



(12) 发明专利

(10) 授权公告号 CN 109948166 B

(45) 授权公告日 2021. 03. 02

(21) 申请号 201910228498.X

审查员 何祥鹏

(22) 申请日 2019.03.25

(65) 同一申请的已公布的文献号

申请公布号 CN 109948166 A

(43) 申请公布日 2019.06.28

(73) 专利权人 腾讯科技(深圳)有限公司

地址 518000 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 王龙跃 涂兆鹏 王星 史树明

(74) 专利代理机构 广州华进联合专利商标代理

有限公司 44224

代理人 李文渊 何平

(51) Int. Cl.

G06F 40/58 (2020.01)

G06F 40/30 (2020.01)

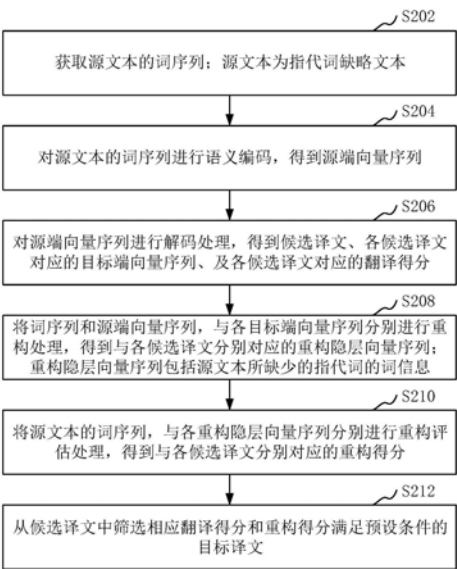
权利要求书7页 说明书34页 附图10页

(54) 发明名称

文本翻译方法、装置、存储介质和计算机设备

(57) 摘要

本申请涉及一种文本翻译方法、装置、存储介质和计算机设备,该方法包括:获取源文本的词序列;源文本为指代词省略文本;对词序列进行语义编码,得到源端向量序列;对源端向量序列进行解码,得到候选译文、各候选译文对应的目标端向量序列、及各候选译文对应的翻译得分;将词序列和源端向量序列,与各目标端向量序列分别进行重构处理,得到与各候选译文分别对应的重构隐层向量序列;重构隐层向量序列包括源文本所缺少指代词的词信息;将源文本的词序列,与各重构隐层向量序列分别进行重构评估处理,得到与各候选译文分别对应的重构得分;从候选译文中筛选相应翻译得分和重构得分满足预设条件的目标译文。本申请提供的方案可以提高翻译质量。



1. 一种文本翻译方法,包括:

获取源文本的词序列;所述源文本为指代词省略文本;

通过机器翻译模型中的编码器,对所述源文本的词序列进行语义编码,得到源端向量序列;

通过所述机器翻译模型中的解码器,对所述源端向量序列进行解码处理,得到候选译文、各所述候选译文对应的目标端向量序列、及各所述候选译文对应的翻译得分;

通过所述机器翻译模型中的重构器,将所述词序列和所述源端向量序列,与各所述目标端向量序列分别进行重构处理,得到与各所述候选译文分别对应的重构隐层向量序列;所述重构处理是将源端向量序列和目标端向量序列重构回源文本的处理;所述重构隐层向量序列包括所述源文本所缺少的指代词的词信息,所述指代词的词信息包括指代词的位置信息和指代词的词本身;

通过所述机器翻译模型中的重构器,将所述源文本的词序列,与各所述重构隐层向量序列分别进行重构评估处理,得到与各所述候选译文分别对应的重构得分;

从所述候选译文中筛选相应翻译得分和重构得分满足预设条件的目标译文;

其中,所述机器翻译模型通过以下步骤训练得到:

获取样本源文本、参考目标译文、以及与所述源文本对应的标注词序列;所述标注词序列是对所述样本源文本进行省略词标注处理所得到的词序列;

将所述样本源文本输入机器翻译模型中进行训练,得到预测目标译文;

依据所述参考目标译文和预测目标译文构建最大似然函数;

通过所述机器翻译模型,获取与所述样本源文本对应的源端样本向量序列、以及目标端样本向量序列;

通过所述机器翻译模型中的重构器,依据所述样本源文本,对所述源端样本向量序列、以及目标端样本向量序列进行重构处理,得到样本重构隐层向量序列;

依据所述样本源文本、所述样本重构隐层向量序列、所述源端样本向量序列、以及所述目标端样本向量序列构建重构评估函数,所述重构评估函数用于衡量省略词的词信息召回程度,辅助机器翻译模型的重构器所输出的重构隐层向量序列学习得到关于省略词的词信息;

依据所述标注词序列和所述样本重构隐层向量序列,构建词序列标注损失函数,所述标注词序列用于在模型训练时引导所述重构器在生成对应的重构隐层向量序列学习到指代词的词信息;

根据所述最大似然函数、所述重构评估函数和所述词序列标注损失函数,确定目标函数;

将所述目标函数最大化时的模型参数作为机器翻译模型的模型参数,返回所述将所述样本源文本输入机器翻译模型中进行训练,得到预测目标译文的步骤并继续训练,直至满足训练停止条件时停止训练。

2. 根据权利要求1所述的方法,其特征在于,所述方法还包括:

获取与所述源文本对应的篇章文本;

对所述篇章文本进行编码处理,得到篇章向量;

将所述篇章向量,与各所述重构隐层向量序列分别进行融合,得到与各所述候选译文

分别对应的第一融合向量序列；

所述将所述源文本的词序列，与各所述重构隐层向量序列分别进行重构评估处理，得到与各所述候选译文分别对应的重构得分，包括：

将所述源文本的词序列，与各所述第一融合向量序列分别进行重构评估处理，得到与各所述候选译文分别对应的重构得分。

3. 根据权利要求1所述的方法，其特征在于，所述方法还包括：

获取与所述源文本对应的篇章文本；

对所述篇章文本进行编码处理，得到篇章向量；

将所述篇章向量，与各所述目标端向量序列分别进行融合，得到与各所述候选译文分别对应的第二融合向量序列；

所述将所述词序列和所述源端向量序列，与各所述目标端向量序列分别进行重构处理，得到与各所述候选译文分别对应的重构隐层向量序列，包括：

将所述词序列和所述源端向量序列，与各所述第二融合向量序列分别进行重构处理，得到与各所述候选译文分别对应的重构隐层向量序列。

4. 根据权利要求2所述的方法，其特征在于，所述对所述篇章文本进行编码处理，得到篇章向量，包括：

确定所述篇章文本所包括的整句；

对各整句分别进行分词处理，得到各整句各自对应的词序列；

对各整句各自对应的词序列分别进行词级编码处理，得到与各整句分别对应的句子向量序列；

将与各整句分别对应的句子向量序列进行句子级编码处理，得到与所述篇章文本对应的篇章向量。

5. 根据权利要求1所述的方法，其特征在于，所述对所述源端向量序列进行解码处理，得到候选译文、各所述候选译文对应的目标端向量序列、及各所述候选译文对应的翻译得分，包括：

依次对所述源端向量序列进行解码得到目标端向量，且每次解码时根据前次确定的候选词的词向量进行解码，并根据当次的目标端向量确定当次的多于一个的候选词；

将依次解码得到的各候选词，按所述候选词的生成时序，拼接得到多于一组的候选译文；

通过构成各候选译文的候选词所对应的目标端向量，确定与各候选译文分别对应的目标端向量序列；

根据所述源端向量序列、历史解码得到的候选词、及历史解码得到的目标端向量序列，确定与各候选词对应的翻译得分；

依据构成各所述候选译文的各候选词所对应的翻译得分，计算得到与各所述候选译文对应的翻译得分。

6. 根据权利要求1所述的方法，其特征在于，所述方法还包括：

获取与所述源端向量序列对应的源端注意力分配权重向量；

根据所述源端注意力分配权重向量、以及所述源端向量序列，计算得到源端内容向量；

分别获取与各目标端向量序列对应的目标端注意力分配权重向量；

根据所述目标端注意力分配权重向量、以及相对应的目标端向量序列,计算得到与各候选译文分别对应的目标端内容向量;

所述将所述词序列和所述源端向量序列,与各所述目标端向量序列分别进行重构处理,得到与各所述候选译文分别对应的重构隐层向量序列,包括:

将所述词序列和所述源端内容向量,与各所述目标端内容向量分别进行重构处理,得到与各所述候选译文分别对应的重构隐层向量序列。

7. 根据权利要求6所述的方法,其特征在于,所述将所述词序列和所述源端内容向量,与各所述目标端内容向量分别进行重构处理,得到与各所述候选译文分别对应的重构隐层向量序列,包括:

对于每组候选译文,分别根据所述源文本中当前词的前一词的词向量、前次重构得到的重构隐层向量、所述源端内容向量和各候选译文各自所对应的目标端内容向量,重构得到当次重构隐层向量;

对于每组候选译文,分别将依次重构得到的重构隐层向量构成重构隐层向量序列;重构得到的重构隐层向量序列包括所述源文本所缺少的指代词的词信息。

8. 根据权利要求6所述的方法,其特征在于,所述将所述源文本的词序列,与各所述重构隐层向量序列分别进行重构评估处理,得到与各所述候选译文分别对应的重构得分,包括:

对于每组候选译文,分别根据所述源文本中当前词的前一词的词向量、所述源端内容向量、以及各候选译文对应的当次的重构隐层向量和目标端内容向量,计算得到与各候选译文分别对应、且与所述源文本中当前词对应的重构得分;

对于每组候选译文,分别对与所述源文本中各词所对应的重构得分进行连乘运算,得到与各候选译文分别对应的重构得分。

9. 根据权利要求1所述的方法,其特征在于,所述从所述候选译文中筛选相应翻译得分和重构得分满足预设条件的目标译文,包括:

对各所述候选译文分别对应的翻译得分和重构得分进行加权求和计算,得到各所述候选译文各自所对应的综合得分;

从所述候选译文中,筛选相应综合得分为最高得分的候选译文;

将筛选出的候选译文作为目标译文并输出。

10. 根据权利要求1所述的方法,其特征在于,所述根据所述最大似然函数、所述重构评估函数和所述词序列标注损失函数,确定目标函数包括:

将所述最大似然函数、所述重构评估函数和所述词序列标注损失函数的加权和函数作为所述机器翻译模型的目标函数;其中,通过以下公式表示所述目标函数:

$$J(\theta, \gamma, \psi) = \arg \max_{\theta, \gamma, \psi} \sum_{n=1}^N \{ \log L(y^n | x^n; \theta) + \lambda \log R(x | h^{enc}, h^{dec}; \theta, \gamma) + \tau \log P(dp | h^{rec}; \theta, \gamma, \psi) \}$$

$$\text{其中, } R(x | h^{enc}, h^{dec}) = \prod_{i=1}^I g_r(x_{i-1}, h_i^{rec}, c_i^{enc}, c_i^{dec}); \quad h_i^{rec} = f_r(x_{i-1}, h_{i-1}^{rec}, c_i^{enc}, c_i^{dec});$$

$$c_i = \sum_{j=1}^J \alpha_{i,j} \bullet h_j; \quad P(dp|h^{rec}) = \prod_{i=1}^I P(dp_i|h_i^{rec}) = \prod_{i=1}^I g_i(dp_i|h_i^{rec});$$

其中, θ 、 γ 和 Φ 分别是模型参数; λ 和 τ 为对应的加权系数; $\{[x^n, y^n]\}_{n=1}^N$ 是模型训练中的样本平行语料对; $g_r()$ 和 $f_r()$ 分别为归一化函数和激活函数; $g_l()$ 是归一化函数; h^{enc} 表示源端样本向量序列; h^{dec} 表示目标端样本向量序列; x_{i-1} 表示所述样本源文本中的第 $i-1$ 个词的词向量; h_i^{rec} 表示与所述样本源文本中的第 i 个词对应的样本重构隐层向量; h_j 表示各源端样本向量或目标端样本向量; $\alpha_{i,j}$ 表示源端注意力分配权重或目标端注意力分配权重; $\bullet$ 表示向量点乘运算; c_i^{enc} 表示与源文本中的第 i 个词对应的源端内容向量; c_i^{dec} 表示与源文本中的第 i 个词对应的目标端内容向量; $dp = \{dp_1, dp_2, \dots, dp_I\}$ 是样本标注词序列。

11. 根据权利要求1所述的方法, 其特征在于, 所述方法还包括:

获取与所述样本源文本对应的样本篇章文本;

对所述样本篇章文本进行编码处理, 得到样本篇章向量;

融合所述样本篇章向量和所述样本重构隐层向量序列, 得到样本融合向量序列;

所述依据所述样本源文本、所述样本重构隐层向量序列、所述源端样本向量序列、以及所述目标端样本向量序列构建重构评估函数, 包括:

依据所述样本源文本、所述样本融合向量序列、所述源端样本向量序列、以及所述目标端样本向量序列构建重构评估函数;

所述依据所述标注词序列和所述样本重构隐层向量序列, 构建词序列标注损失函数, 包括:

依据所述标注词序列和所述样本融合向量序列, 构建词序列标注损失函数。

12. 一种文本翻译装置, 所述装置包括:

获取模块, 用于获取源文本的词序列; 所述源文本为指代词省略文本;

编码模块, 用于通过机器翻译模型中的编码器, 对所述源文本的词序列进行语义编码, 得到源端向量序列;

解码模块, 用于通过所述机器翻译模型中的解码器, 对所述源端向量序列进行解码处理, 得到候选译文、各所述候选译文对应的目标端向量序列、及各所述候选译文对应的翻译得分;

重构模块, 用于通过所述机器翻译模型中的重构器, 将所述词序列和所述源端向量序列, 与各所述目标端向量序列分别进行重构处理, 得到与各所述候选译文分别对应的重构隐层向量序列; 所述重构处理是将源端向量序列和目标端向量序列重构回源文本的处理; 所述重构隐层向量序列包括所述源文本所缺少的指代词的词信息, 所述指代词的词信息包括指代词的位置信息和指代词的词本身;

评估模块, 用于通过所述机器翻译模型中的重构器, 将所述源文本的词序列, 与各所述重构隐层向量序列分别进行重构评估处理, 得到与各所述候选译文分别对应的重构得分;

筛选模块, 用于从所述候选译文中筛选相应翻译得分和重构得分满足预设条件的目标译文;

训练模块, 用于获取样本源文本、参考目标译文、以及与所述源文本对应的标注词序

列;所述标注词序列是对所述样本源文本进行缺略词标注处理所得到的词序列;将所述样本源文本输入机器翻译模型中进行训练,得到预测目标译文;依据所述参考目标译文和预测目标译文构建最大似然函数;通过所述机器翻译模型,获取与所述样本源文本对应的源端样本向量序列、以及目标端样本向量序列;通过所述机器翻译模型中的重构器,依据所述样本源文本,对所述源端样本向量序列、以及目标端样本向量序列进行重构处理,得到样本重构隐层向量序列;依据所述样本源文本、所述样本重构隐层向量序列、所述源端样本向量序列、以及所述目标端样本向量序列构建重构评估函数,所述重构评估函数用于衡量缺略词的词信息召回程度,辅助机器翻译模型的重构器所输出的重构隐层向量序列学习得到关于缺略词的词信息;依据所述标注词序列和所述样本重构隐层向量序列,构建词序列标注损失函数,所述标注词序列用于在模型训练时引导所述重构器在生成对应的重构隐层向量序列学习到指代词的词信息;根据所述最大似然函数、所述重构评估函数和所述词序列标注损失函数,确定目标函数;将所述目标函数最大化时的模型参数作为机器翻译模型的模型参数,返回所述将所述样本源文本输入机器翻译模型中进行训练,得到预测目标译文的步骤并继续训练,直至满足训练停止条件时停止训练。

13. 根据权利要求12所述的装置,其特征在于,所述装置还包括融合模块,其中:

所述获取模块,用于获取与所述源文本对应的篇章文本;

所述编码模块,还用于对所述篇章文本进行编码处理,得到篇章向量;

所述融合模块,用于将所述篇章向量,与各所述重构隐层向量序列分别进行融合,得到与各所述候选译文分别对应的第一融合向量序列;

所述评估模块,还用于将所述源文本的词序列,与各所述第一融合向量序列分别进行重构评估处理,得到与各所述候选译文分别对应的重构得分。

14. 根据权利要求12所述的装置,其特征在于,所述装置还包括融合模块,其中:

所述获取模块还用于获取与所述源文本对应的篇章文本;

所述编码模块,还用于对所述篇章文本进行编码处理,得到篇章向量;

所述融合模块,还用于将所述篇章向量,与各所述目标端向量序列分别进行融合,得到与各所述候选译文分别对应的第二融合向量序列;

所述重构处理,还用于将所述词序列和所述源端向量序列,与各所述第二融合向量序列分别进行重构处理,得到与各所述候选译文分别对应的重构隐层向量序列。

15. 根据权利要求13所述的装置,其特征在于,所述编码模块还用于确定所述篇章文本所包括的整句;对各整句分别进行分词处理,得到各整句各自对应的词序列;对各整句各自对应的词序列分别进行词级编码处理,得到与各整句分别对应的句子向量序列;将与各整句分别对应的句子向量序列进行句子级编码处理,得到与所述篇章文本对应的篇章向量。

16. 根据权利要求12所述的装置,其特征在于,所述解码模块,还用于依次对所述源端向量序列进行解码得到目标端向量,且每次解码时根据前次确定的候选词的词向量进行解码,并根据当次的目标端向量确定当次的多于一个的候选词;将依次解码得到的各候选词,按所述候选词的生成时序,拼接得到多于一组的候选译文;通过构成各候选译文的候选词所对应的目标端向量,确定与各候选译文分别对应的目标端向量序列;根据所述源端向量序列、历史解码得到的候选词、及历史解码得到的目标端向量序列,确定与各候选词对应的翻译得分;依据构成各所述候选译文的各候选词所对应的翻译得分,计算得到与各所述候

选译文对应的翻译得分。

17. 根据权利要求12所述的装置,其特征在于,所述装置还包括计算模块,用于获取与所述源端向量序列对应的源端注意力分配权重向量;根据所述源端注意力分配权重向量、以及所述源端向量序列,计算得到源端内容向量;分别获取与各目标端向量序列对应的目标端注意力分配权重向量;根据所述目标端注意力分配权重向量、以及相对应的目标端向量序列,计算得到与各候选译文分别对应的目标端内容向量;

所述重构模块,还用于将所述词序列和所述源端内容向量,与各所述目标端内容向量分别进行重构处理,得到与各所述候选译文分别对应的重构隐层向量序列。

18. 根据权利要求17所述的装置,其特征在于,所述重构模块,还用于对于每组候选译文,分别根据所述源文本中当前词的前一词的词向量、前次重构得到的重构隐层向量、所述源端内容向量和各候选译文各自所对应的目标端内容向量,重构得到当次重构隐层向量;对于每组候选译文,分别将依次重构得到的重构隐层向量构成重构隐层向量序列;重构得到的重构隐层向量序列包括所述源文本所缺少的指代词的信息。

19. 根据权利要求17所述的装置,其特征在于,所述评估模块,还用于对于每组候选译文,分别根据所述源文本中当前词的前一词的词向量、所述源端内容向量、以及各候选译文对应的当次的重构隐层向量和目标端内容向量,计算得到与各候选译文分别对应、且与所述源文本中当前词对应的重构得分;对于每组候选译文,分别对与所述源文本中各词所对应的重构得分进行连乘运算,得到与各候选译文分别对应的重构得分。

20. 根据权利要求12所述的装置,其特征在于,所述筛选模块,还用于对各所述候选译文分别对应的翻译得分和重构得分进行加权求和计算,得到各所述候选译文各自所对应的综合得分;从所述候选译文中,筛选相应综合得分为最高得分的候选译文;将筛选出的候选译文作为目标译文并输出。

21. 根据权利要求12所述的装置,其特征在于,所述训练模块,还用于将所述最大似然函数、所述重构评估函数和所述词序列标注损失函数的加权和函数作为所述机器翻译模型的目标函数;其中,通过以下公式表示所述目标函数:

$$J(\theta, \gamma, \psi) = \arg \max_{\theta, \gamma, \psi} \sum_{n=1}^N \{ \log L(y^n | x^n; \theta) + \lambda \log R(x | h^{enc}, h^{dec}; \theta, \gamma) + \tau \log P(dp | h^{rec}; \theta, \gamma, \psi) \}$$

$$\text{其中, } R(x | h^{enc}, h^{dec}) = \prod_{i=1}^I g_r(x_{i-1}, h_i^{rec}, c_i^{enc}, c_i^{dec}); \quad h_i^{rec} = f_r(x_{i-1}, h_{i-1}^{rec}, c_i^{enc}, c_i^{dec});$$

$$c_i = \sum_{j=1}^J \alpha_{i,j} \cdot h_j; \quad P(dp | h^{rec}) = \prod_{i=1}^I P(dp_i | h_i^{rec}) = \prod_{i=1}^I g_l(dp_i | h_i^{rec});$$

其中, θ 、 γ 和 ψ 分别是模型参数; λ 和 τ 为对应的加权系数; $\{[x^n, y^n]\}_{n=1}^N$ 是模型训练中的样本平行语料对; $g_r()$ 和 $f_r()$ 分别为归一化函数和激活函数; $g_l()$ 是归一化函数; h^{enc} 表示源端样本向量序列; h^{dec} 表示目标端样本向量序列; x_{i-1} 表示所述样本源文本中的第 $i-1$ 个词的词向量; h_i^{rec} 表示与所述样本源文本中的第 i 个词对应的样本重构隐层向量; h_j 表示各源端样本向量或目标端样本向量; $\alpha_{i,j}$ 表示源端注意力分配权重或目标端注意力分配权

重; $\bullet$ 表示向量点乘运算; c_i^{enc} 表示与源文本中的第 i 个词对应的源端内容向量; c_i^{dec} 表示与源文本中的第 i 个词对应的目标端内容向量; $dp = \{dp_1, dp_2, \dots, dp_T\}$ 是样本标注词序列。

