图4

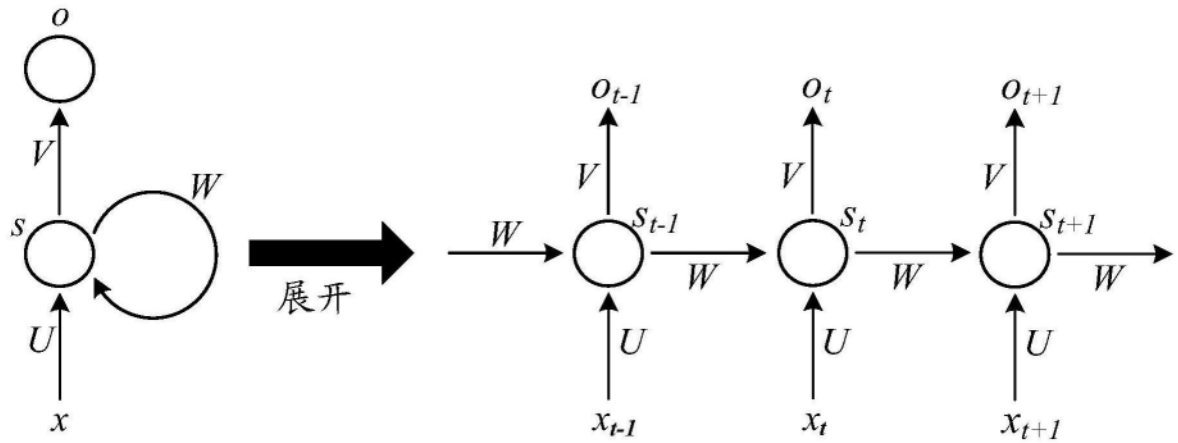

o
V
W
s
U
x
展开
o_{t-1}
o_t
o_{t+1}
s_{t-1}
s_t
s_{t+1}
x_{t-1}
x_t
x_{t+1}

图5

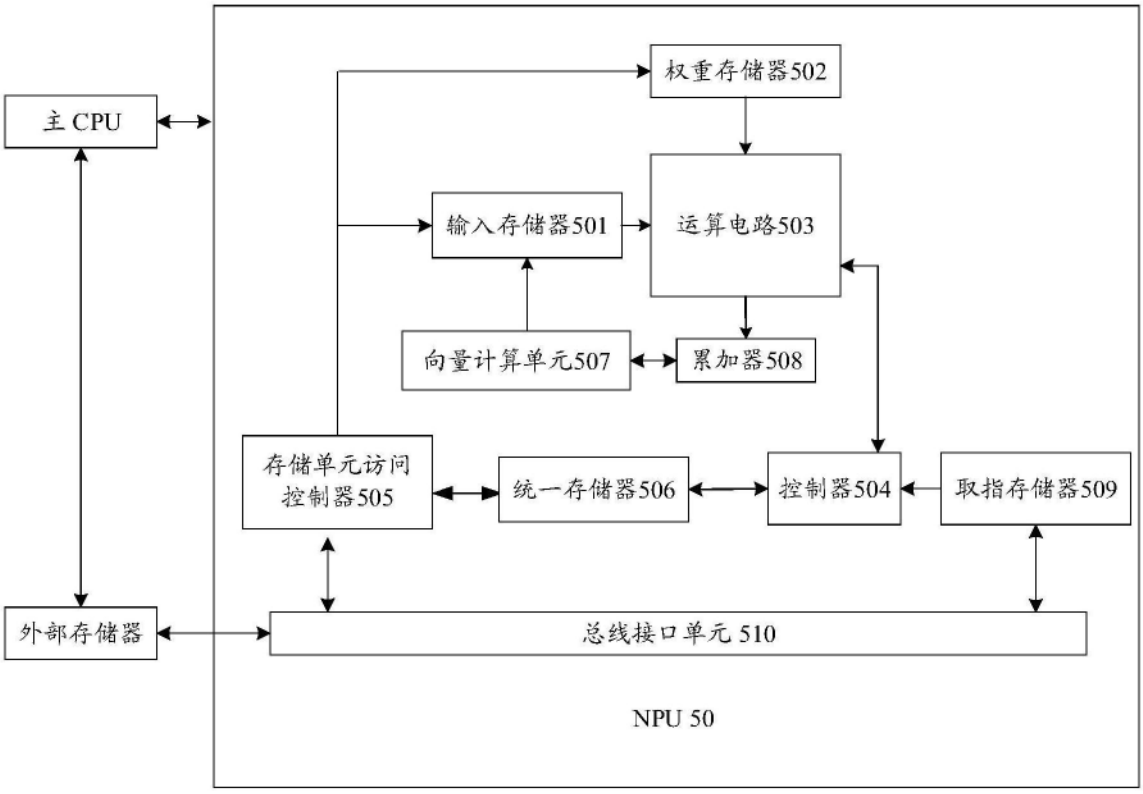

主 CPU
权重存储器502
输入存储器501
运算电路503
向量计算单元507
累加器508
存储单元访问
控制器505
统一存储器506
控制器504
取指存储器509
外部存储器
总线接口单元 510
NPU 50

图6

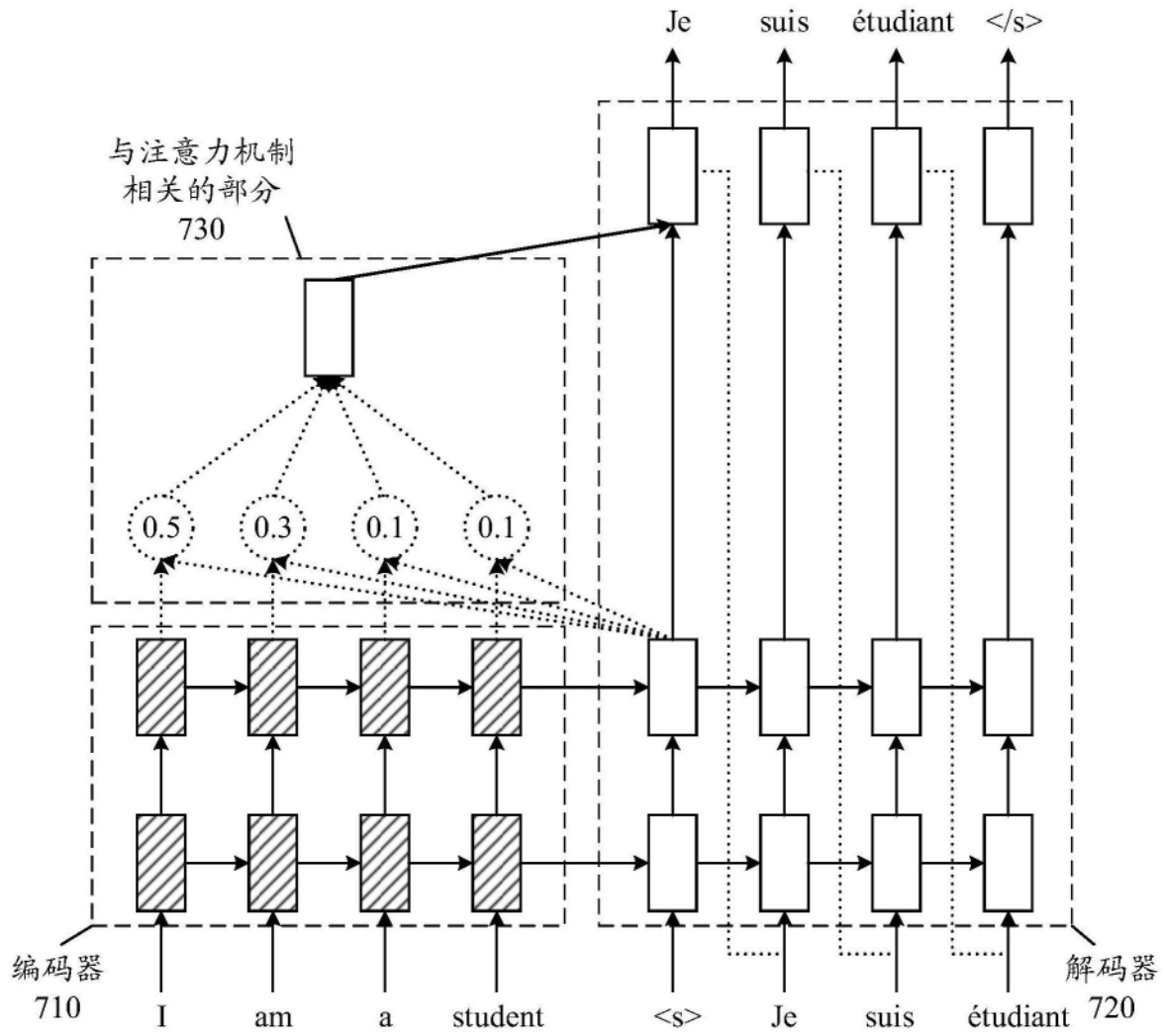

Je
suis
étudiant
</s>
与注意力机制
相关的部分
730
0.5
0.3
0.1
0.1
编码器
710
I
am
a
student
<s>
Je
suis
étudiant
解码器
720