22. 根据权利要求12所述的装置, 其特征在于, 所述训练模块, 还用于获取与所述样本源文本对应的样本篇章文本; 对所述样本篇章文本进行编码处理, 得到样本篇章向量; 融合所述样本篇章向量和所述样本重构隐层向量序列, 得到样本融合向量序列; 依据所述样本源文本、所述样本融合向量序列、所述源端样本向量序列、以及所述目标端样本向量序列构建重构评估函数; 依据所述标注词序列和所述样本融合向量序列, 构建词序列标注损失函数。

23. 一种计算机设备, 包括存储器和处理器, 所述存储器存储有计算机程序, 其特征在于, 所述处理器执行所述计算机程序时实现权利要求1至11中任一项所述的方法的步骤。

24. 一种计算机可读存储介质, 存储有计算机程序, 其特征在于, 所述计算机程序被处理器执行时实现权利要求1至11中任一项所述的方法的步骤。

文本翻译方法、装置、存储介质和计算机设备

技术领域

[0001] 本申请涉及机器学习技术领域,特别是涉及一种文本翻译方法、装置、存储介质和计算机设备,以及模型训练方法、装置、存储介质和计算机设备。

背景技术

[0002] 机器翻译是通过计算机将一种自然语言文本(需翻译的自然语言文本一般称为源文本)翻译为另一种自然语言文本(翻译后的自然语言文本称为目标译文),以实现不同语种的文本之间的转换。而不同的语种之间会有不同的语法及句式习惯,比如,在汉语和日语等语言中,指代词省略是极为常见的现象,与之相反,在英、法等语言中,指代词作为句子成分则必不可少。当从指代词省略语言(比如汉语)到非指代词省略语言(比如英语)翻译时,需要将隐含的省略词(也就是缺少的指代词)召回并正确地翻译出来,这给机器翻译带来了极大挑战,机器翻译很难将源端的隐性信息在目标端显性地解码出来。

[0003] 传统的机器翻译方式,以中译英为例,通常采用词对齐算法和语言模型,自动标注并还原中文句子中的省略代词,将补全了省略代词的句子传递给统计机器翻译系统进行翻译。然而传统的机器翻译方式,常常由于省略词还原的低准确率带来错误传递,导致源端被添加了错误的指代词从而使得翻译结果错误,导致了文本翻译准确率低的问题。

发明内容

[0004] 基于此,有必要针对文本翻译准确率低的技术问题,提供一种文本翻译方法、装置、计算机可读存储介质和计算机设备,以及模型训练方法、装置、计算机可读存储介质和计算机设备。

[0005] 一种文本翻译方法,包括:

[0006] 获取源文本的词序列;所述源文本为指代词省略文本;

[0007] 对所述源文本的词序列进行语义编码,得到源端向量序列;

[0008] 对所述源端向量序列进行解码处理,得到候选译文、各所述候选译文对应的目标端向量序列、及各所述候选译文对应的翻译得分;

[0009] 将所述词序列和所述源端向量序列,与各所述目标端向量序列分别进行重构处理,得到与各所述候选译文分别对应的重构隐层向量序列;所述重构隐层向量序列包括所述源文本所缺少的指代词的词信息;

[0010] 将所述源文本的词序列,与各所述重构隐层向量序列分别进行重构评估处理,得到与各所述候选译文分别对应的重构得分;

[0011] 从所述候选译文中筛选相应翻译得分和重构得分满足预设条件的目标译文。

[0012] 一种文本翻译装置,所述装置包括:

[0013] 获取模块,用于获取源文本的词序列;所述源文本为指代词省略文本;

[0014] 编码模块,用于对所述源文本的词序列进行语义编码,得到源端向量序列;

[0015] 解码模块,用于对所述源端向量序列进行解码处理,得到候选译文、各所述候选译

文对应的目标端向量序列、及各所述候选译文对应的翻译得分；

[0016] 重构模块,用于将所述词序列和所述源端向量序列,与各所述目标端向量序列分别进行重构处理,得到与各所述候选译文分别对应的重构隐层向量序列;所述重构隐层向量序列包括所述源文本所缺少的指代词的词信息;

[0017] 评估模块,用于将所述源文本的词序列,与各所述重构隐层向量序列分别进行重构评估处理,得到与各所述候选译文分别对应的重构得分;

[0018] 筛选模块,用于从所述候选译文中筛选相应翻译得分和重构得分满足预设条件的目标译文。

[0019] 一种计算机可读存储介质,存储有计算机程序,所述计算机程序被处理器执行时,使得所述处理器执行所述文本翻译方法的步骤。

[0020] 一种计算机设备,包括存储器和处理器,所述存储器存储有计算机程序,所述计算机程序被所述处理器执行时,使得所述处理器执行所述文本翻译方法的步骤。

[0021] 上述文本翻译方法、装置、计算机可读存储介质和计算机设备,对缺少指代词的源文本进行语义编码,得到源端向量序列,并继续对源端向量序列进行解码,得到候选译文、各候选译文对应的目标端向量序列、及各候选译文对应的翻译得分,这样可充分提取出源端和目标端的隐层信息。再通过融合源端和目标端的隐层信息可以共享两端的共性信息,在根据源文本进行重构处理时可以减少冗余信息并能避免重要信息的缺失,不需要额外的省略词标注信息就可重构回包括缺少的指代词的词信息,这样使得确定的重构得分可以很好的衡量缺少的指代词信息的召回程度。因此,依据重构得分和翻译得分所筛选出的目标译文,可以减少翻译时缺少省略代词所带来的影响,大大提高了翻译质量。

[0022] 一种模型训练方法,包括:

[0023] 获取样本源文本、参考目标译文、以及与所述源文本对应的标注词序列;所述标注词序列是对所述样本源文本进行省略词标注处理所得到的词序列;

[0024] 将所述样本源文本输入机器翻译模型中进行训练,得到预测目标译文;

[0025] 依据所述参考目标译文和预测目标译文构建最大似然函数;

[0026] 通过所述机器翻译模型,获取与所述样本源文本对应的源端样本向量序列、以及目标端样本向量序列;

[0027] 依据所述样本源文本,对所述源端样本向量序列、以及目标端样本向量序列进行重构处理,得到样本重构隐层向量序列;

[0028] 依据所述样本源文本、所述样本重构隐层向量序列、所述源端样本向量序列、以及所述目标端样本向量序列构建重构评估函数;

[0029] 依据所述标注词序列和所述样本重构隐层向量序列,构建词序列标注损失函数;

[0030] 根据所述最大似然函数、所述重构评估函数和所述词序列标注损失函数,确定目标函数;

[0031] 将所述目标函数最大化时的模型参数作为机器翻译模型的模型参数,返回所述将所述样本源文本输入机器翻译模型中进行训练,得到预测目标译文的步骤并继续训练,直至满足训练停止条件时停止训练。

[0032] 在其中一个实施例中,所述模型训练方法还包括:

- [0033] 获取与所述样本源文本对应的样本篇章文本；
- [0034] 对所述样本篇章文本进行编码处理，得到样本篇章向量；
- [0035] 融合所述样本篇章向量和所述样本重构隐层向量序列，得到样本融合向量序列；
- [0036] 所述依据所述样本源文本、所述样本重构隐层向量序列、所述源端样本向量序列、以及所述目标端样本向量序列构建重构评估函数，包括：
- [0037] 依据所述样本源文本、所述样本融合向量序列、所述源端样本向量序列、以及所述目标端样本向量序列构建重构评估函数；
- [0038] 所述依据所述标注词序列和所述样本重构隐层向量序列，构建词序列标注损失函数，包括：
- [0039] 依据所述标注词序列和所述样本融合向量序列，构建词序列标注损失函数。
- [0040] 在其中一个实施例中，所述融合所述样本篇章向量和所述样本重构隐层向量序列，得到样本融合向量序列，包括：
- [0041] 通过以下公式融合所述样本篇章向量和所述样本重构隐层向量序列：
- [0042] $\hat{h}_i^{rec} = f_c(h_i^{rec}, M)$ ；
- [0043] 其中， $\hat{h}_i^{rec}$ 表示与所述样本源文本中的第i个词对应的样本融合向量； $f_c()$ 表示串联操作函数；M表示样本篇章向量； h_i^{rec} 表示与所述样本源文本中的第i个词对应的样本重构隐层向量。
- [0044] 在其中一个实施例中，所述模型训练方法还包括：
- [0045] 获取待翻译文本的词序列；
- [0046] 通过已训练的机器翻译模型的编码器，对所述待翻译文本的词序列进行语义编码处理，得到语义向量序列；
- [0047] 通过所述已训练的机器翻译模型的解码器，根据所述机器翻译模型前次输出的目标词的词向量，对所述语义向量序列进行解码，得到当次的解码隐层向量；
- [0048] 根据所述解码隐层向量，确定所述机器翻译模型当次输出的目标词；
- [0049] 根据所述机器翻译模型输出的各目标词，生成所述待翻译文本对应的目标译文。
- [0050] 一种模型训练装置，包括：
- [0051] 样本获取模块，用于获取样本源文本、参考目标译文、以及与所述源文本对应的标注词序列；所述标注词序列是对所述样本源文本进行省略词标注处理所得到的词序列；
- [0052] 输入模块，用于将所述样本源文本输入机器翻译模型中进行训练，得到预测目标译文；
- [0053] 构建模块，用于依据所述参考目标译文和预测目标译文构建最大似然函数；
- [0054] 所述获取模块还用于通过所述机器翻译模型，获取与所述样本源文本对应的源端样本向量序列、以及目标端样本向量序列；
- [0055] 重构模块，用于依据所述样本源文本，对所述源端样本向量序列、以及目标端样本向量序列进行重构处理，得到样本重构隐层向量序列；
- [0056] 所述构件模块还用于依据所述样本源文本、所述样本重构隐层向量序列、所述源端样本向量序列、以及所述目标端样本向量序列构建重构评估函数；

[0057] 所述构件模块还用于依据所述标注词序列和所述样本重构隐层向量序列,构建词序列标注损失函数;

[0058] 确定模块,用于根据所述最大似然函数、所述重构评估函数和所述词序列标注损失函数,确定目标函数;

[0059] 所述确定模块还用于将所述目标函数最大化时的模型参数作为机器翻译模型的模型参数,返回所述将所述样本源文本输入机器翻译模型中进行训练,得到预测目标译文的步骤并继续训练,直至满足训练停止条件时停止训练。

[0060] 一种计算机可读存储介质,存储有计算机程序,所述计算机程序被处理器执行时,使得所述处理器执行所述模型训练方法的步骤。

[0061] 一种计算机设备,包括存储器和处理器,所述存储器存储有计算机程序,所述计算机程序被所述处理器执行时,使得所述处理器执行所述模型训练方法的步骤。

[0062] 上述模型训练方法、装置、计算机可读存储介质和计算机设备,在模型训练过程中一方面将源端样本向量序列和目标端样本向量序列重构回样本源文本,另一方面依据标注词序列对重构得到的隐层向量序列进行缺略词标注。训练时,最大似然函数、重构评估函数和词序列标注损失函数叠加一起优化,使得翻译模型不仅可以学习到缺略词的位置信息,更具备还原缺略词的能力,使得训练得到的机器翻译模型具有更高的翻译准确性。

附图说明

[0063] 图1为一个实施例中文本翻译方法和/或模型训练方法的应用环境图;

[0064] 图2为一个实施例中文本翻译方法的流程示意图;

[0065] 图3为一个实施例中对汉语表达中缺略词被省略的实例示意图;

[0066] 图4为一个实施例中对源端向量序列进行解码处理的结构框图;

[0067] 图5为一个实施例中通过机器翻译模型对源文本进行翻译得到目标译文的流程示意图;

[0068] 图6为一个实施例中对篇章文本进行编码的结构示意图;

[0069] 图7为一个实施例中通过包括编码器-解码器结构的机器翻译模型进行文本翻译的流程示意图;

[0070] 图8为一个实施例中模型训练步骤的流程示意图;

[0071] 图9为一个实施例中机器翻译模型在训练时的训练任务示意图;

[0072] 图10为一个实施例中在模型训练过程中引入篇章信息的结构示意图;

[0073] 图11为一个具体实施例中文本翻译方法的流程示意图;

[0074] 图12为一个实施例中通过已训练的机器翻译模型翻译文本的步骤的流程示意图;

[0075] 图13为一个实施例中文本翻译装置的结构框图;

[0076] 图14为另一个实施例中文本翻译装置的结构框图;

[0077] 图15为一个实施例中模型训练装置的结构框图;

[0078] 图16为另一个实施例中模型训练装置的结构框图;

[0079] 图17为一个实施例中计算机设备的结构框图。

具体实施方式

[0080] 为了使本申请的目的、技术方案及优点更加清楚明白,以下结合附图及实施例,对本申请进行进一步详细说明。应当理解,此处所描述的具体实施例仅仅用以解释本申请,并不用于限定本申请。

[0081] 图1为一个实施例中文本翻译方法和/或模型训练方法的应用环境图。参照图1,该文本翻译方法和/或模型训练方法应用于文本处理系统。该文本处理系统包括终端110和服务器120。终端110和服务器120通过网络连接。终端110具体可以是台式终端或移动终端,移动终端具体可以手机、平板电脑、笔记本电脑等中的至少一种。服务器120可以用独立的服务器或者是多个服务器组成的服务器集群来实现。终端110和服务器120均可单独用于执行本申请实施例中提供的文本翻译方法和/或模型训练方法。终端110和服务器120也可协同用于执行本申请实施例中提供的文本翻译方法和/或模型训练方法。

[0082] 如图2所示,在一个实施例中,提供了一种文本翻译方法。本实施例主要以该方法应用于计算机设备来举例说明,该计算机设备具体可以是上图中的终端110或者服务器120。参照图2,该文本翻译方法具体包括如下步骤:

[0083] S202,获取源文本的词序列;源文本为指代词省略文本。

[0084] 其中,源文本是待翻译的文本,具体可以是词、句子、段落或篇章等文本。源文本可以是任意语种的文本,比如,中文文本或日文文本等。指代词省略文本是指缺少至少一部分的指代词文本,文本中缺少的指代词可称作省略词,通常是省略代词(DP, Dropped Pronoun)。

[0085] 在特定语言(如,汉语、日语)表达中,一些指代词常常会被省略从而使表达更加简洁。虽然这些指代词被省略了,但可以通过上下文语境推断出缺失的信息(即,省略代词及其指代的先行词),因而这种表达并不会影响人们对内容的理解。此现象属于篇章理论中的零指代或空指代范畴。参考图3,图3示出了一个实施例中在汉语表达中省略词被省略的实例示意图。如图3所示,汉语端括号中的主格代词“你”、“我”和宾格代词“它”都被省略了(括号中的文字被省略了),但在英文端对应的指代词则没有被省略。

[0086] 具体地,计算机设备可从本地或其他计算机设备处获取待翻译的源文本,并对源文本进行分词处理后得到的各个词所组成的词序列。其中,对于源文本为中文文本的分词处理,可采用基于词典或基于统计的分词方式,在本申请中不做限定。

[0087] 在一个实施例中,计算机设备在获取源文本的词序列后,可通过词嵌入(word embedding)处理将离散的词序列转换成连续的空间表示向量序列。再将空间表示向量序列输入至预训练的机器翻译模型的输入层,通过预训练的机器翻译模型执行文本翻译方法,以得到目标译文。

[0088] S204,对源文本的词序列进行语义编码,得到源端向量序列。

[0089] 其中,对源文本进行语义编码,是将源文本转换为向量的过程。源端向量序列,是指对源文本的词序列进行语义编码后得到的向量序列。具体地,计算机设备可分别对源文本的词序列中的各词进行语义编码,提取各词的语义特征,将各词转换成向量表示。拼接各词所对应的向量表示,得到源端向量序列。

[0090] 在一个实施例中,在一个实施例中,计算机设备在获取源文本的词序列后,可通过词嵌入处理将离散的词序列转换成连续的空间表示向量序列。再将空间表示向量序列输入

至预训练的机器翻译模型的输入层,通过预训练的机器翻译模型中的编码器对与源文本对应的空间表示向量序列进行语义编码处理,得到源端向量序列。其中,源端向量序列中的每个源端向量分别对应源文本的词序列中的各词。

[0091] 其中,预训练的机器翻译模型采用基于神经网络的Sequence-to-Sequence (序列-序列) 框架。Sequence-to-Sequence框架是包含Encoder-Decoder (编码器-解码器) 结构的框架。Encoder-Decoder结构将输入序列转换为另一个序列输出。在该框架中,编码器将输入序列转化成向量,解码器则接受向量并按时间顺序依次生成输出序列。编码器以及解码器可以采用相同类型的神经网络模型,也可以是不同类型的神经网络模型。神经网络模型比如CNN (Convolutional Neural Networks, 卷积神经网络) 模型、RNN (Recurrent Neural Networks) 模型、长短时记忆模型 (Long Short-Term Memory, LSTM)、时延网络模型或闸控卷积神经网络模型等。

[0092] 在一个实施例中,计算机设备可将源文本的词序列输入到预训练的机器翻译模型中编码器的输入层,通过编码器中各隐层进行语义编码处理,得到隐层所输出的向量序列。其中,隐层是神经网络模型中的一种术语,是相对于输入层和输出层的中间层,隐层中包括对神经网络模型训练得到的模型参数。这里编码器的隐层是相对于编码器的输入层和编码器的输出层的中间层。编码器的隐层可包括多个神经网络层。这里将编码器的隐层当作“黑盒”进行描述。

[0093] 举例说明,以源文本的词序列为一个长度为 m 的序列为例,将源文本的词序列输入到编码器中进行语义编码,得到词序列中各词对应的向量。编码器的处理过程如下:计算机设备可以对源文本进行分词,得到源文本的词序列为 $x = (x_1, x_2, \dots, x_m)$,通过词嵌入方式将词序列转换成连续的空间表示向量,可记为 $p = (p_1, p_2, \dots, p_m)$ 。通过编码器的神经网络结构,对空间表示向量进行语义编码,得到源端向量序列,可记为 $H^{enc} = (h_1^{enc}, h_2^{enc}, \dots, h_m^{enc})$ 。

[0094] S206,对源端向量序列进行解码处理,得到候选译文、各候选译文对应的目标端向量序列、及各候选译文对应的翻译得分。

[0095] 其中,解码处理是将向量转换为序列输出的过程。候选译文是解码所得到的与源文本对应的备选译文。可以理解,源文本具体可以是词、句子、段落或者篇章等。相应的,候选译文也可以是与源文本对应的词、句子、段落或者篇章等。源文本可以是任意语种的文本,候选译文则为异于源文本所属语种的其他语种的文本。

[0096] 目标端向量序列,是指对源文本对应的源端向量进行解码计算得到的向量所组成的序列。候选译文对应的目标端向量序列,是指构成候选译文的各候选词所对应的目标端向量所组成的序列。可以理解,计算机设备的解码过程是有时序的,在解码的过程中,计算机设备会根据前次解码得到的候选词的词向量,对源端向量序列进行解码,得到当次的目标端向量,进而根据当次的目标端向量确定当次的候选词。对于各候选译文而言,候选译文中的候选词所对应的目标端向量是可以确定的。因而,计算机设备在解码得到候选译文时,则可相应地确定与各候选译文分别对应的目标端向量序列。

[0097] 翻译得分是衡量候选译文的翻译质量的量化值。计算机设备在对源端向量序列进行解码处理时,可根据源端向量序列、历史解码得到的候选词、及历史解码得到的目标端向量序列,确定与各候选词对应的翻译得分。具体通过以下公式计算各候选词的翻译得分:

$$P(y|x) = \prod_{i=1}^I g(y_{i-1}, s_i, c_i)$$
。其中, y_{i-1} 是前次解码得到的候选词; s_i 是当次的目标端向量;

c_i 是对源端向量序列进行当次的注意力机制处理后得到的当次的内容向量; g 表示 softmax 函数; $\Pi()$ 函数表示连乘运算。关于目标端向量的具体计算方式及源端内容向量的具体计算方式在后面的实施例中进行详细介绍。

[0098] 进一步地, 计算机设备在计算出各候选词所对应的翻译得分后, 确定构成各候选译文各候选词所对应的翻译得分, 将候选译文各候选词所对应的翻译得分相加即可得到该候选译文对应的翻译得分。

[0099] 在一个实施例中, 计算机设备可通过预训练的机器翻译模型中的解码器依次对源端向量序列进行解码, 以得到目标端向量, 进而根据当次的目标端向量确定当次的候选词、以及与该候选词对应的翻译得分。其中, 机器翻译模型可将源端向量序列输入到解码器的隐层中计算得到目标端向量、候选词及翻译得分。这里解码器的隐层是相对于解码器的输入层和解码器的输出层的中间层。解码器的隐层可包括多个神经网络层。这里将解码器的隐层当做“黑盒”进行描述。在一个实施例中, 解码器在初始时刻的隐层状态向量可以是模型训练得到的模型参数, 也可以是默认设置的模型参数。

[0100] 在一个实施例中, 计算机设备根据目标端向量确定的候选词可以是一个, 也可以是多个。具体地, 解码器可根据当次(也可称作当前时刻)的目标端向量, 计算得到当次(也可称作当前时刻)的输出概率序列。其中, 机器翻译模型当次确定的输出概率序列是指在输出端词集中各候选词为当次输出的目标词的概率所组成的序列。进一步地, 机器翻译模型可选择输出概率序列中的最大概率所对应的候选词作为当次的候选词, 或者选择输出概率序列中的概率值最大的前几名(比如前十名)所对应的候选词作为当次的候选词。