图7

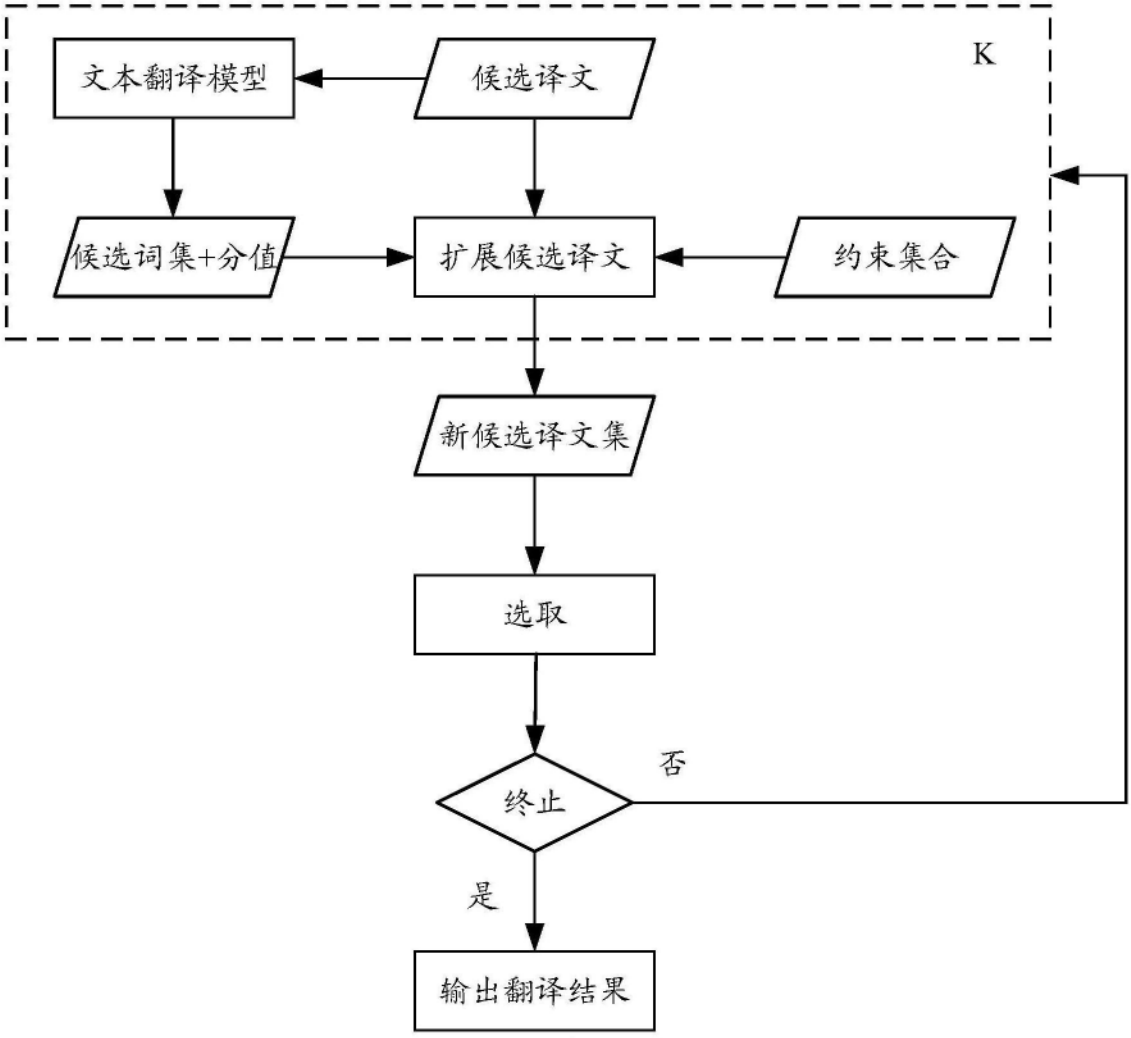


图8

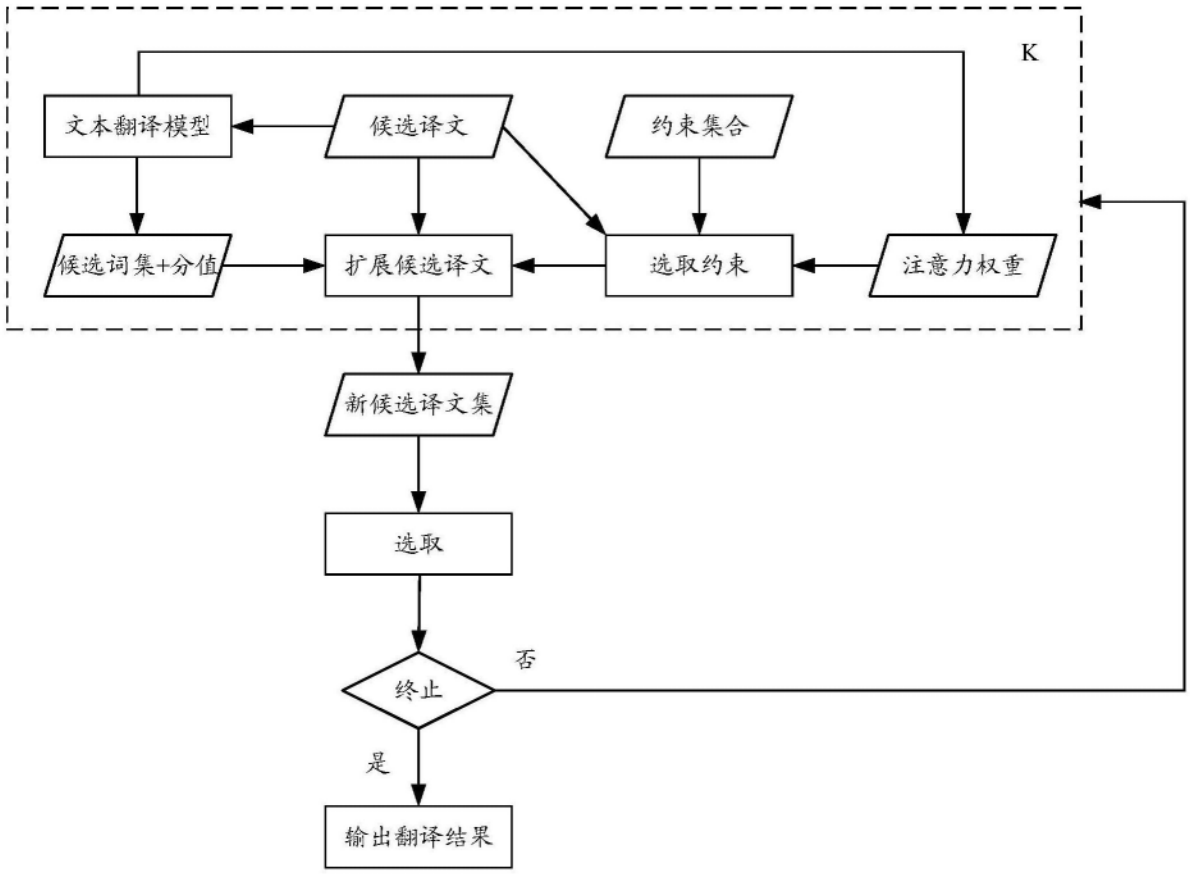

K
文本翻译模型
候选译文
约束集合
候选词集+分值
扩展候选译文
选取约束
注意力权重
新候选译文集
选取
终止
否
是
输出翻译结果

图9

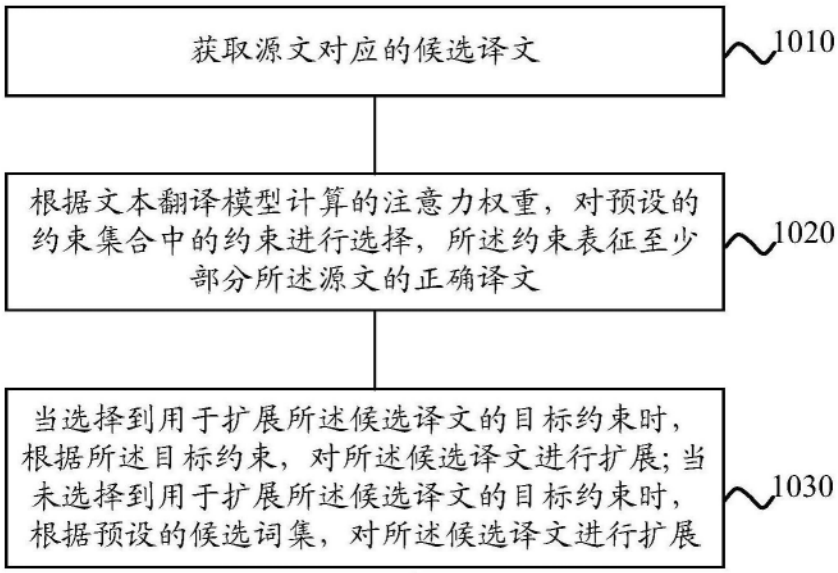

获取源文对应的候选译文
1010
根据文本翻译模型计算的注意力权重，对预设的约束集合中的约束进行选择，所述约束表征至少部分所述源文的正确译文
1020
当选择到用于扩展所述候选译文的目标约束时，根据所述目标约束，对所述候选译文进行扩展；当未选择到用于扩展所述候选译文的目标约束时，根据预设的候选词集，对所述候选译文进行扩展
1030