[0101] 在一个实施例中, 解码器可根据当次的目标端向量, 确定当次的翻译得分序列。其中, 机器翻译模型当次确定的翻译得分序列是指在输出端词集中各候选词所对应的翻译得分的序列。翻译得分用于评价该候选词作为目标词时的翻译准确性。候选词所对应的翻译得分越大, 表示该候选词作为目标词的概率越大。进一步地, 机器翻译模型可选择翻译得分序列中的最大分数所对应的候选词作为当次的候选词。计算机设备也可选择翻译得分序列中的翻译得分最大的前几名(比如前十名)所对应的候选词作为当次的候选词。

[0102] 当当次确定的候选词的数量为多个时, 计算机设备在下次解码时, 可分别并行进行多组解码处理。机器翻译模型的解码器可分别根据前次确定的各候选词分别计算当次的目标端向量。这样, 计算得到的当次的目标端向量的数量也是多个, 再根据当次的目标端向量确定当次的候选目标词。依次类推, 直到解码得到多组候选译文。

[0103] 在一个实施例中, 为减轻计算机的计算压力, 在得到多组候选译文后, 可根据各组候选译文所对应的翻译得分, 筛选翻译得分较高的预设数量的候选译文。筛选的候选译文所对应的目标端向量序列才参与到后续的重构处理中。这样, 在计算各候选译文所对应的重构得分时, 预先筛选掉一部分的候选译文, 仅留下翻译质量较高的候选译文, 再计算相应的重构得分, 从而确定目标译文, 可大大减轻计算机设备的计算压力, 提高了翻译效率。

[0104] 在一个实施例中, 计算机设备在对源端向量序列进行解码时, 可采用贪心算法(greedy search)或集束搜索算法(beam search)等算法进行解码处理, 从而得到多组候选

译文。

[0105] 在一个实施例中,步骤S206具体包括:依次对源端向量序列进行解码得到目标端向量,且每次解码时根据前次确定的候选词的词向量进行解码,并根据当次的目标端向量确定当次的多于一个的候选词;将依次解码得到的各候选词,按候选词的生成时序,拼接得到多于一组的候选译文;通过构成各候选译文的候选词所对应的目标端向量,确定与各候选译文分别对应的目标端向量序列;根据源端向量序列、历史解码得到的候选词、及历史解码得到的目标端向量序列,确定与各候选词对应的翻译得分;依据构成各候选译文的各候选词所对应的翻译得分,计算得到与各候选译文对应的翻译得分。

[0106] 具体地,计算机设备每次解码时根据前次确定的候选词进行解码,得到当次的目标端向量,并根据当次的目标端向量确定当次的多于一个的候选词。计算机设备再将当次的多于一个的候选词分别作为下次解码的前次候选词并行解码,每组解码过程都会得到多于一个的候选词,直到结束解码。在解码过程中,每个候选词所对应的目标端向量是确定的,该候选词的前一词的词向量也是确定的,因而,每个候选词所对应的翻译得分也是确定的。

[0107] 进一步地,计算机设备可采用反向回溯的方式,从最后得到的候选词进行反向回溯,确定该候选词的前一词,依次回溯直到解码得到的初始候选词。这样,每一组回溯路径所包括的各候选词按相应的生成时序,即可拼接得到一组候选译文。对应地,将构成该候选译文的各候选词所对应的目标端向量进行拼接处理,即可得到与该候选译文对应的目标端向量序列。对应地,将构成该候选译文的各候选词所对应的翻译得分相加,即可得到与该候选译文对应的翻译得分。

[0108] 参考图4,图4为一个实施例中源端向量序列进行解码处理的结构框图。如图4所示,计算机设备每次解码时根据前次确定的候选词进行解码,并确定当次的候选词,比如当次确定的候选词有3个。在下次解码时,则分别进行三组解码,每组解码依据一个候选词确定下次的3个候选词,依次解码直到解码过程的结束。采用这样并行的多组解码方式进行解码,可得到多组候选译文,以及相应的翻译得分 $\{P_1, P_2, \dots, P_k\}$ 。

[0109] 在一个实施例中,在每次根据前次解码得到的候选词进行解码的过程中,计算机设备可依次获取当次的与源端向量序列对应的注意力分配权重向量,根据注意力分配权重向量、以及源端向量序列,计算得到当次的内容向量;依次根据当次的内容向量、前次的目标端向量、以及前次确定的候选词的词向量,计算得到当次的目标端向量,并根据当次的目标端向量确定当次的候选词。

[0110] 在一个实施例中,计算机设备在生成当次(也称当前时刻)的候选词之前,可获取当次(也称当前时刻)的与源端向量序列对应的注意力分配权重向量。其中,当次的注意力分配权重向量可根据源端向量序列、以及解码器中前次(前一时刻)的目标端向量所确定。

具体地,可参考以下公式:

$$\alpha_{i,j} = \frac{\exp(e_{i,j})}{\sum_{k=1}^{T_s} \exp(e_{i,k})}; \quad e_{i,j} = \text{ATT}(s_{i-1}, h_j)。$$

其中, i 指第 i 次, j 指源

端向量序列中第 j 个向量, $\alpha_{i,j}$ 指第 i 次中源端向量序列中第 j 个向量所对应的注意力分配权重。将每次源端向量序列中各个向量所对应的注意力分配权重进行拼接,即可得到当次与

源端向量序列对应的注意力分配权重向量。

[0111] 进一步地,计算机设备可根据当次的注意力分配权重向量、以及源端向量序列,计算得到当次的内容向量。当次的内容向量是对各源端向量加权求和得到的,相应的加权系数就是注意力分配权重。举例说明,对于当次的注意力分配权重向量,可以表示为 $\alpha_{i,j} = \{\alpha_1, \alpha_2, \dots, \alpha_J\}$ 。其中,源端向量序列可以表示为 $H^{enc} = \{h_1^{enc}, h_2^{enc}, \dots, h_J^{enc}\}$,则当次的内容向量可以通过以下公式计算: $C_i^{enc} = \sum_{j=1}^J \alpha_{i,j} \cdot h_j^{enc}$ 。相应的,对于每一次的内容向量都可以通过上述公式计算得到。

[0112] 进一步地,计算机设备可依次根据当次的内容向量、前次的目标端向量、以及前次确定的候选词的词向量,计算得到当次的目标端向量,并根据当次的目标端向量确定当次的候选词。举例说明,对于当次的目标端向量,可以用 S_i 表示,前次的候选词可以用 y_{i-1} 表示,当次的内容向量可以用 C_i 表示,则 $S_i = f(S_{i-1}, y_{i-1}, C_i)$;其中, $f()$ 表示激活函数。相应的,对于每一次的目标端向量都可以通过上述公式计算得到。进一步地,可通过以下公式计算确定当次的候选词: $y_i = g(y_{i-1}, S_i, C_i)$;其中, $g()$ 是softmax函数。对于每个候选词所对

应的翻译得分,计算机设备可通过以下公式计算: $P(y|x) = \prod_{i=1}^I g(y_{i-1}, S_i, C_i)$ 。

[0113] 这样通过注意力机制处理,可以使得当次的目标端向量包含源文本中对当次而言更重要的语义信息,再综合了前次所输出候选词的语义信息,可以使得生成的候选译文更为连贯,信息量更为丰富。

[0114] 可以理解的是,解码器首次解码得到候选词的解码方式和后面时刻解码的方式一样,对于首次解码时所需的前一时刻的相应的数据则可以根据经验预设初始值或随机生成初始值,以进行下一步的处理。

[0115] 可以理解,本申请所使用的“当次”或“当前时刻”用于描述机器翻译模型的解码器进行解码并确定当次的目标词的时刻,“前次”或“前一时刻”用于描述解码器进行解码并确定前次的目标词的时刻。比如,前次为第 $i-1$ 次,机器翻译模型所输出的目标词则为 y_{i-1} ;当次为第 i 次,机器翻译模型所输出的目标词则为 y_i 。并且,当次是一个相对变化的次数,比如,机器翻译模型在输出当次 i 的下次 $i+1$ 的目标词时,则可将该下次 $i+1$ 作为新的当次,将当次 i 作为新的前次。

[0116] 上述实施例中,依次对源端向量序列进行解码得到目标端向量,且每次解码时根据前次确定的候选目标词的词向量进行解码,并根据当次的目标端向量确定当次的多于一个候选词。可使得解码生成的候选词不仅包含了源文本的语义信息,还综合了前次的候选词的语义信息,使得生成的候选译文更为连贯,翻译结果更准确。由于在每次解码时都可解码得到多于一个的候选词,在不断地解码过程中,根据每个候选词又可解码得到下次的多于一个的候选词,因而根据各候选词的生成时序可拼接得到多个候选译文,根据各候选词的目标端向量可拼接得到与候选译文对应的目标端向量序列,根据各候选词的翻译得分可计算得到候选译文的翻译得分。

[0117] S208,将词序列和源端向量序列,与各目标端向量序列分别进行重构处理,得到与各候选译文分别对应的重构隐层向量序列;重构隐层向量序列包括源文本所缺少的指代词

的词信息。

[0118] 其中,重构处理是将源端向量序列和目标端向量序列重构回源文本的处理,可以理解成一个自动编码器(auto-encoder)操作。重构隐层向量序列是将源端向量序列和目标端向量序列重构回源文本的过程中隐层所输出的向量序列。指代词的词信息具体可以是指代词的位置信息和/或指代词的词本身。

[0119] 具体地,计算机设备可根据源端向量序列和各目标端向量序列,重新构建与源文本的词序列对应的重构隐层向量序列。对于每组候选译文,计算机设备均采用以下方式计算各组候选译文对应的重构隐层向量序列:计算机设备可根据源文本中当前词的前一词的词向量、前次重构得到的重构隐层向量、源端向量序列和该组候选译文对应的目标端向量序列,重构得到当次重构隐层向量。依次重构处理,直到得到每次重构处理的重构隐层向量,将每次得到的重构隐层向量构成重构隐层向量序列。

[0120] 在一个实施例中,可通过预训练的机器翻译模型执行文本翻译方法。计算机设备可将源端向量序列和目标端向量序列输入至机器翻译模型的重构器中,通过重构器的隐层将编码器端的隐层向量序列和解码器端的隐层向量序列重构回源语端句子。这相当于一个自动编码器(auto-encoder)操作,我们假设此过程可以使源端向量序列和目标端向量序列学习到源文本中缺少的指代词的词信息。其中,重构器可以由深层神经网络所构成。这里重构器的隐层是相对于重构器的输入层和重构器的输出层的中间层。重构器的隐层可包括多个神经网络层。这里将重构器的隐层当做“黑盒”进行描述。

[0121] 在一个实施例中,机器翻译模型在训练时,可联合词序列标注任务联合训练,也就是,在训练时,一方面将源端向量序列和目标端向量序列重构回源文本;另一方面通过预先与源文本进行序列标注所得到的标注词序列,引导重构隐层向量序列学习到指代词的词信息。这样训练得到的机器翻译模型的模型参数,在通过机器翻译模型重构得到重构隐层向量序列时,无需额外输入源文本所缺少的指代词的词信息作为参考,就可通过已训练好的机器翻译模型的模型参数,使得构建的重构隐层向量序列包括源文本所缺少的指代词的词信息。避免了由于额外对源文本进行还原所带来的错误而影响翻译质量。虽然通过平行语料训练还原模型,可以实现对源文本中缺少的指代词的还原和位置预测,但一般来说,对源文本中缺少的指代词的词还原的准确率只有66%左右,对源文本中缺少的指代词的位置预测的准确率最高也只能达到88%左右。因而在对源文本进行翻译的过程中,由于引入源文本所缺少的指代词的词信息作为参考,在模型计算的迭代过程中,引入的词信息中的错误部分会不断迭代放大,最终影响到机器翻译的翻译质量。而本申请所提供的各种实施例中,在通过机器翻译模型对源文本进行翻译时,无需引入额外的缺略词的词信息作为参考,通过训练得到的机器翻译模型的模型参数,使得机器翻译模型在构建重构隐层向量序列时,自动包括了源文本所缺少的指代词的词信息,大大提高了翻译质量。其中,关于机器翻译模型的训练步骤在后面的实施例中再做详细介绍。

[0122] S210,将源文本的词序列,与各重构隐层向量序列分别进行重构评估处理,得到与各候选译文分别对应的重构得分。

[0123] 其中,重构评估处理是通过对经重构处理所得到的重构隐层向量序列和源文本进行比较,以评估从重构隐层向量序列到生成源文本的代价的负。其中,代价的负是指重构得分越大,重构的代价越小,说明重构隐层向量序列越容易重构需要的内容。而重构得分是衡

量省略代词信息召回程度的分值。重构得分可用来衡量各候选译文的翻译质量,重构得分越高,则表示该候选译文作为目标译文的可能性越高。

[0124] 在一个实施例中,对于每组候选译文的重构得分,计算机设备可通过如下方式计算:首先,计算机设备可根据源文本中当前词的前一词的词向量、源端向量序列、以及该候选译文对应的当次的重构隐层向量和目标端向量序列,计算得到与该候选译文对应、且与源文本中当前词对应的重构得分。再对与源文本中各词所对应的重构得分进行连乘运算,得到与该候选译文对应的重构得分。

[0125] 在一个实施例中,可通过预训练的机器翻译模型执行文本翻译方法。对于每组候选译文的重构得分,计算机设备可通过如下方式计算:计算机设备可将源端向量序列和该候选译文所对应的目标端向量序列输入至机器翻译模型的重构器中,通过源文本的词序列,重构得到重构隐层向量序列。再通过机器翻译模型的重构器,对源文本的词序列、重构隐层向量序列、源端向量序列和目标端向量序列进行重构评估处理,得到重构得分。

[0126] S212,从候选译文中筛选相应翻译得分和重构得分满足预设条件的目标译文。

[0127] 其中,目标译文是对源文本进行机器翻译得到的最终文本。目标译文为异于源文本所属语种的其他语种的文本。目标译文具体可以是与源文本对应的词、句子、段落或者篇章等。

[0128] 预设条件是预先设置的条件,具体可以通过翻译得分和重构得分计算得到的综合得分大于或等于阈值。其中,综合得分的计算方式不限定,比如可对翻译得分和重构得分进行加权求和得到综合得分,或将翻译得分和重构得分相乘得到综合得分等。或者,翻译得分及重构得分各自分别大于预设的相应阈值。又或者,将根据翻译得分和重构得分计算得到的综合得分进行从高到低排序后,排序名次小于预设名次等。

[0129] 具体地,计算机设备可根据各候选译文各自所对应的翻译得分和重构得分,筛选出翻译得分和重构得分,满足预设条件的候选译文,将筛选出的候选译文作为目标译文。

[0130] 在一个实施例中,步骤S212,也就是从候选译文中筛选相应翻译得分和重构得分满足预设条件的目标译文的步骤具体包括:对各候选译文分别对应的翻译得分和重构得分进行加权求和计算,得到各候选译文各自所对应的综合得分;从候选译文中,筛选相应综合得分为最高得分的候选译文;将筛选出的候选译文作为目标译文并输出。

[0131] 具体地,计算机设备可对各候选译文各自所对应的翻译得分和重构得分进行加权求和计算,得到各候选译文各自所对应的综合得分。进而选择综合得分最大的候选译文为与源文本对应的目标译文。这样,依据翻译得分和重构得分,从候选译文中筛选出目标译文,可以极大地缓解源文本中代词缺失的痛点问题,大大提高了翻译的准确性。

[0132] 参考图5,图5示出了一个实施例中,通过机器翻译模型对源文本进行翻译得到目标译文的流程示意图。如图5所示,计算机设备可将源文本(source sentence)输入至机器翻译模型的编码器-解码器(encoder-decoder)中,得到候选译文 $\{cand_1, cand_2, \dots, cand_k\}$,以及各候选译文所对应的翻译得分 $\{P_1, P_2, \dots, P_k\}$ 。其中,候选译文对应的翻译得分可以通过将候选译文中各候选目标词对应的翻译得分进行加法计算得到。再将源文本、源端向量序列和目标端向量序列输入至机器翻译模型的重构器中进行重构处理,得到各候选译文对应的重构隐层向量序列。再通过重构器对源文本的词序列、重构隐层向量序列、源端向量序列、及目标端向量序列进行重构评估处理,得到各候选译文对应的重构得分 $\{R_1,$

$R_2, \dots, R_k$ 。计算机设备可对各候选译文的翻译得分和重构得分进行加权求和,得到综合得分,比如综合得分 $y = P_k + \lambda R_k$,其中, λ 为加权系数。最后通过选择器将最高的综合得分所对应的候选译文作为目标译文输出。

[0133] 上述文本翻译方法,对缺少指代词的源文本进行语义编码,得到源端向量序列,并继续对源端向量序列进行解码,得到候选译文、各候选译文对应的目标端向量序列、及各候选译文对应的翻译得分,这样可充分提取出源端和目标端的隐层信息。再通过融合源端和目标端的隐层信息可以共享两端的共性信息,在根据源文本进行重构处理时可以减少冗余信息并能避免重要信息的缺失,不需要额外的省略词标注信息就可重构回包括缺少的指代词的词信息的重构隐层向量序列,这样使得确定的重构得分可以很好的衡量缺少的指代词信息的召回程度。因此,依据重构得分和翻译得分所筛选出的目标译文,可以减少翻译时缺少省略代词所带来的影响,大大提高了翻译质量。

[0134] 在一个实施例中,该文本翻译方法还包括分别对源端向量序列和目标端向量序列进行注意力机制处理的步骤,该步骤具体包括:获取与源端向量序列对应的源端注意力分配权重向量;根据源端注意力分配权重向量、以及源端向量序列,计算得到源端内容向量;分别获取与各目标端向量序列对应的目标端注意力分配权重向量;根据目标端注意力分配权重向量、以及相对应的目标端向量序列,计算得到与各候选译文分别对应的目标端内容向量。将词序列和源端向量序列,与各目标端向量序列分别进行重构处理,得到与各候选译文分别对应的重构隐层向量序列的步骤,包括:将词序列和源端内容向量,与各目标端内容向量分别进行重构处理,得到与各候选译文分别对应的重构隐层向量序列。

[0135] 在一个实施例中,计算机设备对源端向量序列、以及目标端向量序列进行注意力机制处理的过程中可采用独立注意力机制处理或交叉注意力机制处理的方式。其中独立注意力机制处理是指对源端向量序列、以及目标端向量序列进行注意力机制处理的注意力分配权重各自独立计算;交叉注意力机制处理是指对源端向量序列、以及目标端向量序列进行注意力机制处理的注意力分配权重交叉计算。

[0136] 在一个实施例中,获取与源端向量序列对应的源端注意力分配权重向量的步骤具体包括:对源文本中当前词的前一词的词向量、重构器所输出的与源文本中前一词对应的重构隐层向量、以及源端向量序列进行注意力机制处理,得到与源端向量序列对应的源端注意力分配权重向量;或,对源文本中当前词的前一词的词向量、重构器所输出的与源文本中前一词对应的重构隐层向量、源端向量序列、以及目标端内容向量进行注意力机制处理,得到与源端向量序列对应的源端注意力分配权重向量。

[0137] 在一个实施例中,获取与目标端向量序列对应的目标端注意力分配权重向量的步骤包括:对源文本中当前词的前一词的词向量、重构器所输出的与源文本中前一词对应的重构隐层向量、以及目标端向量序列进行注意力机制处理,得到与目标端向量序列对应的目标端注意力分配权重向量;或,对源文本中当前词的前一词的词向量、重构器所输出的与源文本中前一词对应的重构隐层向量、目标端向量序列、以及源端内容向量进行注意力机制处理,得到与目标端向量序列对应的目标端注意力分配权重向量。

[0138] 下面通过举例说明来介绍计算机设备对源端向量序列、以及目标端向量序列进行独立注意力机制处理的方式。以源文本的词序列的词数为次数基准,源文本的第 i 个词对应第 i 次重构处理,本申请实施例中的第 i 次重构处理也可以理解成当次重构处理。计算机设

备可采用以下公式计算得到与各源端向量分别对应的源端注意力分配权重： $\alpha_{i,j}^{enc} = ATT(x_{i-1}, h_{i-1}^{rec}, h_j^{enc})$ ，其中，ATT表示注意力机制处理； x_{i-1} 表示源文本中的第i-1个词的词向量； h_{i-1}^{rec} 表示与源文本中的第i-1个词对应的重构隐层向量； h_j^{enc} 表示各源端向量。计算机设备可采用以下公式计算得到与各目标端向量分别对应的目标端注意力分配权重： $\alpha_{i,j}^{dec} = ATT(x_{i-1}, h_{i-1}^{rec}, h_j^{dec})$ ，其中，ATT表示注意力机制处理； x_{i-1} 表示源文本中的第i-1个词的词向量； h_{i-1}^{rec} 表示与源文本中的第i-1个词对应的重构隐层向量； h_j^{dec} 表示各目标端向量。

[0139] 在一个实施例中，计算机设备对源端向量序列、以及目标端向量序列进行交叉注意力机制处理过程中注意力分配权重的计算有多种方式，下面通过多种举例来进行说明。

[0140] 方式一：计算机设备可采用以下公式计算得到与各源端向量分别对应的源端注意力分配权重： $\alpha_{i,j}^{enc} = ATT(x_{i-1}, h_{i-1}^{rec}, h_j^{enc})$ ，其中，ATT表示注意力机制处理； x_{i-1} 表示源文本中的第i-1个词的词向量； h_{i-1}^{rec} 表示与源文本中的第i-1个词对应的重构隐层向量； h_j^{enc} 表示各源端向量。计算机设备可采用以下公式计算得到与各目标端向量分别对应的目标端注意力分配权重： $\alpha_{i,j}^{dec} = ATT(x_{i-1}, h_{i-1}^{rec}, h_j^{dec}, c_i^{enc})$ ， $c_i^{enc} = \sum_{j=1}^J \alpha_{i,j}^{enc} \cdot h_j^{enc}$ ；其中，ATT表示