图10

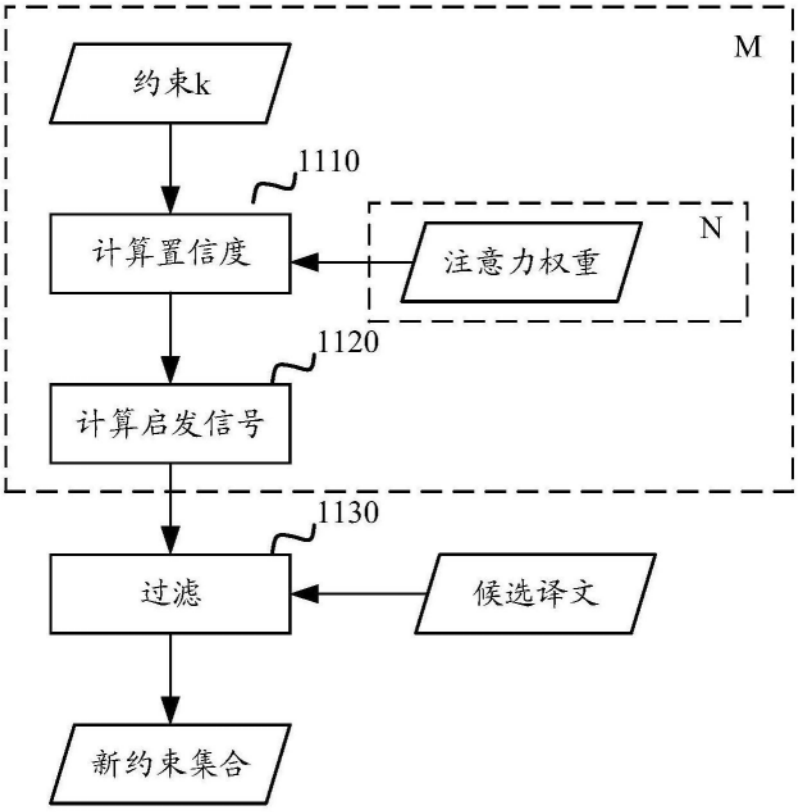

M
约束k
1110
计算置信度
N
注意力权重
1120
计算启发信号
1130
过滤
候选译文
新约束集合

图11

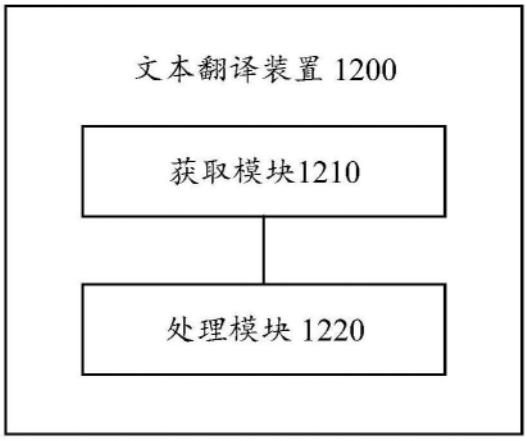

文本翻译装置 1200
获取模块1210
处理模块 1220

图12

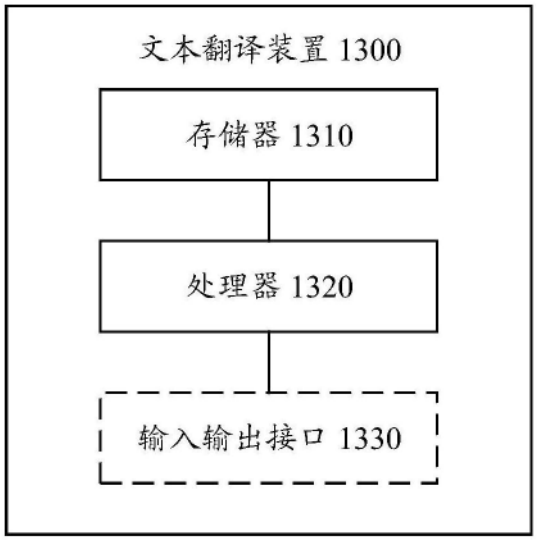

文本翻译装置 1300
存储器 1310
处理器 1320
输入输出接口 1330

图13

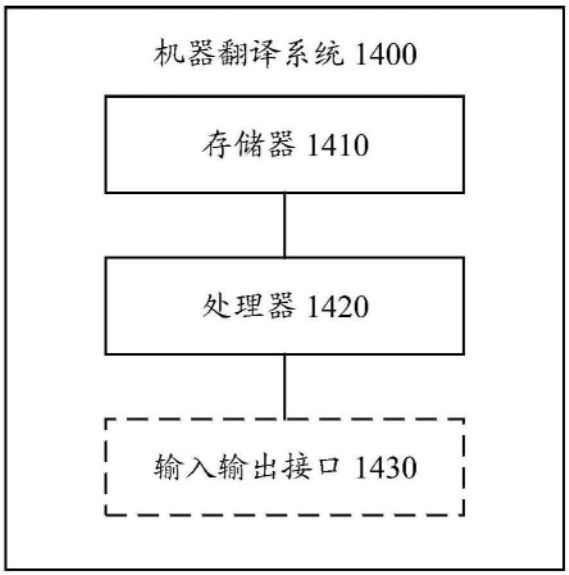

机器翻译系统 1400
存储器 1410
处理器 1420
输入输出接口 1430

图14