注意力机制处理； x_{i-1} 表示源文本中的第i-1个词的词向量； h_{i-1}^{rec} 表示与源文本中的第i-1个词对应的重构隐层向量； h_j^{dec} 表示各目标端向量； c_i^{enc} 表示当次的，也就是与源文本中的第i个词对应的源端内容向量； $\alpha_{i,j}^{enc}$ 表示当次的源端注意力分配权重； h_j^{enc} 表示各源端向量； $\cdot$ 表示向量点乘运算。

[0141] 方式二：计算机设备可采用以下公式计算得到与各源端向量分别对应的源端注意力分配权重： $\alpha_{i,j}^{enc} = ATT(x_{i-1}, h_{i-1}^{rec}, h_j^{enc}, c_i^{dec})$ ， $c_i^{dec} = \sum_{j=1}^J \alpha_{i,j}^{dec} \cdot h_j^{dec}$ ；其中，ATT表示注意力机制处理； x_{i-1} 表示源文本中的第i-1个词的词向量； h_{i-1}^{rec} 表示与源文本中的第i-1个词对应的重构隐层向量； h_j^{enc} 表示各源端向量； c_i^{dec} 表示当次的也就是与源文本中的第i个词对应的目标端内容向量； h_j^{dec} 表示各目标端向量； $\alpha_{i,j}^{dec}$ 表示当次的目标端注意力分配权重； $\cdot$ 表示向量点乘运算。计算机设备可采用以下公式计算得到与各目标端向量分别对应的目标端注意力分配权重： $\alpha_{i,j}^{dec} = ATT(x_{i-1}, h_{i-1}^{rec}, h_j^{dec})$ ，其中，ATT表示注意力机制处理； x_{i-1} 表示源文本中的第i-1个词的词向量； h_{i-1}^{rec} 表示与源文本中的第i-1个词对应的重构隐层向量； h_j^{dec} 表示各目标端向量。

[0142] 方式三：计算机设备可采用以下公式计算得到与各源端向量对应的源端注

意力分配权重: $\alpha_{i,j}^{enc} = ATT(x_{i-1}, h_{i-1}^{rec}, h_j^{enc}, c_i^{dec})$, $c_i^{dec} = \sum_{j=1}^J \alpha_{i,j}^{dec} \bullet h_j^{dec}$; 其中, ATT 表示注意力机制处理; x_{i-1} 表示源文本中的第 $i-1$ 个词的词向量; h_{i-1}^{rec} 表示与源文本第 $i-1$ 个词对应的重构隐层向量; h_j^{enc} 表示各源端向量; c_i^{dec} 表示当次的, 也就是与源文本中的第 i 个词对应的目标端内容向量; h_j^{dec} 表示各目标端向量; $\alpha_{i,j}^{dec}$ 表示与各目标端向量分别对应的目标端注意力分配权重; $\bullet$ 表示向量点乘运算。计算机设备可采用以下公式计算得到与各目标端向量分别对应的目标端注

意力分配权重: $\alpha_{i,j}^{dec} = ATT(x_{i-1}, h_{i-1}^{rec}, h_j^{dec}, c_i^{enc})$, $c_i^{enc} = \sum_{j=1}^J \alpha_{i,j}^{enc} \bullet h_j^{enc}$; 其中, ATT 表示注意力机制处理; x_{i-1} 表示源文本中的第 $i-1$ 个词的词向量; h_{i-1}^{rec} 表示与源文本第 $i-1$ 个词对应的重构隐层向量; c_i^{enc} 表示当次的, 也就是与源文本中的第 i 个词对应的源端内容向量; $\alpha_{i,j}^{enc}$ 表示与各源端向量分别对应的源端注意力分配权重; h_j^{enc} 表示各源端向量; $\bullet$ 表示向量点乘运算。

[0143] 进一步地, 当计算机设备分别确定了当次的源端注意力分配权重向量和当次的目标端注意力分配权重向量时, 可根据源端注意力分配权重向量、以及源端向量序列, 计算得到当次的源端内容向量; 根据目标端注意力分配权重向量、以及目标端向量序列, 计算得到当次的目标端内容向量。这样, 通过独立或交叉的关注机制处理源端向量序列和目标端向量序列, 可更好地利用源端和目标端的共性知识, 能更好的共享源端和目标端的隐层特征, 提高文本翻译的准确性。

[0144] 在一个实施例中, 计算机设备可通过以下公式计算当次的源端内容向量:

$c_i^{enc} = \sum_{j=1}^J \alpha_{i,j}^{enc} \bullet h_j^{enc}$; 其中, h_j^{enc} 表示各源端向量; $\alpha_{i,j}^{enc}$ 表示与各源端向量分别对应的源端注意力分配权重; $\bullet$ 表示向量点乘运算。计算机设备可通过以下公式计算当次的目标端

内容向量: $c_i^{dec} = \sum_{j=1}^J \alpha_{i,j}^{dec} \bullet h_j^{dec}$; 其中, h_j^{dec} 表示各目标端向量; $\alpha_{i,j}^{dec}$ 表示与各目标端向量分别对应的目标端注意力分配权重; $\bullet$ 表示向量点乘运算。

[0145] 进一步地, 计算机设备在计算得到当次的源端内容向量和当次的目标端内容向量后, 可按照源文本, 对源端内容向量和目标端内容向量进行重构处理, 得到当次的重构隐层向量。

[0146] 在一个实施例中, 将词序列和源端内容向量, 与各目标端内容向量分别进行重构处理, 得到与各候选译文分别对应的重构隐层向量序列的步骤具体包括: 对于每组候选译文, 分别根据源文本中当前词的前一词的词向量、前次重构得到的重构隐层向量、源端内容向量和各候选译文各自所对应的目标端内容向量, 重构得到当次重构隐层向量; 对于每组候选译文, 分别将依次重构得到的重构隐层向量构成重构隐层向量序列; 重构得到的重构隐层向量序列包括源文本所缺少的指代词的词信息。

[0147] 在一个实施例中,对于每组候选译文分别所对应的重构隐层向量序列,计算机设备均可通过下面的方式计算得到;首选,计算机设备可通过以下公式计算与源文本中当前词对应的重构隐层向量: $h_i^{rec} = f_r(x_{i-1}, h_{i-1}^{rec}, c_i^{enc}, c_i^{dec})$; 其中, $f_r()$ 为激活函数; x_{i-1} 表示源文本中的第 $i-1$ 个词向量; h_i^{rec} 表示与源文本中的第 i 个词对应的重构隐层向量; c_i^{enc} 表示当次的源端内容向量; c_i^{dec} 表示当次的目标端内容向量。依次计算,直到得到与源文本中各个词分别对应的重构隐层向量。接着,计算机设备可将依次重构得到的重构隐层向量构成重构隐层向量序列。其中,重构得到的重构隐层向量序列包括源文本所缺少的指代词的词信息。这样,通过将各候选译文分别对应的目标端内容向量,依次与源文本的词序列和源端内容向量进行重构处理,可以很好地构建包括省略词的词信息重构隐层向量序列,可以用于后续步骤中重构得分的计算。

[0148] 可以理解,计算机设备可通过预训练的机器翻译模型对源文本进行处理,而机器翻译模型在训练时,可结合源文本重构任务和词序列标注任务联合训练,也就是,在训练时,一方面将源端向量序列和目标端向量序列重构回源文本;另一方面通过预先与源文本进行序列标注所得到的标注词序列,引导重构隐层向量序列学习到指代词的位置信息,并具备还原省略词的能力。这样训练得到的机器翻译模型的模型参数,在通过机器翻译模型重构得到重构隐层向量序列时,无需额外输入源文本所缺少的指代词的词信息作为参考,就可通过已训练好的机器翻译模型的模型参数,使得构建的重构隐层向量序列包括源文本所缺少的指代词的词信息。

[0149] 上述实施例中,分别对源端向量序列和各目标端向量序列进行注意力机制处理,得到对应的源端内容向量和目标端内容向量。将源文本和源端内容向量,与各目标端内容向量分别进行重构处理,可以得到融合了源端和目标端信息重构隐层向量序列,且该隐层向量序列还包括了源文本所缺少的指代词的词信息,可以帮助生成能准确衡量省略词找回程度的重构得分。

[0150] 在一个实施例中,将源文本的词序列,与各重构隐层向量序列分别进行重构评估处理,得到与各候选译文分别对应的重构得分的步骤具体包括:对于每组候选译文,分别根据源文本中当前词的前一词的词向量、源端内容向量、以及各候选译文对应的当次的重构隐层向量和目标端内容向量,计算得到与各候选译文分别对应、且与源文本中当前词对应的重构得分;对于每组候选译文,分别对与源文本中各词所对应的重构得分进行连乘运算,得到与各候选译文分别对应的重构得分。

[0151] 具体地,对于每组候选译文所对应的重构得分,计算机设备都可通过以下方式计算得到:计算机设备可根据源文本的词序列、源端内容向量、该组候选译文对应的目标端内容向量、以及该组候选译文对应的重构隐层向量序列,计算得到相应的重构得分。

[0152] 在一个实施例中,对于每组候选译文所对应的重构得分均可通过以下公式计算得

到: $R(x|h^{enc}, h^{dec}) = \prod_{i=1}^I g_r(x_{i-1}, h_i^{rec}, c_i^{enc}, c_i^{dec})$; 其中, $g_r()$ 为归一化函数; x_{i-1} 表示源文

本中的第 $i-1$ 个词向量; h_i^{rec} 表示与源文本中的第 i 个词对应的重构隐层向量; c_i^{enc} 表示当

次的源端内容向量; c_i^{dec} 表示当次的目标端内容向量; $\Pi()$ 表示连乘运算。

[0153] 上述实施例中,根据源文本的词序列、源端内容向量、该组候选译文对应的目标端内容向量、以及该组候选译文对应的重构隐层向量序列,计算得到相应的重构得分,可以通过量化的重构得分的分值来衡量重构隐层向量序列和源文本的差异,也就是可以评估出将重构隐层向量序列重构回源文本的代价。

[0154] 在实际应用中,对源文本进行翻译的过程中,与源文本中缺少的指代词对应的先行词可能在该源文本的前述文本中。比如,“这块蛋糕很美味!你烤的吗?”,假如当前句子为“你烤的吗?”,很明显,当前句子所缺少的指代词是“它”,该句子完整的表述应是“你烤的(它)吗”,而指代词(也就是该句中的省略词)“它”的先行词为前一句中的“蛋糕”。如果仅考虑当前句子,在机器翻译的时候可能根据当前句很难将省略词还原准确。

[0155] 基于上述的情况,可通过引入篇章信息来进一步提高机器翻译的准确率。在本申请的各实施例中,篇章信息具体包括跨句子的信息,比如当前句子的前N句话的词级建模信息,以及句子之前起承转合的关系。这些都将有助于机器翻译模型对省略词预测和还原的学习。

[0156] 在一个实施例中,计算机设备可通过层次级网络对源文本的篇章文本进行建模,将篇章信息融入到重构隐层向量中,通过模型的训练,使机器翻译模型的重构器学习到省略词补充的能力。其中,计算机设备将篇章信息融入到重构隐层向量中,具体可以是将篇章信息分别和各候选译文对应的目标端向量序列进行融合,还可以是将篇章信息分别和各重构隐层向量序列进行融合。

[0157] 在一个实施例中,该文本翻译方法还包括融合篇章信息的步骤,该步骤具体包括:获取与源文本对应的篇章文本;对篇章文本进行编码处理,得到篇章向量;将篇章向量,与各重构隐层向量序列分别进行融合,得到与各候选译文分别对应的第一融合向量序列。将源文本的词序列,与各重构隐层向量序列分别进行重构评估处理,得到与各候选译文分别对应的重构得分的步骤包括:将源文本的词序列,与各第一融合向量序列分别进行重构评估处理,得到与各候选译文分别对应的重构得分。

[0158] 其中,与源文本对应的篇章文本具体可以是当前待翻译的句子之前和/或之后的预设数量的句子。具体地,计算机设备可从待翻译文本中,确定位于当前待翻译的句子之前的篇章文本(也就是该句子的上下文)。继而,计算机设备可对篇章文本进行编码处理,得到篇章向量。通过建模的方式,将篇章文本的信息压缩到篇章向量中。这样,在计算重构得分时,可分别融合篇章向量来进行计算。对于每组候选译文各自对应的重构得分,计算机设备都可采用如下方式计算:将篇章向量与该候选译文所对应的重构隐层向量序列进行融合,得到第一融合向量序列。再根据第一融合向量序列计算得到该候选译文所对应的重构得分。

[0159] 在一个实施例中,计算机设备可通过神经网络结构,比如RNN网络(Recurrent Neural Network,循环神经网络)结构、CNN网络(Convolutional Neural Networks,卷积神经网络)结构、或者SAN网络(Storage Area Networks)结构等,对篇章文本进行编码处理,输出相应的篇章向量。

[0160] 在一个实施例中,对篇章文本进行编码处理,得到篇章向量的步骤具体包括:确定

篇章文本所包括的整句;对各整句分别进行分词处理,得到各整句各自对应的词序列;对各整句各自对应的词序列分别进行词级编码处理,得到与各整句分别对应的句子向量序列;将与各整句分别对应的句子向量序列进行句子级编码处理,得到与篇章文本对应的篇章向量。

[0161] 具体地,计算机设备可根据篇章文本中包括的结束符来确定篇章文本中的各整句。其中,结束符是表示句子结束的符号,比如句号、问号或感叹号等。计算机设备可将两个结束符之间的字符构建成一个整句。

[0162] 进一步地,计算机设备对分别对各整句进行分词处理,得到各整句各自对应的词序列。计算机设备可对各整句各自对应的词序列分别进行词级编码处理,得到与各整句分别对应的句子向量序列。将与各整句分别对应的句子向量序列进行句子级编码处理,得到与篇章文本对应的篇章向量。

[0163] 举例说明,将源文本的前K句话作为篇章文本 $X = \{X^{-K}, \dots, X^{-1}\}$,计算机设备可通过层次级神经网络对篇章文本进行建模。对每一个整句,都通过词级编码器对句子进行编码,得到对应的句子向量序列: $h^{-K} = \text{ENCODER}_{\text{word}}(X^{-K})$ 。当得到所有的句子向量序列 $H^X = \{h^{-K}, \dots, h^{-1}\}$ 后,将所有的句子向量序列输入至句子级编码器进行编码处理,得到篇章向量 M ,其中, $M = \text{ENCODER}_{\text{sentence}}(H^X)$ 。其中,这里编码得到的句子向量 M 不仅包含了词与词之间的语义关系,更隐含了句子之间的逻辑关系。

[0164] 在一个实施例中,参考图6,图6示出了一个实施例中对篇章文本进行编码的结构示意图。如图6所示,计算机设备可通过词级编码器对各个整句分别进行编码,得到各个整句各自对应的句子向量序列。再根据多组句子向量序列构建出一组长度为句子数的综合句子向量序列。比如,计算机设备可分别提取将各组句子向量序列中的一个向量,共同拼接组成一组包含所有句子信息的综合句子向量序列。其中,计算机设备提取各组句子向量序列中的一个向量的方式,具体可以是提取各句子向量序列中相同位置处的向量,比如各组句子向量序列中的第一个向量或最后一个向量;也可以是提取各句子向量序列中不同位置处的向量等,在此不做限定。还比如,计算机设备可将各组句子向量序列分别再次进行编码,得到代表该组句子向量序列的单个句子向量,根据得到的单个句子向量共同拼接生成综合句子向量序列。进一步地,参考图6,计算机设备可通过句子级编码器对综合句子向量序列进行句子级编码,得到表示篇章信息的篇章向量 M 。

[0165] 上述实施例中,分层级对篇章文本进行编码,可得到包含了词与词之间的语义关系,更隐含了句子之间的逻辑关系的篇章向量。

[0166] 在一个实施例中,计算机设备可直接将篇章向量融入到重构器中。具体地,计算机设备可以将篇章向量与重构隐层向量序列进行融合。比如,采用以下公式融合篇章向量:

$\hat{h}_i^{rec} = f_c(h_i^{rec}, M)$; 其中, $\hat{h}_i^{rec}$ 表示源文本中的第 i 个词对应的第一融合向量; $f_c()$ 表示串联(concatenation)操作函数; M 表示篇章向量; h_i^{rec} 表示与源文本中的第 i 个词对应的重构隐层向量。

[0167] 在一个实施例中,计算机设备可将 M 的长度复制多次与重构隐层向量的长度一致,从而进行concatenation操作。融合后的每个位置的隐层状态的维度为 M 的维度与该位置处原重构隐层状态的维度之和。

[0168] 上述实施例中,通过对篇章文本进行编码处理,得到篇章向量,再将篇章向量融合到重构隐层向量序列中,加强了在翻译过程中对篇章信息的感知,从而使得确定的重构得分更为准确,进而提高了目标译文的准确性。

[0169] 在一个实施例中,该文本翻译方法还包括将篇章信息与编码端融合的步骤,该步骤具体包括:获取与源文本对应的篇章文本;对篇章文本进行编码处理,得到篇章向量;将篇章向量,与各目标端向量序列分别进行融合,得到与各候选译文分别对应的第二融合向量序列。将词序列和源端向量序列,与各目标端向量序列分别进行重构处理,得到与各候选译文分别对应的重构隐层向量序列的步骤具体包括:将词序列和源端向量序列,与各第二融合向量序列分别进行重构处理,得到与各候选译文分别对应的重构隐层向量序列。

[0170] 具体地,计算机设备可从待翻译文本中,确定位于当前待翻译的句子之前和/或之后的预设数据的多个句子作为与源文本对应的篇章文本。继而,计算机设备可对篇章文本进行编码处理,得到篇章向量。通过建模的方式,将篇章文本的信息压缩到篇章向量中。再将篇章向量,与各目标端向量序列分别进行融合,得到与各候选译文分别对应的第二融合向量序列。在重构隐层向量序列时,可根据融合了篇章信息的第二融合向量序列,重构得到相应的重构隐层向量序列。这样,在计算重构得分时,也充分考虑了篇章信息。

[0171] 关于将篇章向量与目标端向量序列融合的方式,具体可参考前述实施例中将篇章向量与重构隐层向量序列进行融合的方式。比如,采用以下公式融合篇章向量:

$\hat{h}_i^{dec} = f_c(h_i^{dec}, M)$; 其中, $\hat{h}_i^{dec}$ 表示源文本中的第i个词对应的第一融合向量; $f_c()$ 表示串联(concatenation)操作函数; M 表示篇章向量; h_i^{dec} 表示与源文本中的第i个词对应的重构隐层向量。相应的,重构隐层向量序列的计算公式相应更新为: $\hat{h}_i^{rec} = f_r(x_{i-1}, h_{i-1}^{rec}, c_i^{enc}, \hat{c}_i^{dec})$; 其中, $\hat{c}_i^{dec}$ 为对第二融合向量序列进行注意力机制处理后所得到的目标端内容向量。

[0172] 上述实施例中,通过对篇章文本进行编码处理,得到篇章向量,再融合篇章向量和各目标端向量序列,从而得到第二融合向量序列。再第二融合向量序列,计算得到与各候选译文分别对应的重构隐层向量序列,加强了在翻译过程中对篇章信息的感知,从而使得确定的重构得分更为准确,进而提高了目标译文的准确性。

[0173] 在一个实施例中,参考图7,图7示出了一个实施例中,通过包括编码器-解码器结构的机器翻译模型进行文本翻译的流程示意图。首先将源文本(也就是输入句子)输入机器翻译模型的编码器中,通过编码器模块输出源端向量序列。再将源端向量序列输入注意力模块,通过注意力机制处理模块对源端向量序列进行注意力机制处理,得到当次的源端内容向量,也就是当前时刻的源端上下文。再将当前时刻的源端上下文输入机器翻译模型的解码器中,通过解码器模块对当前时刻的源端上下文进行解码处理,输出当前时刻的目标端向量,并根据当前时刻的目标端向量确定当前时刻的候选词。依次重复上述步骤,直到多组候选译文、各候选译文对应的目标端向量序列、及各候选译文对应的翻译得分。通过注意力机制处理模块对源端向量序列进行注意力机制处理,得到源端内容向量;通过注意力机制处理模块依次对各组目标端向量序列进行注意力机制处理,得到相应的目标端内容向量。将源文本、源端内容向量和目标端内容向量输入至机器翻译模型的重构器中进行重构处理,得到与各组候选译文分别对应的重构隐层向量序列。另一方面,通过层次级神经网络

结构对篇章文本进行建模,得到篇章向量表示(也就是篇章向量)。通过重构器融合重构隐层向量序列和篇章向量表示,进而通过融合后的向量序列确定与各候选译文分别对应的重构得分。最后通过选择器,根据重构得分和翻译得分,从候选译文中筛选出目标译文并输出。

[0174] 在一个实施例中,该文本翻译方法通过已训练的机器翻译模型执行,该机器翻译模型的训练步骤包括:

[0175] S802,获取样本源文本、参考目标译文、以及与源文本对应的标注词序列;标注词序列是对样本源文本进行省略词标注处理所得到的词序列。

[0176] 其中,样本源文本、参考目标译文、以及与源文本对应的标注词序列为训练数据。参考目标译文是样本源文本的标准翻译译文,是低噪声或无噪声的翻译文本。参考目标译文具体可以是人工根据样本源文本进行翻译所得到的翻译文本,也可以是人工对机器翻译文本进行校正后所得到的翻译文本。计算机设备可从网络爬取低噪声或无噪声的平行语料作为样本源文本和相应的参考目标译文。可以理解,参考目标译文至少满足语句通顺、所表达的含义与源文本的表达含义相同或相近等条件。源文本具体可以是词、句子、段落或者篇章等。相应的,参考目标译文可以是与源文本对应的词、句子、段落或者篇章等。源文本可以是任意语种的文本,参考目标译文则为异于源文本所属语种的其他语种的文本。

[0177] 标注词序列是对样本源文本进行省略词标注处理后所得到的词序列。对样本源文本进行省略词标注处理,是指标注出样本源文本中缺少的指代词的过程。在一个实施例中,计算机设备可通过对样本源文本和参考目标译文进行词对齐处理,再确定出样本源文本中所缺少的指代词。或者,计算机设备可通过省略代词语料库自动标注方法,得到平行语料中的样本源文本的省略词的信息。

[0178] 举例说明,当样本源文本为“你烤的吗?</s>”(其中,“</s>”为结束符),参考目标译文为“Did you bake it?”,计算机设备可通过对样本源文本和参考目标译文进行词对齐处理,识别出样本源文本中省略词的位置并在该位置处生成相应的指代词,得到“你烤的(它)吗?”,从而得到标注词序列“NNN'它NN”。其中,标注词序列中,除还原的指代词外的其他词均可用同一指代符比如字母“N”替代,从而达到标注省略词的效果。标注词序列中包含了省略词的位置信息和省略词本身。

[0179] S804,将样本源文本输入机器翻译模型中进行训练,得到预测目标译文。

[0180] 具体地,计算机设备可将样本源文本的词序列输入到机器翻译模型中,根据机器翻译模型执行前述的文本翻译方法,得到预测目标译文。计算机设备可朝着减小预测目标译文和参考目标译文之间的差异的方向,调整模型参数。这样,通过不断的输入样本源文本的词序列,得到预测目标译文,根据预测目标译文和参考目标译文之间的差异调整模型参数,以训练机器翻译模型。

[0181] S806,依据参考目标译文和预测目标译文构建最大似然函数。

[0182] 其中,最大似然函数用于评估参考目标译文与预测目标译文之间的差异程度。最大似然函数承担了训练得到好的机器翻译模型的责任,使模型生成语法正确、文字流畅的目标文本。

[0183] 在一个实施例中,可通过以下公式构建最大似然函数:

$J_{likelihood}(\theta) = \arg \max_{\theta} \sum_{n=1}^N \log P(y^n | x^n; \theta)$; 其中, $\arg \max$ 是表示寻找具有最大值的参

量的函数, 就是求使得 $\log P(y^n | x^n; \theta)$ 的值最大时的 θ 值。 $\{[x^n, y^n]\}_{n=1}^N$ 是模型训练中的训练样本对。 $\log$ 为对数函数, $\Sigma()$ 为叠加运算, θ 为模型参数。

[0184] S808, 通过机器翻译模型, 获取与样本源文本对应的源端样本向量序列、以及目标端样本向量序列。

[0185] 具体地, 计算机设备可将源文本的词序列输入至机器翻译模型, 通过机器翻译模型中的编码器对源文本进行语义编码, 得到与源文本对应的源端向量序列。通过机器翻译模型的解码器, 依次对源端向量序列进行解码得到目标端向量, 且每次解码时根据前次确定的候选词的词向量进行解码, 并根据当次的目标端向量确定当次的候选词。将依次解码得到的目标端向量拼接成目标端向量序列。

[0186] S810, 依据样本源文本, 对源端样本向量序列、以及目标端样本向量序列进行重构处理, 得到样本重构隐层向量序列。

[0187] 具体地, 计算机设备可将源端样本向量序列和目标端样本向量序列输入至机器翻译模型的重构器中, 通过重构器将编码器端的隐层向量序列和解码器端的隐层向量序列重构回源语端句子。这相当于一个自动编码器 (auto-encoder) 操作。计算机设备可根据源文本中当前词的前一词的词向量、前次重构得到的重构隐层向量、源端向量序列和该组候选译文对应的目标端向量序列, 重构得到当次重构隐层向量。依次重构处理, 直到得到每次重构处理的重构隐层向量, 将每次得到的重构隐层向量构成重构隐层向量序列。

[0188] 在一个实施例中, 计算机设备可通过以下公式计算与源文本中当前词对应的重构隐层向量: $h_i^{rec} = f_r(x_{i-1}, h_{i-1}^{rec}, c_i^{enc}, c_i^{dec})$; 其中, $f_r()$ 为激活函数; x_{i-1} 表示源文本中的第 $i-1$ 个词向量; h_i^{rec} 表示与源文本中的第 i 个词对应的重构隐层向量; c_i^{enc} 表示当次的源端内容向量; c_i^{dec} 表示当次的目标端内容向量。依次计算, 直到得到与源文本中各个词分别对应的重构隐层向量。接着, 计算机设备可将依次重构得到的重构隐层向量构成重构隐层向量序列。

[0189] S812, 依据样本源文本、样本重构隐层向量序列、源端样本向量序列、以及目标端样本向量序列构建重构评估函数。

[0190] 其中, 重构评估函数用来衡量省略代词信息召回程度, 可以辅助机器翻译模型的重构器所输出的隐层表示可以更好的学习得关于省略词的词信息。具体地, 计算机设备可通过样本源文本的词序列、样本重构隐层向量序列、源端样本向量序列、以及目标端样本向量序列构建重构评估函数。

[0191] 在一个实施例中, 可通过以下公式构建重构评估函数:

[0192] $J_{reconstruction}(\theta, \gamma) = \arg \max_{\theta, \gamma} \log R(x | h^{enc}, h^{dec}; \theta, \gamma)$;

[0193] $R(x | h^{enc}, h^{dec}) = \prod_{i=1}^I g_r(x_{i-1}, h_i^{rec}, c_i^{enc}, c_i^{dec})$; $h_i^{rec} = f_r(x_{i-1}, h_{i-1}^{rec}, c_i^{enc}, c_i^{dec})$;

$$[0194] \quad c_i = \sum_{j=1}^J \alpha_{i,j} \bullet h_j;$$

[0195] 其中, $R(x|h^{enc}, h^{dec})$ 表示重构得分; θ 和 γ 分别是模型参数; $g_r()$ 和 $f_r()$ 分别为归一化函数和激活函数; x_{i-1} 表示源文本中的第 $i-1$ 个词的词向量; h_i^{rec} 表示与源文本中的第 i 个词对应的隐层状态向量序列; h 表示各源端向量或目标端向量; $\alpha_{i,j}$ 表示源端注意力分配权重或目标端注意力分配权重; $\bullet$ 表示点乘运算。

[0196] S814, 依据标注词序列和样本重构隐层向量序列, 构建词序列标注损失函数。

[0197] 其中, 词序列标注损失函数用来衡量省略词信息标注的程度, 可以辅助机器翻译模型的重构器可以的学习关于省略词的标注信息。具体地, 计算机设备可依据标注词序列和样本重构隐层向量序列, 构建词序列标注损失函数。

[0198] 在一个实施例中, 可通过以下公式构建词序列标注损失函数:

$$[0199] \quad J_{\text{labeling}}(\theta, \gamma, \psi) = \arg \max_{\theta, \gamma, \psi} \log P(dp|h^{rec}; \theta, \gamma, \psi);$$

$$[0200] \quad P(dp|h^{rec}) = \prod_{i=1}^I P(dp_i|h_i^{rec}) = \prod_{i=1}^I g_i(dp_i|h_i^{rec});$$

[0201] 其中, $g_i()$ 是归一化函数; $dp = \{dp_1, dp_2, \dots, dp_I\}$ 是样本标注词序列; h_i^{rec} 表示与源文本中的第 i 个词对应的重构隐层向量。

[0202] S816, 根据最大似然函数、重构评估函数和词序列标注损失函数, 确定目标函数。

[0203] 具体地, 计算机设备可对最大似然函数、重构评估函数和词序列标注损失函数进行加权求和计算, 将加权求和函数作为机器翻译模型的目标函数。机器翻译模型可通过以下公式得到机器学习模型在训练过程中的目标函数: $J = J_{\text{likelihood}} + \lambda J_{\text{reconstruction}} + \tau J_{\text{labeling}}$ 。其中, λ 和 τ 分别是加权系数。

[0204] 在一个实施例中, 可通过以下函数表示目标函数:

$$[0205] \quad J(\theta, \gamma, \psi) = \arg \max_{\theta, \gamma, \psi} \sum_{n=1}^N \{\log L(y^n|x^n; \theta) + \lambda \log R(x|h^{enc}, h^{dec}; \theta, \gamma) + \tau \log P(dp|h^{rec}; \theta, \gamma, \psi)\};$$

$$[0206] \quad \text{其中, } R(x|h^{enc}, h^{dec}) = \prod_{i=1}^I g_r(x_{i-1}, h_i^{rec}, c_i^{enc}, c_i^{dec}); \quad h_i^{rec} = f_r(x_{i-1}, h_{i-1}^{rec}, c_i^{enc}, c_i^{dec});$$

$$[0207] \quad c_i = \sum_{j=1}^J \alpha_{i,j} \bullet h_j; \quad P(dp|h^{rec}) = \prod_{i=1}^I P(dp_i|h_i^{rec}) = \prod_{i=1}^I g_i(dp_i|h_i^{rec})$$

[0208] 其中, θ 、 γ 和 τ 分别是模型参数; $\{[x^n, y^n]\}_{n=1}^N$ 是模型训练中的样本平行语料对; $g_r()$ 和 $f_r()$ 分别为归一化函数和激活函数; $g_i()$ 是归一化函数; x_{i-1} 表示样本源文本中的第 $i-1$ 个词的词向量; h_i^{rec} 表示与样本源文本中的第 i 个词对应的样本重构隐层向量; h_j 表示各源端样本向量或目标端样本向量; $\alpha_{i,j}$ 表示源端注意力分配权重或目标端注意力分配权重; $\bullet$ 表示向量点乘运算; $dp = \{dp_1, dp_2, \dots, dp_I\}$ 是样本标注词序列。

[0209] S818, 将目标函数最大化时的模型参数作为机器翻译模型的模型参数, 返回将样

本源文本输入机器翻译模型中进行训练,得到预测目标译文的步骤并继续训练,直至满足训练停止条件时停止训练。

[0210] 其中,训练停止条件是结束模型训练的条件。训练停止条件可以是达到预设的迭代次数,或者是调整模型参数后的机器翻译模型的性能指标达到预设指标。调整机器翻译模型的模型参数,是对机器翻译模型的模型参数进行调整。

[0211] 具体地,对于每个样本序列对应的目标函数,取目标函数最大时的模型参数作为机器翻译模型的模型参数,然后在该模型参数的基础上对下一个样本序列进行预测,以对模型参数继续进行训练,直至满足训练停止条件时停止训练。

[0212] 在一个实施例中,参考图9,图9示出了机器翻译模型在训练时的训练任务示意图。如图9所示,计算机设备可通过编码器(encoder)和解码器(decoder)对样本源文本x“你烤的吗?</s>”进行编解码处理,得到预测目标译文y“Did you bake it?</s>”。并将编码器输出的与样本源文本对应的源端样本向量序列,以及解码器输出的与预测目标译文对应的目标端样本向量序列输入至机器翻译模型的重构器中。重构器依据样本源文本的词序列,根据源端样本向量序列和目标端样本向量序列,重构得到样本重构隐层向量序列。机器翻译模型在进行重构处理的过程中,会做两个任务。参考图9,图9左侧上方表示重构回样本源文本的训练任务;图9左侧下方表示的是依据标注词序列做省略词标注的训练任务。这样,就能通过多任务的联合训练的方式,使得机器翻译模型在重构得到样本重构隐层向量序列时可以学习到省略词的词信息。进而重构器可通过样本重构隐层向量序列和样本源文本的词序列计算得到重构得分,用于辅助衡量编码器和解码器对省略代词的还原能力。

[0213] 机器翻译模型引入额外的共享式重构层(也就是重构器),其输入为编码器端和解码器端的隐层状态向量序列和样本源文本的词序列,输出为重构得分。训练时,最大似然函数可以衡量翻译质量;重构评估函数可以衡量源端表示和目标端表示对源语端句子的总结能力;词序列标注损失函数可以辅助编解码器和重构器学习更好的预测省略词的词信息。合并三个学习函数可以统一地加强整体翻译质量,并且有效地引导模型参数朝还原省略代词的方向发展。

[0214] 模型训练完毕后,通过标准的编码器-解码器框架、重构器、及训练好的模型参数进行文本翻译,源端和目标端的隐层向量表示在通过自动编码器操作后,已经具有了对源文本所缺少的指代词的召回能力,因此,可以提高翻译质量。

[0215] 上述实施例中,在模型训练过程中一方面将源端样本向量序列和目标端样本向量序列重构回样本源文本,另一方面依据标注词序列对重构得到的隐层向量序列进行省略词标注。训练时,最大似然函数、重构评估函数和词序列标注损失函数叠加一起优化,使得翻译模型不仅可以学习到省略词的位置信息,更具备还原省略词的能力,使得训练得到的机器翻译模型具有更高的翻译准确性。

[0216] 在一个实施例中,该机器翻译模型的训练步骤还包括:获取与样本源文本对应的样本篇章文本;对样本篇章文本进行编码处理,得到样本篇章向量;融合样本篇章向量和样本重构隐层向量序列,得到样本融合向量序列。依据样本源文本、样本重构隐层向量序列、源端样本向量序列、以及目标端样本向量序列构建重构评估函数的步骤包括:依据样本源文本、样本融合向量序列、源端样本向量序列、以及目标端样本向量序列构建重构评估函数。依据标注词序列和样本重构隐层向量序列,构建词序列标注损失函数的步骤包括:依据

标注词序列和样本融合向量序列,构建词序列标注损失函数。

[0217] 其中,与样本源文本对应的样本篇章文本具体可以是当前用于训练机器翻译模型的源文本之前的预设数量的句子。具体地,计算机设备在获取样本源文本时,可一并获取该样本源文本之前的预设数量的句子作为样本篇章文本。继而,计算机设备可对样本篇章文本进行编码处理,得到样本篇章向量。通过建模的方式,将样本篇章文本的信息压缩到样本篇章向量中。关于对样本篇章文本进行编码的具体处理方式,可参考前述实施例中计算机设备对篇章文本进行编码得到篇章向量的说明。

[0218] 进一步地,计算机设备在训练机器学习模型时,可将样本篇章向量融入到样本重构隐层向量中进行训练,以此加强机器翻译模型对省略词的标注信息和翻译的学习。计算机设备可采取不同的融合方式来融合样本篇章向量,比如,计算机设备可将样本篇章向量和目标端样本向量进行融合,或者将样本篇章向量和样本重构隐层向量序列进行融合等。

[0219] 在一个实施例中,计算机设备可融合样本篇章向量和样本重构隐层向量序列,得到样本融合向量序列。具体可通过以下公式融合样本篇章向量和样本重构隐层向量序列:

$$[0220] \quad \hat{h}_i^{rec} = f_c(h_i^{rec}, M);$$

[0221] 其中, $\hat{h}_i^{rec}$ 表示源文本中的第i个词对应的样本融合向量; $f_c()$ 表示串联操作函数; M 表示样本篇章向量; h_i^{rec} 表示与源文本中的第i个词对应的样本重构隐层向量。这样,通过拼接各样本融合向量,得到样本融合向量序列。

[0222] 进一步地,步骤S812,也就是依据样本源文本、样本重构隐层向量序列、源端样本向量序列、以及目标端样本向量序列构建重构评估函数的步骤,具体可包括:依据样本源文本、样本融合向量序列、源端样本向量序列、以及目标端样本向量序列构建重构评估函数。具体可通过以下公式构建重构评估函数:

$$[0223] \quad \hat{J}_{reconstruction}(\theta, \gamma) = \arg \max_{\theta, \gamma} \log R(x | h^{enc}, h^{dec}, M; \theta, \gamma);$$

$$[0224] \quad R(x | h^{enc}, h^{dec}) = \prod_{i=1}^I g_r(x_{i-1}, \hat{h}_i^{rec}, c_i^{enc}, c_i^{dec}); \quad \hat{h}_i^{rec} = f_c(h_i^{rec}, M)$$

$$[0225] \quad h_i^{rec} = f_r(x_{i-1}, h_{i-1}^{rec}, c_i^{enc}, c_i^{dec}); \quad c_i = \sum_{j=1}^J \alpha_{i,j} \bullet h_j;$$

[0226] 其中, $R(x | h^{enc}, h^{dec}, M)$ 表示重构得分; θ 和 γ 分别是模型参数; $g_r()$ 和 $f_r()$ 分别为归一化函数和激活函数; $f_c()$ 为串联操作函数; x_{i-1} 表示样本源文本中的第i-1个词的词向量; h_i^{rec} 表示与样本源文本中的第i个词对应的样本隐层状态向量; M 表示样本篇章向量; $\hat{h}_i^{rec}$ 表示与样本源文本中的第i个词对应的样本融合向量; h_j 表示各源端样本向量或目标端样本向量; $\alpha_{i,j}$ 表示源端注意力分配权重或目标端注意力分配权重; $\bullet$ 表示点乘运算。

[0227] 进一步地,步骤S814,也就是根据最大似然函数、重构评估函数和词序列标注损失函数,确定目标函数的步骤,具体包括:依据标注词序列和样本融合向量序列,构建词序列标注损失函数。具体可通过以下公式构建词序列标注损失函数:

$$[0228] \quad \hat{J}_{labeling}(\theta, \gamma, \psi) = \arg \max_{\theta, \gamma, \psi} \log P(dp | \hat{h}^{rec}; \theta, \gamma, \psi);$$

$$[0229] \quad P(dp | \hat{h}^{rec}) = \prod_{i=1}^I P(dp_i | \hat{h}_i^{rec}) = \prod_{i=1}^I g_l(dp_i | \hat{h}_i^{rec}); \quad \hat{h}_i^{rec} = f_c(h_i^{rec}, M)$$

[0230] 其中, $g_l()$ 是归一化函数; $dp = \{dp_1, dp_2, \dots, dp_I\}$ 是样本标注词序列; $\hat{h}_i^{rec}$ 表示与源文本中的第 i 个词对应的重构隐层向量; M 表示样本篇章向量; $\hat{h}_i^{rec}$ 表示与样本源文本中的第 i 个词对应的样本融合向量。

[0231] 可以理解, 对于融入了篇章信息的机器学习模型在训练过程中的目标函数可通过以下公式计算:

$$[0232] \quad J(\theta, \gamma, \psi) = \arg \max_{\theta, \gamma, \psi} \sum_{n=1}^N \{\log L(y^n | x^n; \theta) + \lambda \log R(x | h^{enc}, h^{dec}, M; \theta, \gamma) + \tau \log P(dp | \hat{h}^{rec}; \theta, \gamma, \psi)\}。$$

[0233] 可以理解, 对于将样本篇章向量和目标端样本向量进行融合处理的机器学习模型, 其训练过程可以参考上述将样本篇章向量和样本重构隐层向量序列进行融合处理的机器学习模型的训练步骤, 其区别仅在于将样本篇章向量和目标端样本向量进行融合的机器学习模型, 在将样本篇章向量和目标端样本向量进行融合后, 再根据融合后的向量序列计算得到相应的样本重构隐层向量。此时的样本重构隐层向量间接地融入了样本篇章的信息。再根据样本重构隐层向量构建相应的重构评估函数和词序列标注损失函数, 从而构建相应的目标函数。

[0234] 进而, 在机器翻译模型的训练过程中, 计算机设备可对于每个样本序列对应的目标函数, 取目标函数最大时的模型参数作为机器翻译模型的模型参数, 然后在该模型参数的基础上对下一个样本序列进行预测, 以对模型参数继续进行训练, 直至满足训练停止条件时停止训练。

[0235] 在一个实施例中, 参考图10, 图10示出了一个实施例中在模型训练过程中, 引入篇章信息的结构示意图。参考图10, 计算机设备可通过编码器 (encoder) 和解码器 (decoder) 对样本源文本 x “你烤的吗?” 进行编解码处理, 得到预测目标译文 y “Did you bake it?”。并将编码器输出的与样本源文本对应的源端向量序列, 以及解码器输出的与预测目标译文对应的目标端向量序列输入至机器翻译模型的重构器 (reconstructor) 中。重构器依据样本源文本的词序列, 根据源端样本向量序列和目标端样本向量序列, 重构得到样本重构隐层向量序列。如图10中间框架中的上部分所示, 机器翻译模型会依据样本源文本的词序列, 将源端样本向量序列和目标端样本向量序列, 重构回样本重构隐层向量序列。或者, 计算机设备可依据融合了篇章向量的样本重构隐层向量, 来重构得到下次的样本重构隐层向量, 从而更好地融合篇章的信息。

[0236] 在另一方面, 计算机设备可对样本源文本的样本篇章文本进行编码处理, 得到样本篇章向量。比如, 将篇章文本 $X = \{X^K, \dots, X^1\}$, “你在烤什么?” …… “这块蛋糕很美味” 进行编码, 得到样本篇章向量 M 。计算机设备可将样本篇章 M 的长度复制多次与样本重构隐层向量的长度一致, 从而进行 concatenation 操作。如图10中间框架中的下部分所示, 机器翻译模型根据标注词序列和样本融合向量序列, 做缺略词标注 (DP labeling) 的训

练任务。这样,就能通过多任务的联合训练的方式,使得机器翻译模型在重构得到样本重构隐层向量序列时可以学习到省略词的词信息。进而重构器可通过样本重构隐层向量序列和样本源文本的词序列计算得到重构得分,用于辅助衡量编码器和解码器对省略代词的还原能力。

[0237] 机器翻译模型引入额外的共享式重构层(也就是重构器),其输入为编码器端和解码器端的隐层状态向量序列和样本源文本的词序列,输出为重构得分。训练时,最大似然函数可以衡量翻译质量;重构评估函数可以衡量源端表示和目标端表示对源语端句子的总结能力;词序列标注损失函数可以辅助编解码器和重构器学习更好的预测省略词的词信息。合并三个学习函数可以统一地加强整体翻译质量,并且有效地引导模型参数朝还原省略代词的方向发展。

[0238] 在一个实施例中,本申请提供的依据机器翻译模型实施的各种实施例,一方面共享编码器端和解码器端的隐层状态向量信息,另一方面可以对省略代词的词表层还原任务和翻译任务进行统一建模。这样基于共享重构机制的省略代词还原与神经网络机器翻译联合学习的文本翻译方法,进一步增强了机器翻译模型对省略代词的关注和学习能力,从而进一步提高翻译质量。通过上述各种实施例中的模型训练方法训练得到的机器翻译模型,是一种端到端模型,也就是在通过机器翻译模型进行翻译任务时,不需要额外的部件进行流水线作业(Pipeline Model),也就是,在翻译时无需先利用额外的DP还原装置对输入的源文本进行DP补充,再借助补充发DP词信息辅助翻译。

[0239] 本申请提供的方法可以用在所有主流神经网络机器翻译系统中,比如,RNN模型或Transformer(变压器)框架系统,并适用于所有代词省略语言到非代词省略语言的翻译任务(如,日语到英语、阿拉伯语到英语等)。并且,本申请各种实施例在文本翻译上,翻译质量都有显著提升。

[0240] 在实际应用中,通过本发明各种实施例所提供的方法可极大地提高文本翻译质量。如下表1:

	模型	翻译效果	
		BLEU	Δ
[0241]	基线系统	33. 18	--
	模型 1 (不进行篇章融合处理)	36. 04	+2. 86
	模型 2 (进行篇章融合处理)	37. 37	+4. 19

[0242] 表1多种翻译模型的翻译效果比对

[0243] 其中,BLEU(bilingual evaluation understudy,双语评价指标)是一种译文评价指标,一般提高超过1个点即是显著提高; Δ 是指提高的绝对数值。其中,模型1是通过采用不进行篇章融合方式训练得到的机器翻译模型;模型2是通过采用进行篇章融合处理方式训练得到的机器翻译模型。这两种模型都能很好地提升翻译质量。

[0244] 在一个具体的实施例中,参考图11,该文本翻译方法具体包括以下步骤:

- [0245] S1102,获取源文本的词序列;源文本为指代词省略文本。
- [0246] S1104,对源文本的词序列进行语义编码,得到源端向量序列。
- [0247] S1106,依次对源端向量序列进行解码得到目标端向量,且每次解码时根据前次确定的候选词的词向量进行解码,并根据当次的目标端向量确定当次的多于一个的候选词。
- [0248] S1108,将依次解码得到的各候选词,按候选词的生成时序,拼接得到多于一组的候选译文。
- [0249] S1110,通过构成各候选译文的候选词所对应的目标端向量,确定与各候选译文分别对应的目标端向量序列。
- [0250] S1112,根据源端向量序列、历史解码得到的候选词、及历史解码得到的目标端向量序列,确定与各候选词对应的翻译得分。
- [0251] S1114,依据构成各候选译文的各候选词所对应的翻译得分,计算得到与各候选译文对应的翻译得分。
- [0252] S1116,获取与源端向量序列对应的源端注意力分配权重向量。
- [0253] S1118,根据源端注意力分配权重向量、以及源端向量序列,计算得到源端内容向量。
- [0254] S1120,分别获取与各目标端向量序列对应的目标端注意力分配权重向量。
- [0255] S1122,根据目标端注意力分配权重向量、以及相对应的目标端向量序列,计算得到与各候选译文分别对应的目标端内容向量。
- [0256] S1124,对于每组候选译文,分别根据源文本中当前词的前一词的词向量、前次重构得到的重构隐层向量、源端内容向量和各候选译文各自所对应的目标端内容向量,重构得到当次重构隐层向量。
- [0257] S1126,对于每组候选译文,分别将依次重构得到的重构隐层向量构成重构隐层向量序列;重构得到的重构隐层向量序列包括源文本所缺少的指代词的词信息。
- [0258] S1128,获取与源文本对应的篇章文本。
- [0259] S1130,确定篇章文本所包括的整句。
- [0260] S1132,对各整句分别进行分词处理,得到各整句各自对应的词序列。
- [0261] S1134,对各整句各自对应的词序列分别进行词级编码处理,得到与各整句分别对应的句子向量序列。
- [0262] S1136,将与各整句分别对应的句子向量序列进行句子级编码处理,得到与篇章文本对应的篇章向量。
- [0263] S1138,将篇章向量,与各重构隐层向量序列分别进行融合,得到与各候选译文分别对应的第一融合向量序列。
- [0264] S1140,对于每组候选译文,分别根据源文本中当前词的前一词的词向量、源端内容向量、以及各候选译文对应的当次的融合向量序列和目标端内容向量,计算得到与各候选译文分别对应、且与源文本中当前词对应的重构得分。
- [0265] S1142,对于每组候选译文,分别对与源文本中各词所对应的重构得分进行连乘运算,得到与各候选译文分别对应的重构得分。
- [0266] S1144,对各候选译文分别对应的翻译得分和重构得分进行加权求和计算,得到各候选译文各自所对应的综合得分。

[0267] S1146,从候选译文中,筛选相应综合得分为最高得分的候选译文。

[0268] S1148,将筛选出的候选译文作为目标译文并输出。

[0269] 上述文本翻译方法,对缺少指代词的源文本进行语义编码,得到源端向量序列,并继续对源端向量序列进行解码,得到候选译文、各候选译文对应的目标端向量序列、及各候选译文对应的翻译得分,这样可充分提取出源端和目标端的隐层信息。再通过融合源端和目标端的隐层信息可以共享两端的共性信息,在根据源文本进行重构处理时可以减少冗余信息并能避免重要信息的缺失,不需要额外的缺略词标注信息就可重构回包括缺少的指代词的词信息的重构隐层向量序列,这样使得确定的重构得分可以很好的衡量缺少的指代词信息的召回程度。因此,依据重构得分和翻译得分所筛选出的目标译文,可以减少翻译时缺少缺略代词所带来的影响,大大提高了翻译质量。

[0270] 图11为一个实施例中文本翻译方法的流程示意图。应该理解的是,虽然图11的流程图中的各个步骤按照箭头的指示依次显示,但是这些步骤并不是必然按照箭头指示的顺序依次执行。除非本文中有明确的说明,这些步骤的执行并没有严格的顺序限制,这些步骤可以以其它的顺序执行。而且,图11中的至少一部分步骤可以包括多个子步骤或者多个阶段,这些子步骤或者阶段并不必然是在同一时刻执行完成,而是可以在不同的时刻执行,这些子步骤或者阶段的执行顺序也不必然是依次进行,而是可以与其它步骤或者其它步骤的子步骤或者阶段的至少一部分轮流或者交替地执行。

[0271] 在一个实施例中,提供了一种模型训练方法。本实施例主要以该方法应用于计算机设备来举例说明,该计算机设备具体可以是上图中的终端110或者服务器120。

[0272] 该模型训练方法具体包括如下步骤:获取样本源文本、参考目标译文、以及与源文本对应的标注词序列;标注词序列是对样本源文本进行缺略词标注处理所得到的词序列;将样本源文本输入机器翻译模型中进行训练,得到预测目标译文;依据参考目标译文和预测目标译文构建最大似然函数;通过机器翻译模型,获取与样本源文本对应的源端样本向量序列、以及目标端样本向量序列;依据样本源文本,对源端样本向量序列、以及目标端样本向量序列进行重构处理,得到样本重构隐层向量序列;依据样本源文本、样本重构隐层向量序列、源端样本向量序列、以及目标端样本向量序列构建重构评估函数;依据标注词序列和样本重构隐层向量序列,构建词序列标注损失函数;根据最大似然函数、重构评估函数和词序列标注损失函数,确定目标函数;将目标函数最大化时的模型参数作为机器翻译模型的模型参数,返回将样本源文本输入机器翻译模型中进行训练,得到预测目标译文的步骤并继续训练,直至满足训练停止条件时停止训练。

[0273] 关于模型训练方法中各个步骤的详细说明可参考前述实施例中机器翻译模型的模型训练步骤的说明,训练方式是一致的,在此不做重复说明。

[0274] 在一个实施例中,该机器训练方法还包括融合篇章信息的步骤,该步骤具体包括:获取与样本源文本对应的样本篇章文本;对样本篇章文本进行编码处理,得到样本篇章向量;融合样本篇章向量和样本重构隐层向量序列,得到样本融合向量序列。依据样本源文本、样本重构隐层向量序列、源端样本向量序列、以及目标端样本向量序列构建重构评估函数的步骤,包括:依据样本源文本、样本融合向量序列、源端样本向量序列、以及目标端样本向量序列构建重构评估函数。依据标注词序列和样本重构隐层向量序列,构建词序列标注损失函数的步骤,包括:依据标注词序列和样本融合向量序列,构建词序列标注损失函数。

[0275] 关于模型训练方法中对样本篇章文本进行编码、及融合样本篇章向量的步骤的说明可参考前述实施例中对样本篇章文本进行编码、及融合样本篇章向量的步骤的说明,以及对融合了篇章信息的机器学习模型的训练方式的说明可参考前述实施例中融合了篇章信息的机器学习模型的训练方式,训练方式是一致的,在此不做重复说明。

[0276] 上述模型训练方法,在模型训练过程中一方面将源端样本向量序列和目标端样本向量序列重构回样本源文本,另一方面依据标注词序列对重构得到的隐层向量序列进行省略词标注。训练时,最大似然函数、重构评估函数和词序列标注损失函数叠加一起优化,使得翻译模型不仅可以学习到省略词的位置信息,更具备还原省略词的能力,使得训练得到的机器翻译模型具有更高的翻译准确性。

[0277] 在一个实施例中,通过上述模型训练方法训练得到的机器翻译模型,可以用来翻译文本。参照图12,通过该机器翻译模型翻译文本的步骤具体包括以下步骤:

[0278] S1202,获取待翻译文本的词序列。

[0279] 具体地,计算机设备可从本地或其他计算机设备处获取待翻译文本,并对待翻译文本进行分词处理,得到待翻译文本的词序列。

[0280] S1204,通过已训练的机器翻译模型的编码器,对待翻译文本的词序列进行语义编码处理,得到语义向量序列。

[0281] 具体地,计算机设备可将待翻译文本的词序列输入至通过前述各实施例所描述的模型训练方法训练得到的机器翻译模型中。通过已训练的机器翻译模型的编码器对源文本的词序列进行语义编码,得到相应的语义向量序列。

[0282] S1206,通过已训练的机器翻译模型的解码器,根据机器翻译模型前次输出的目标词的词向量,对语义向量序列进行解码,得到当次的解码隐层向量。

[0283] 其中,解码隐层向量是将词序列中各词的词向量对应的语义向量序列输入到解码器的隐层中计算得到的向量。这里解码器的隐层是相对于解码器的输入层和解码器的输出层的中间层。解码器的隐层可包括多个神经网络层。

[0284] 具体地,当机器翻译模型在生成当次的目标词之前,可获取前次输出的目标词。依据前次机器翻译模型所输出目标词的词向量,通过解码器对编码器输出的语义向量序列进行解码,得到当次的解码隐层向量。解码隐层向量不仅包含了待翻译文本的词序列中各词的语义信息,且综合了前次所输出目标词的语义信息,使得最终生成的目标译文更为连贯,翻译结果更准确。

[0285] 在一个实施例中,机器翻译模型可在要生成当次的解码隐层向量时,获取前次所输出目标词。机器翻译模型可对前次输出的目标词做词嵌入处理,将目标词转换成用实数表示的词向量。

[0286] 关于机器翻译模型的解码器对语义向量序列进行解码的详细步骤,可参考前述实施例中机器翻译模型的解码器对源端向量序列进行解码的描述,在此不再赘述。

[0287] S1208,根据解码隐层向量,确定机器翻译模型当次输出的目标词。

[0288] 具体地,当机器翻译模型在生成当次的待输出的目标词时,可通过机器翻译模型的输出层,根据解码器解码得到的解码隐层向量,计算得到当次的输出概率序列。其中,机器翻译模型输出的当次的输出概率序列是指在输出端词集中各候选词为当次输出的目标词的概率所组成的序列。进一步地,机器翻译模型可选择输出概率序列中的最大概率所对

应的候选词作为当次的目标词。依次重复上述解码步骤,直到输出结束词为止。

[0289] 关于机器翻译模型的解码器根据解码隐层向量,确定机器翻译模型当次输出的目标词步骤的详细内容,可参考前述实施例中机器翻译模型的解码器生成当次输出的目标词的描述,在此不再赘述。

[0290] S1210,根据机器翻译模型输出的各目标词,生成待翻译文本对应的目标译文。

[0291] 具体地,计算机设备可通过机器翻译模型的输出层,按照输出各个目标词的顺序,将输出的各个目标词进行拼接,生成与待翻译对应的目标译文。在一个实施例中,目标译文与待翻译文本为不同语言的文本。

[0292] 在一个实施例中,通过前述模型训练方法训练得到的机器翻译模型,在执行翻译操作时,可只使用该机器翻译模型中编码器和解码器结构相关的模型参数,通过encoder-decoder(编码器-解码器)结构,对输入的待翻译文本进行处理,输出目标译文,也可以得到翻译质量高的目标译文。因为机器翻译模型中encoder-decoder结构的部分模型参数,在模型训练时已经学习到了一定省略词信息,因而在对待翻译文本进行编解码处理时,在一定程度上也可以提升机器翻译的翻译质量。

[0293] 上述实施例中,通过已训练的机器翻译模型中的通过编码-解码框架对待翻译文本进行翻译,可以学习到源文本的语义和句法信息,更充分利用机器翻译模型中各隐层表示,减少了在模型处理中有效信息的丢失,大大提高了文本翻译的准确率。

[0294] 如图13所示,在一个实施例中,提供了文本翻译装置1300,包括获取模块1301、编码模块1302、解码模块1303、重构模块1304、评估模块1305和筛选模块1306。

[0295] 获取模块1301,用于获取源文本的词序列;源文本为指代词省略文本。

[0296] 编码模块1302,用于对源文本的词序列进行语义编码,得到源端向量序列。

[0297] 解码模块1303,用于对源端向量序列进行解码处理,得到候选译文、各候选译文对应的目标端向量序列、及各候选译文对应的翻译得分。

[0298] 重构模块1304,用于将词序列和源端向量序列,与各目标端向量序列分别进行重构处理,得到与各候选译文分别对应的重构隐层向量序列;重构隐层向量序列包括源文本所缺少的指代词的词信息。

[0299] 评估模块1305,用于将源文本的词序列,与各重构隐层向量序列分别进行重构评估处理,得到与各候选译文分别对应的重构得分。

[0300] 筛选模块1306,用于从候选译文中筛选相应翻译得分和重构得分满足预设条件的目标译文。

[0301] 在一个实施例中,该文本翻译装置1300还包括融合模块1307,其中:

[0302] 获取模块1301还用于获取与源文本对应的篇章文本。

[0303] 编码模块1302还用于对篇章文本进行编码处理,得到篇章向量。

[0304] 融合模块1307,用于将篇章向量,与各重构隐层向量序列分别进行融合,得到与各候选译文分别对应的第一融合向量序列。

[0305] 评估模块1305还用于将源文本的词序列,与各第一融合向量序列分别进行重构评估处理,得到与各候选译文分别对应的重构得分。

[0306] 在一个实施例中,获取模块1301还用于获取与源文本对应的篇章文本。编码模块1302还用于对篇章文本进行编码处理,得到篇章向量。融合模块1307还用于将篇章向量,与

各目标端向量序列分别进行融合,得到与各候选译文分别对应的第二融合向量序列。重构模块1304还用于将词序列和源端向量序列,与各第二融合向量序列分别进行重构处理,得到与各候选译文分别对应的重构隐层向量序列。

[0307] 在一个实施例中,编码模块1302还用于确定篇章文本所包括的整句;对各整句分别进行分词处理,得到各整句各自对应的词序列;对各整句各自对应的词序列分别进行词级编码处理,得到与各整句分别对应的句子向量序列;将与各整句分别对应的句子向量序列进行句子级编码处理,得到与篇章文本对应的篇章向量。

[0308] 在一个实施例中,解码模块1303还用于依次对源端向量序列进行解码得到目标端向量,且每次解码时根据前次确定的候选词的词向量进行解码,并根据当次的目标端向量确定当次的多于一个的候选词;将依次解码得到的各候选词,按候选词的生成时序,拼接得到多于一组的候选译文;通过构成各候选译文的候选词所对应的目标端向量,确定与各候选译文分别对应的目标端向量序列;根据源端向量序列、历史解码得到的候选词、及历史解码得到的目标端向量序列,确定与各候选词对应的翻译得分;依据构成各候选译文的各候选词所对应的翻译得分,计算得到与各候选译文对应的翻译得分。

[0309] 在一个实施例中,该文本翻译方法还包括计算模块1308,其中:获取模块1301还用于获取与源端向量序列对应的源端注意力分配权重向量。计算模块1308,用于根据源端注意力分配权重向量、以及源端向量序列,计算得到源端内容向量。获取模块1301还用于分别获取与各目标端向量序列对应的目标端注意力分配权重向量。计算模块1308还用于根据目标端注意力分配权重向量、以及相对应的目标端向量序列,计算得到与各候选译文分别对应的目标端内容向量。重构模块1304还用于将词序列和源端内容向量,与各目标端内容向量分别进行重构处理,得到与各候选译文分别对应的重构隐层向量序列。

[0310] 在一个实施例中,重构模块1304还用于对于每组候选译文,分别根据源文本中当前词的前一词的词向量、前次重构得到的重构隐层向量、源端内容向量和各候选译文各自所对应的目标端内容向量,重构得到当次重构隐层向量;对于每组候选译文,分别将依次重构得到的重构隐层向量构成重构隐层向量序列;重构得到的重构隐层向量序列包括源文本所缺少的指代词的词信息。

[0311] 在一个实施例中,评估模块1305还用于对于每组候选译文,分别根据源文本中当前词的前一词的词向量、源端内容向量、以及各候选译文对应的当次的重构隐层向量和目标端内容向量,计算得到与各候选译文分别对应、且与源文本中当前词对应的重构得分;对于每组候选译文,分别对与源文本中各词所对应的重构得分进行连乘运算,得到与各候选译文分别对应的重构得分。

[0312] 在一个实施例中,筛选模块1306还用于对各候选译文分别对应的翻译得分和重构得分进行加权求和计算,得到各候选译文各自所对应的综合得分;从候选译文中,筛选相应综合得分为最高得分的候选译文;将筛选出的候选译文作为目标译文并输出。

[0313] 参考图14,在一个实施例中,文本翻译方法还包括训练模块1309,用于获取样本源文本、参考目标译文、以及与源文本对应的标注词序列;标注词序列是对样本源文本进行缺略词标注处理所得到的词序列;将样本源文本输入机器翻译模型中进行训练,得到预测目标译文;依据参考目标译文和预测目标译文构建最大似然函数;通过机器翻译模型,获取与样本源文本对应的源端样本向量序列、以及目标端样本向量序列;依据样本源文本,对源端

样本向量序列、以及目标端样本向量序列进行重构处理,得到样本重构隐层向量序列;依据样本源文本、样本重构隐层向量序列、源端样本向量序列、以及目标端样本向量序列构建重构评估函数;依据标注词序列和样本重构隐层向量序列,构建词序列标注损失函数;根据最大似然函数、重构评估函数和词序列标注损失函数,确定目标函数;将目标函数最大化时的模型参数作为机器翻译模型的模型参数,返回将样本源文本输入机器翻译模型中进行训练,得到预测目标译文的步骤并继续训练,直至满足训练停止条件时停止训练。

[0314] 在一个实施例中,训练模块1309还用于将最大似然函数、重构评估函数和词序列标注损失函数的加权和函数作为机器翻译模型的目标函数;其中,通过以下公式表示目标函数:

$$[0315] \quad J(\theta, \gamma, \psi) = \arg \max_{\theta, \gamma, \psi} \sum_{n=1}^N \{ \log L(y^n | x^n; \theta) + \lambda \log R(x | h^{enc}, h^{dec}; \theta, \gamma) + \tau \log P(dp | h^{rec}; \theta, \gamma, \psi) \}$$

$$[0316] \quad \text{其中, } R(x | h^{enc}, h^{dec}) = \prod_{i=1}^I g_r(x_{i-1}, h_i^{rec}, c_i^{enc}, c_i^{dec}); \quad h_i^{rec} = f_r(x_{i-1}, h_{i-1}^{rec}, c_i^{enc}, c_i^{dec});$$

$$[0317] \quad c_i = \sum_{j=1}^J \alpha_{i,j} \bullet h_j; \quad P(dp | h^{rec}) = \prod_{i=1}^I P(dp_i | h_i^{rec}) = \prod_{i=1}^I g_l(dp_i | h_i^{rec})$$

[0318] 其中,其中, θ 、 γ 和 τ 分别是模型参数; $\{[x^n, y^n]\}_{n=1}^N$ 是模型训练中的样本平行语料对; $g_r()$ 和 $f_r()$ 分别为归一化函数和激活函数; $g_l()$ 是归一化函数; x_{i-1} 表示样本源文本中的第*i*-1个词的词向量; h_i^{rec} 表示与样本源文本中的第*i*个词对应的样本重构隐层向量; h_j 表示各源端样本向量或目标端样本向量; $\alpha_{i,j}$ 表示源端注意力分配权重或目标端注意力分配权重; $\bullet$ 表示向量点乘运算; $dp = \{dp_1, dp_2, \dots, dp_I\}$ 是样本标注词序列。

[0319] 在一个实施例中,训练模块1309还用于获取与样本源文本对应的样本篇章文本;对样本篇章文本进行编码处理,得到样本篇章向量;融合样本篇章向量和样本重构隐层向量序列,得到样本融合向量序列;依据样本源文本、样本融合向量序列、源端样本向量序列、以及目标端样本向量序列构建重构评估函数;依据标注词序列和样本融合向量序列,构建词序列标注损失函数。

[0320] 在一个实施例中,训练模块1309还用于通过以下公式融合样本篇章向量和样本重构隐层向量序列: $\hat{h}_i^{rec} = f_c(h_i^{rec}, M)$;其中, $\hat{h}_i^{rec}$ 表示样本源文本中的第*i*个词对应的样本融合向量; $f_c()$ 表示串联操作函数; M 表示样本篇章向量; h_i^{rec} 表示与样本源文本中的第*i*个词对应的样本重构隐层向量。

[0321] 上述文本翻译装置,对缺少指代词的源文本进行语义编码,得到源端向量序列,并继续对源端向量序列进行解码,得到候选译文、各候选译文对应的目标端向量序列、及各候选译文对应的翻译得分,这样可充分提取出源端和目标端的隐层信息。再通过融合源端和目标端的隐层信息可以共享两端的共性信息,在根据源文本进行重构处理时可以减少冗余信息并能避免重要信息的缺失,不需要额外的省略词标注信息就可重构回包括缺少的指代词的词信息的重构隐层向量序列,这样使得确定的重构得分可以很好的衡量缺少的指代词

信息的召回程度。因此,依据重构得分和翻译得分所筛选出的目标译文,可以减少翻译时缺少省略代词所带来的影响,大大提高了翻译质量。

[0322] 如图15所示,在一个实施例中,提供了模型训练装置1500,包括样本获取模块1501、输入模块1502、构建模块1503、确定模块1504和重构模块1505。

[0323] 样本获取模块1501,用于获取样本源文本、参考目标译文、以及与源文本对应的标注词序列;标注词序列是对样本源文本进行省略词标注处理所得到的词序列。

[0324] 输入模块1502,用于将样本源文本输入机器翻译模型中进行训练,得到预测目标译文。

[0325] 构建模块1503,用于依据参考目标译文和预测目标译文构建最大似然函数。

[0326] 确定模块1504,用于通过机器翻译模型,获取与样本源文本对应的源端样本向量序列、以及目标端样本向量序列。

[0327] 重构模块1505,用于依据样本源文本,对源端样本向量序列、以及目标端样本向量序列进行重构处理,得到样本重构隐层向量序列。

[0328] 构建模块1503还用于依据样本源文本、样本重构隐层向量序列、源端样本向量序列、以及目标端样本向量序列构建重构评估函数。

[0329] 构建模块1503还用于依据标注词序列和样本重构隐层向量序列,构建词序列标注损失函数。

[0330] 确定模块1505还用于根据最大似然函数、重构评估函数和词序列标注损失函数,确定目标函数。

[0331] 确定模块1505还用于将目标函数最大化时的模型参数作为机器翻译模型的模型参数,返回将样本源文本输入机器翻译模型中进行训练,得到预测目标译文的步骤并继续训练,直至满足训练停止条件时停止训练。

[0332] 在一个实施例中,该模型训练装置1500还包括编码模块1506和融合模块1507,其中:

[0333] 样本获取模块1501还用于获取与样本源文本对应的样本篇章文本。

[0334] 编码模块1506,用于对样本篇章文本进行编码处理,得到样本篇章向量。

[0335] 融合模块1507,用于融合样本篇章向量和样本重构隐层向量序列,得到样本融合向量序列。

[0336] 构建模块1503还用于依据样本源文本、样本融合向量序列、源端样本向量序列、以及目标端样本向量序列构建重构评估函数。

[0337] 构建模块1503还用于依据标注词序列和样本融合向量序列,构建词序列标注损失函数。

[0338] 参考图16,在一个实施例中,该模型训练装置1500还包括使用模块1508,用于获取待翻译文本的词序列;通过已训练的机器翻译模型的编码器,对待翻译文本的词序列进行语义编码处理,得到语义向量序列;通过已训练的机器翻译模型的解码器,根据机器翻译模型前次输出的目标词,对语义向量序列进行解码,得到当次的解码隐层向量;根据解码隐层向量,确定机器翻译模型当次输出的目标词;根据机器翻译模型输出的各目标词,生成待翻译文本对应的目标译文。

[0339] 上述模型训练装置,在模型训练过程中一方面将源端样本向量序列和目标端样本

向量序列重构回样本源文本,另一方面依据标注词序列对重构得到的隐层向量序列进行缺略词标注。训练时,最大似然函数、重构评估函数和词序列标注损失函数叠加一起优化,使得翻译模型不仅可以学习到缺略词的位置信息,更具备还原缺略词的能力,使得训练得到的机器翻译模型具有更高的翻译准确性。

[0340] 图17示出了一个实施例中计算机设备的内部结构图。该计算机设备具体可以是图1中的终端110或服务器120。如图17所示,该计算机设备包括该计算机设备包括通过系统总线连接的处理器、存储器、和网络接口。其中,存储器包括非易失性存储介质和内存存储器。该计算机设备的非易失性存储介质存储有操作系统,还可存储有计算机程序,该计算机程序被处理器执行时,可使得处理器实现文本翻译方法和/或模型训练方法。该内存存储器中也可储存有计算机程序,该计算机程序被处理器执行时,可使得处理器执行文本翻译方法和/或模型训练方法。

[0341] 本领域技术人员可以理解,图17中示出的结构,仅仅是与本申请方案相关的部分结构的框图,并不构成对本申请方案所应用于其上的计算机设备的限定,具体的计算机设备可以包括比图中所示更多或更少的部件,或者组合某些部件,或者具有不同的部件布置。

[0342] 在一个实施例中,本申请提供的文本翻译装置可以实现为一种计算机程序的形式,计算机程序可在如图17所示的计算机设备上运行。计算机设备的存储器中可存储组成该文本翻译装置的各个程序模块,比如,图13所示的获取模块、编码模块、解码模块、重构模块、评估模块、和筛选模块。各个程序模块构成的计算机程序使得处理器执行本说明书中描述的本申请各个实施例的文本翻译方法中的步骤。

[0343] 例如,图17所示的计算机设备可以通过如图13所示的文本翻译装置中的获取模块执行步骤S202。计算机设备可通过编码模块执行步骤S204。计算机设备可通过解码模块执行步骤S206。重构模块执行步骤S208。计算机设备可通过评估模块执行步骤S210。计算机设备可通过筛选模块执行步骤S212。

[0344] 在一个实施例中,本申请提供的模型训练装置可以实现为一种计算机程序的形式,计算机程序可在如图17所示的计算机设备上运行。计算机设备的存储器中可存储组成该模型训练装置的各个程序模块,比如,图15所示的样本获取模块、输入模块、构建模块、确定模块和重构模块。各个程序模块构成的计算机程序使得处理器执行本说明书中描述的本申请各个实施例的模型训练方法中的步骤。例如,图17所示的计算机设备可以通过如图15所示的模型训练装置中的各个模块执行该模型训练方法中的各个步骤

[0345] 在一个实施例中,提供了一种计算机设备,包括存储器和处理器,存储器存储有计算机程序,计算机程序被处理器执行时,使得处理器执行上述文本翻译方法和/或模型训练方法的步骤。此处文本翻译方法和/或模型训练方法的步骤可以是上述各个实施例的文本翻译方法和/或模型训练方法中的步骤。

[0346] 在一个实施例中,提供了一种计算机可读存储介质,存储有计算机程序,计算机程序被处理器执行时,使得处理器执行上述文本翻译方法和/或模型训练方法的步骤。此处文本翻译方法和/或模型训练方法的步骤可以是上述各个实施例的XX方法中的步骤。

[0347] 本领域普通技术人员可以理解实现上述实施例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,所述的程序可存储于一非易失性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,本申请所提供

的各实施例中所使用的对存储器、存储、数据库或其它介质的任何引用,均可包括非易失性和/或易失性存储器。非易失性存储器可包括只读存储器 (ROM)、可编程ROM (PROM)、电可编程ROM (EPROM)、电可擦除可编程ROM (EEPROM) 或闪存。易失性存储器可包括随机存取存储器 (RAM) 或者外部高速缓冲存储器。作为说明而非局限,RAM以多种形式可得,诸如静态RAM (SRAM)、动态RAM (DRAM)、同步DRAM (SDRAM)、双数据率SDRAM (DDRSDRAM)、增强型SDRAM (ESDRAM)、同步链路 (Synchlink) DRAM (SLDRAM)、存储器总线 (Rambus) 直接RAM (RDRAM)、直接存储器总线动态RAM (DRDRAM)、以及存储器总线动态RAM (RDRAM) 等。

[0348] 以上实施例的各技术特征可以进行任意的组合,为使描述简洁,未对上述实施例中的各个技术特征所有可能的组合都进行描述,然而,只要这些技术特征的组合不存在矛盾,都应当认为是本说明书记载的范围。

[0349] 以上所述实施例仅表达了本申请的几种实施方式,其描述较为具体和详细,但并不能因此而理解为对本申请专利范围的限制。应当指出的是,对于本领域的普通技术人员来说,在不脱离本申请构思的前提下,还可以做出若干变形和改进,这些都属于本申请的保护范围。因此,本申请专利的保护范围应以所附权利要求为准。

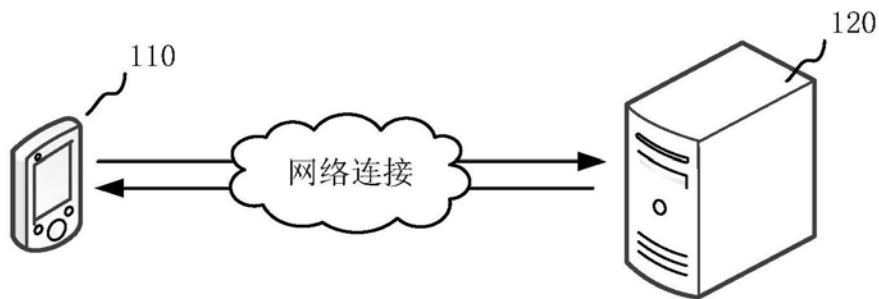


图1

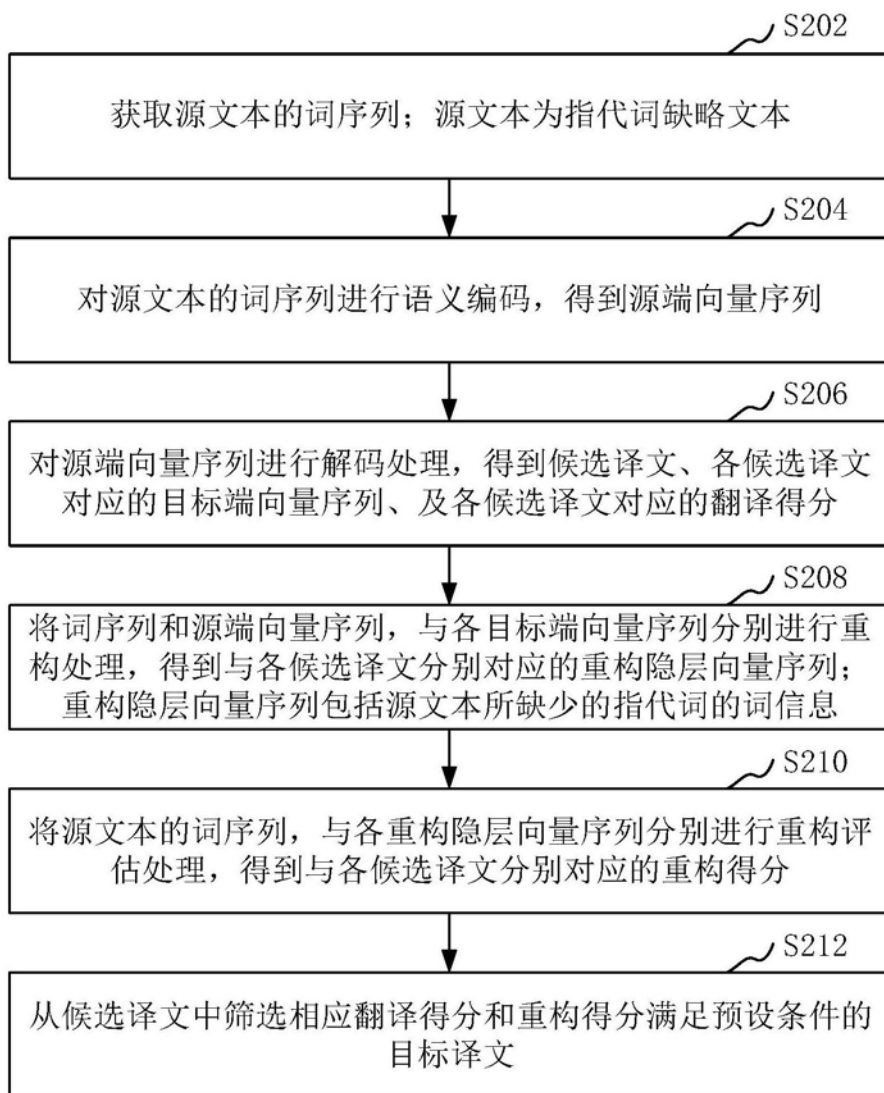


图2

- 1 (a) (你) 喜欢 这块 蛋糕 吗 ?
1 (b) Do you like this cake ?
- 2 (a) (我) 很喜欢 (它) 。你 烤的 (它) 吗 ?
2 (b) I like it . Did you bake it ?

图3

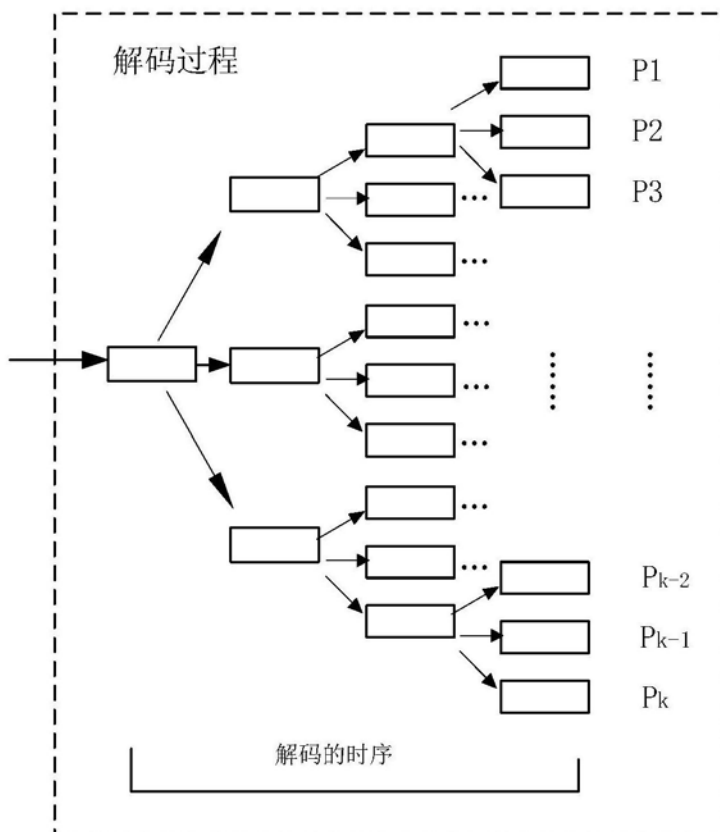


图4

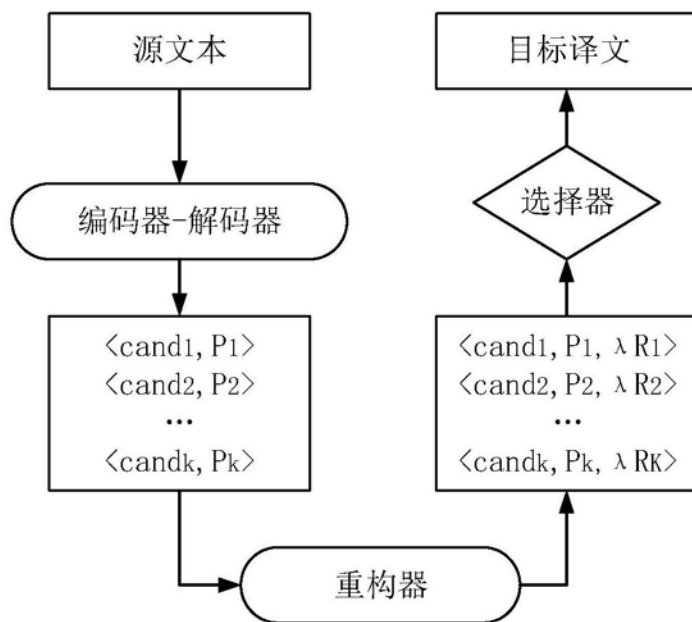


图5

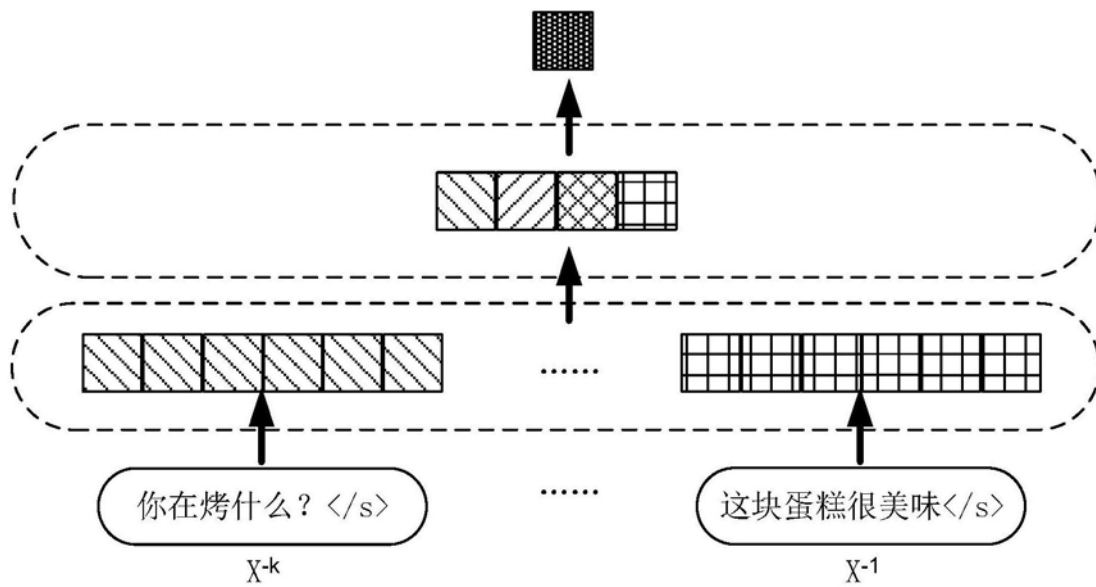


图6

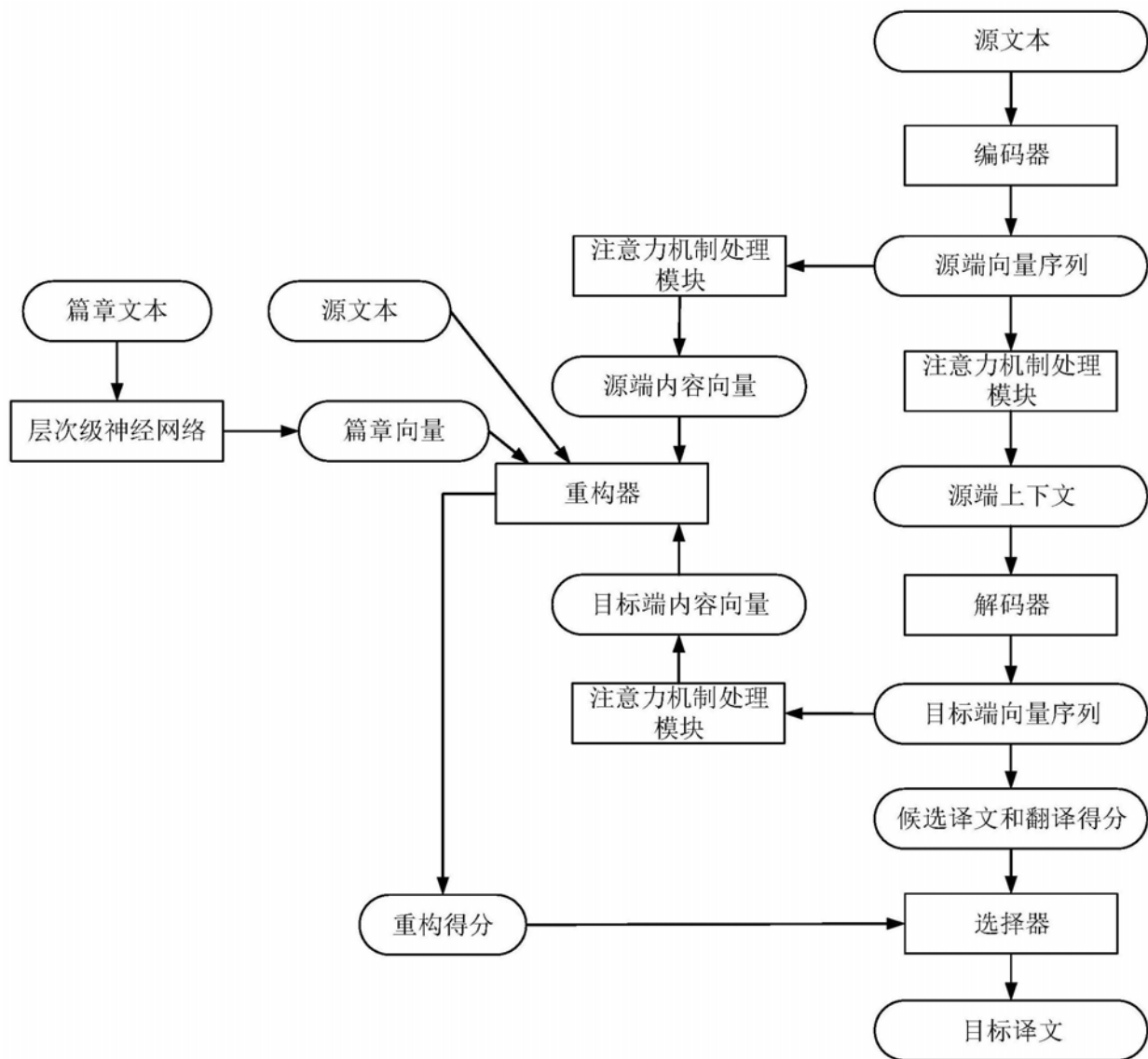


图7

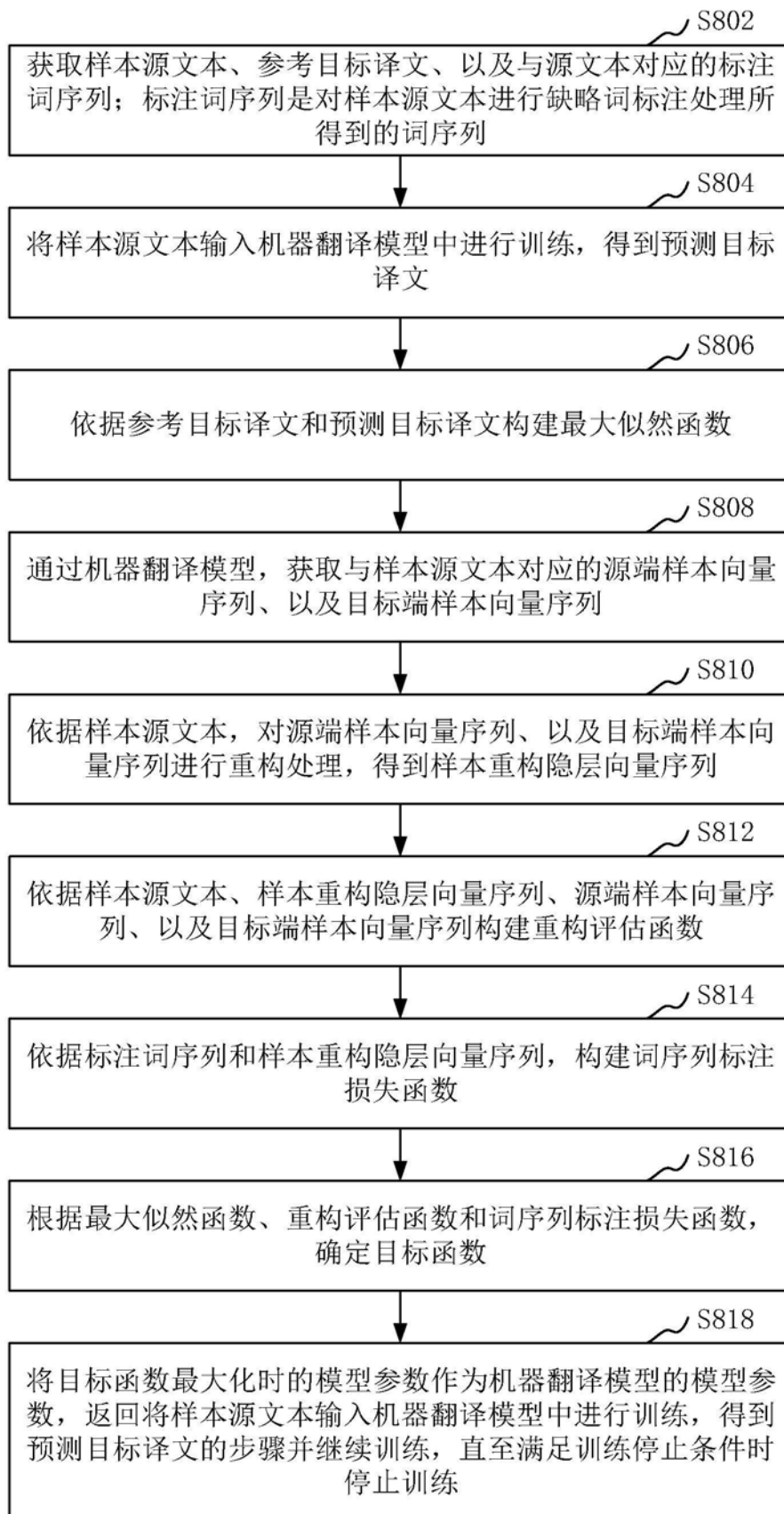


图8

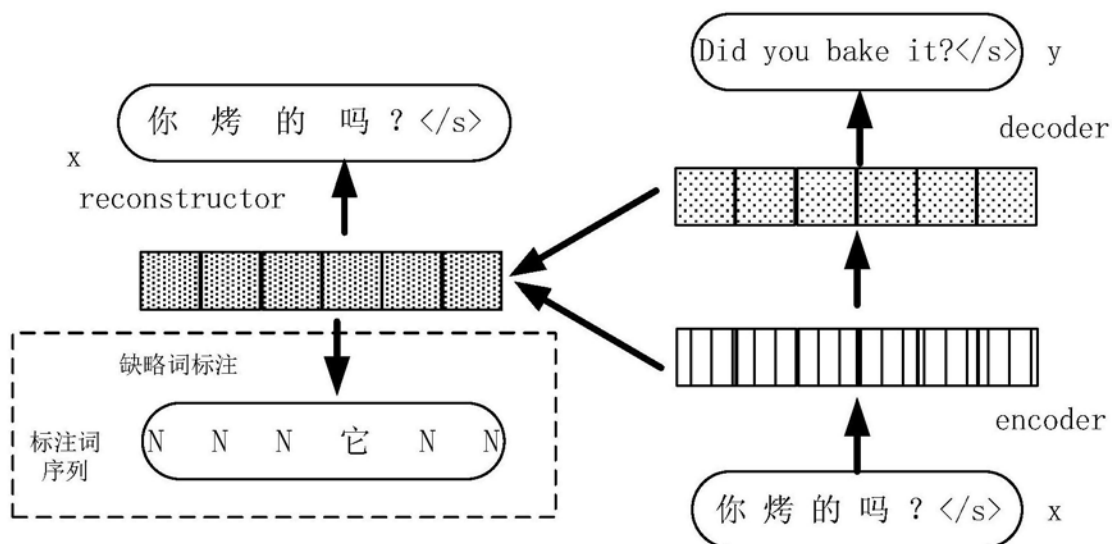


图9

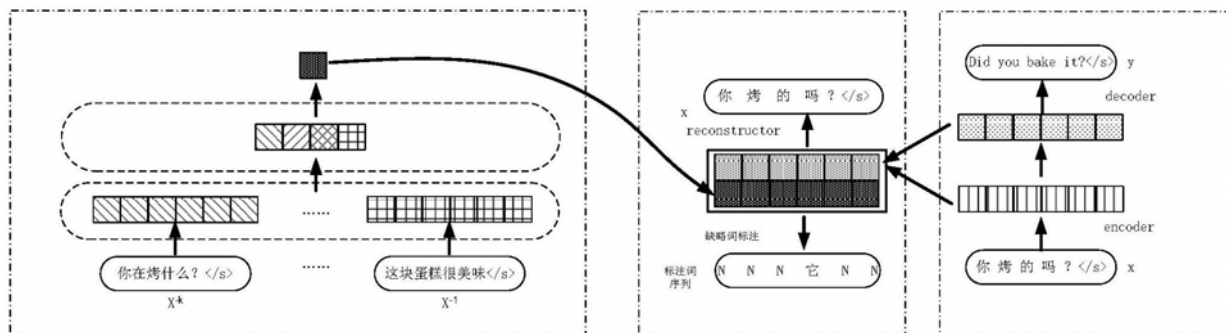


图10

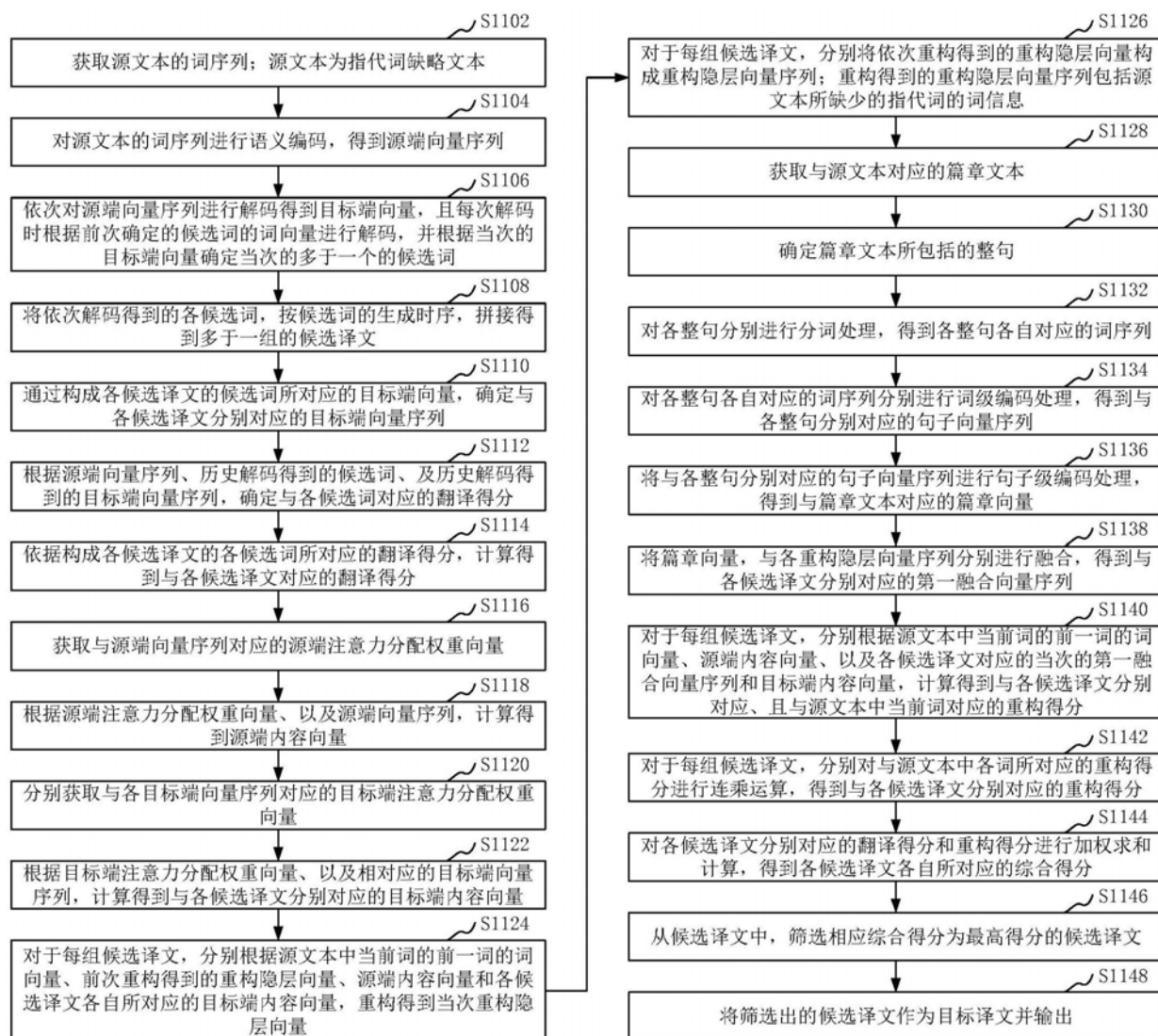


图11

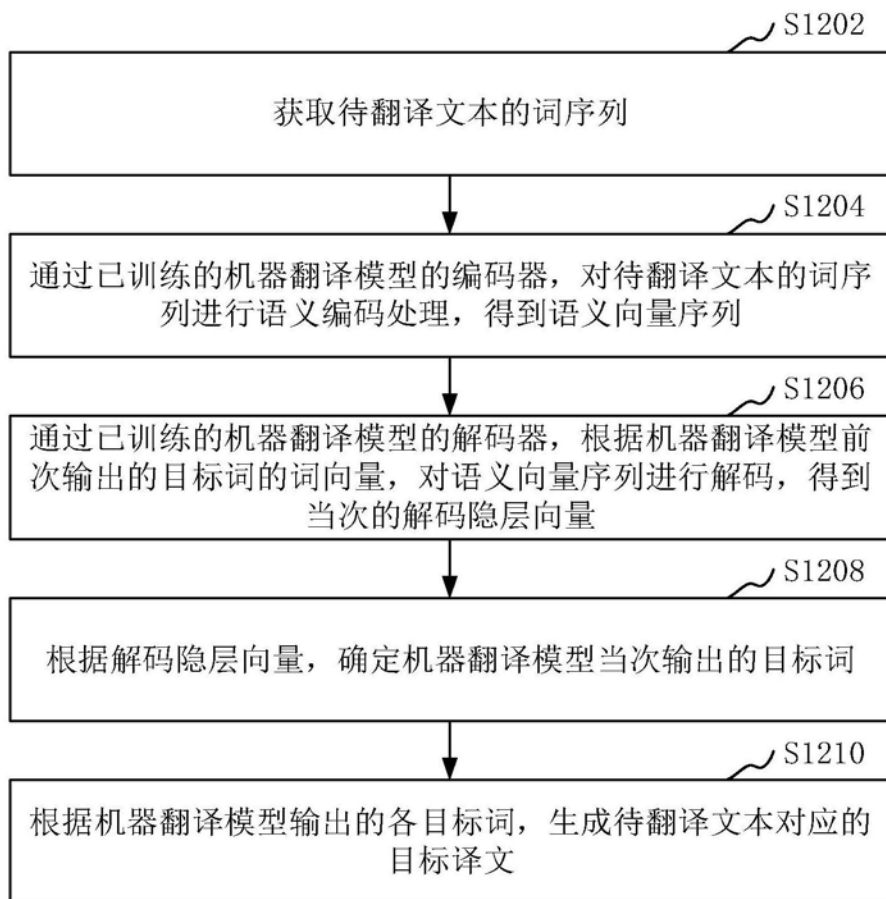


图12

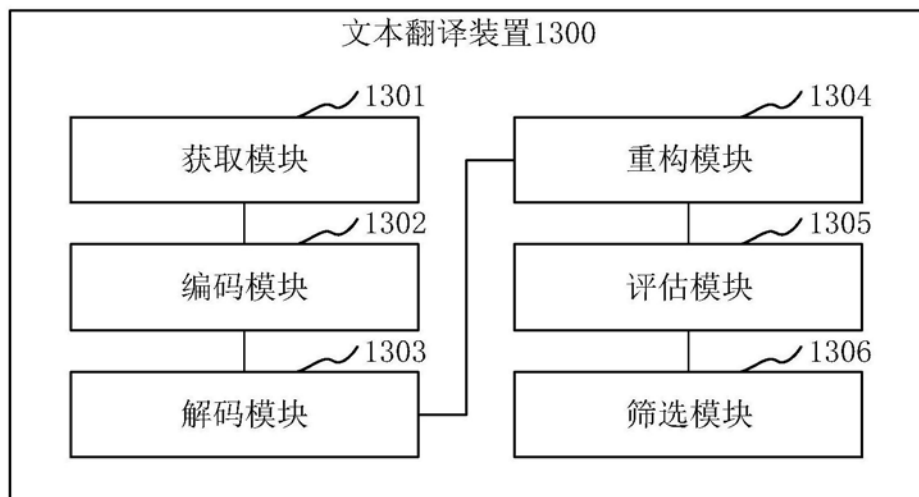


图13

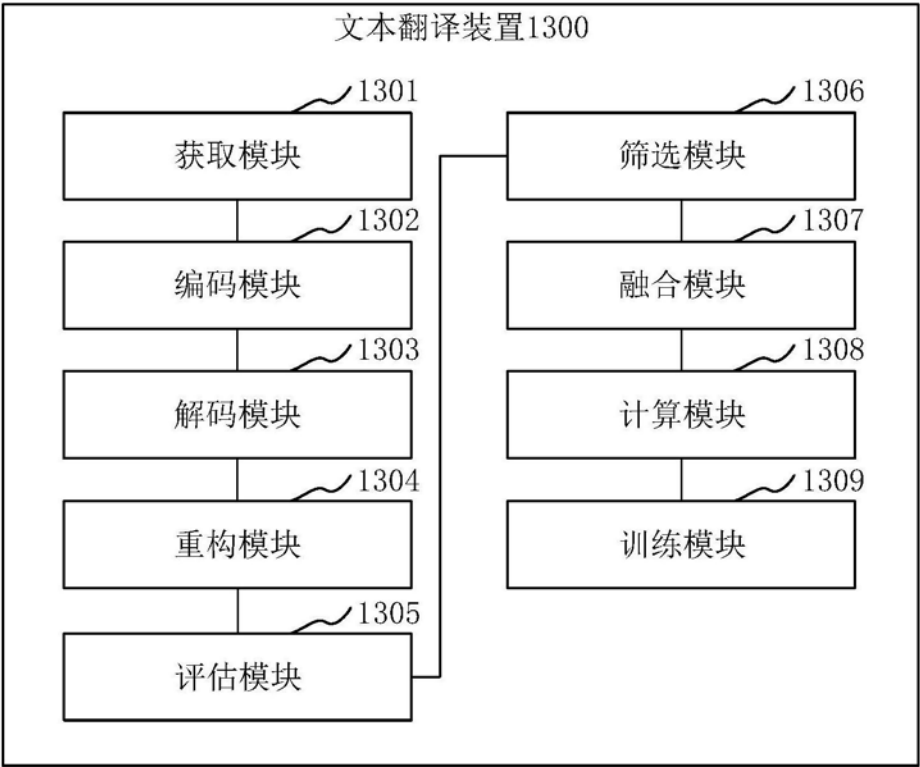


图14

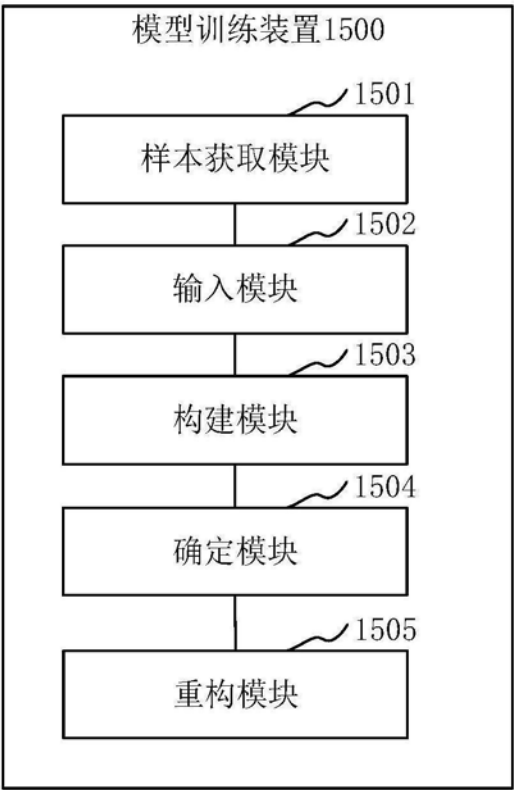


图15

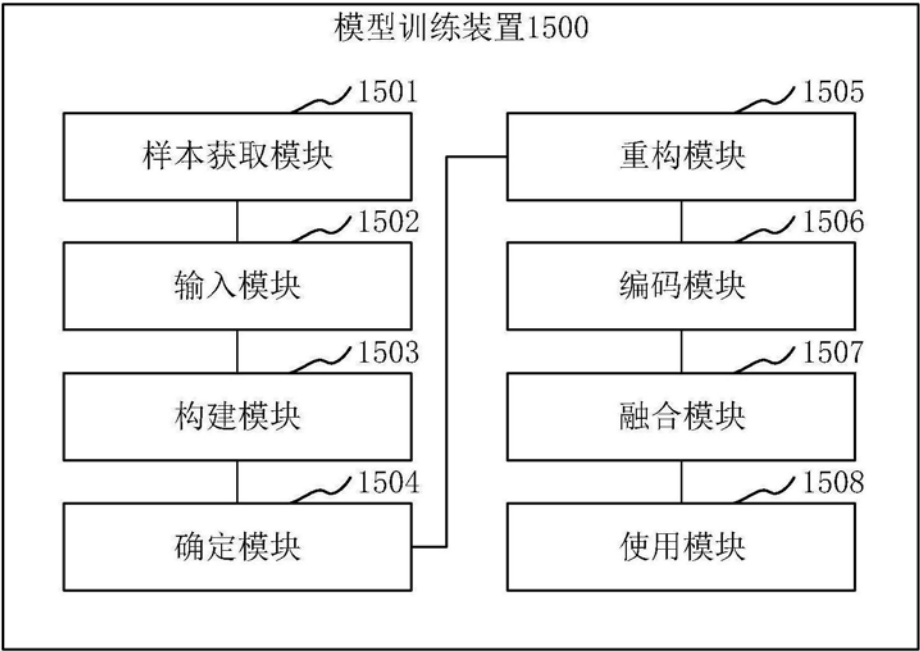


图16

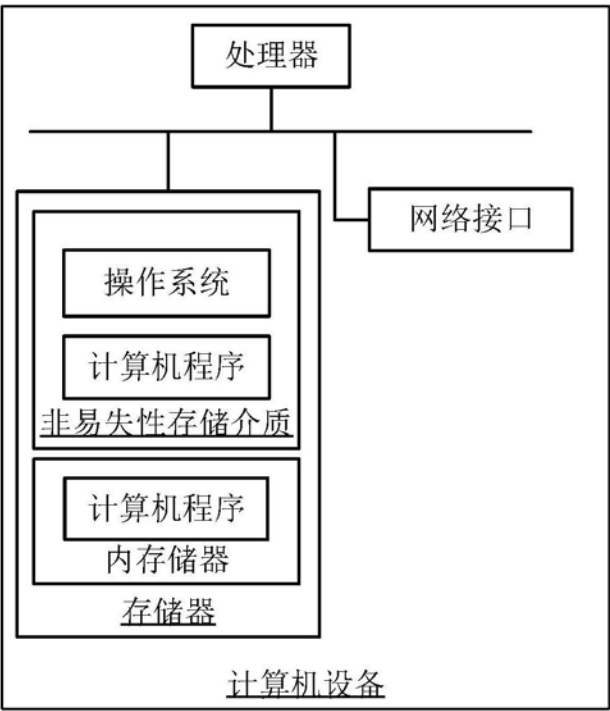


图17