(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 112541343 B
(45)授权公告日 2022.06.14

(21)申请号 202011408948.2

(22)申请日 2020.12.03

(65)同一申请的已公布的文献号
申请公布号 CN 112541343 A

(43)申请公布日 2021.03.23

(73)专利权人 昆明理工大学
地址 650093 云南省昆明市五华区学府路253号

(72)发明人 余正涛 张莹 黄于欣 高盛祥 郭军军 相艳

(74)专利代理机构 昆明人从众知识产权代理有限公司 53204
专利代理师 何娇

(51)Int.Cl.
G06F 40/242(2020.01)
G06F 40/30(2020.01)
G06F 40/44(2020.01)
G06F 16/34(2019.01)

(56)对比文件
CN 110378409 A,2019.10.25
CN 111930930 A,2020.11.13
CN 109614480 A,2019.04.12
CN 111310480 A,2020.06.19
CN 110196903 A,2019.09.03
CN 108733682 A,2018.11.02
US 2020311122 A1,2020.10.01
CN 111639175 A,2020.09.08
王 剑 等.使用词对齐半监督对抗学习的汉越跨语言摘要生成方法.《https://kns.cnki.net/kcms/detail/21.1106.TP.20210516.1345.002.html》.2021,
Yang Liu 等.Text summarization with pretrained encoders.《https://arxiv.org/abs/1908.08345》.2019,

审查员 王悦

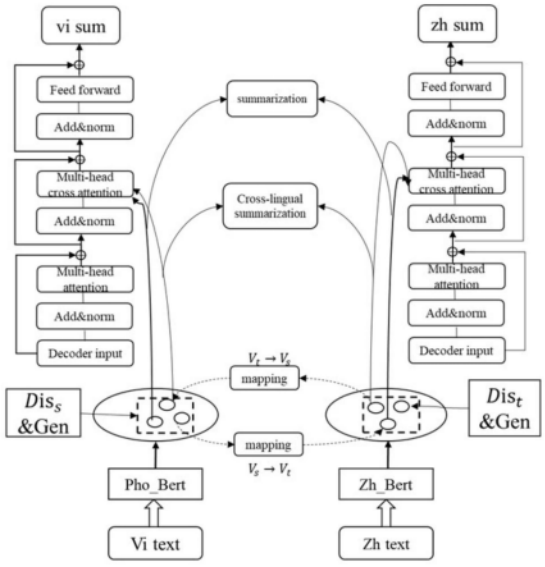
权利要求书2页 说明书8页 附图1页

(54)发明名称

基于词对齐的半监督对抗学习跨语言摘要生成方法

(57)摘要

本发明涉及基于词对齐的半监督对抗学习跨语言摘要生成方法,属于自然语言处理技术领域。本发明包括步骤:收集用于训练汉越跨语言摘要生成的新闻文本,以及获取已有的汉越双语词向量;利用汉越新闻文本和汉越双语词向量分别进行单语摘要模型和半监督对抗学习的预训练;利用Bert编码器分别对输入的汉越伪平行语料进行向量表征;利用编码器获得的向量结合汉越双语种子词典进行半监督对抗学习,获取映射到同一语义空间的向量;把映射在同一语义空间下的上下文文本向量和参考摘要作为transformer解码器的输入,解码输出目标语言摘要。本发明实现了跨语言摘要生成任务,优化了跨语言摘要的效果。



CN 112541343 B

1.基于词对齐的半监督对抗学习跨语言摘要生成方法,其特征在于,所述方法包括:

Step1、收集用于训练汉越跨语言摘要生成的新闻文本,以及获取已有的汉越双语词向量;

Step2、利用汉越新闻文本和汉越双语词向量分别进行单语摘要模型和半监督对抗学习的预训练;

Step3、利用Bert编码器分别对输入的汉越伪平行语料进行向量表征;

Step4、利用编码器获得的向量结合汉越双语种子词典进行半监督对抗学习,获取映射到同一语义空间的向量;

Step5、把映射在同一语义空间下的上下文文本向量和参考摘要作为transformer解码器的输入,解码输出目标语言摘要;

所述步骤Step2的具体步骤为:

Step2.1、首先使用一定数量的越南语和中文新闻文本语料,分别进行两种语言的单语摘要模型训练,摘要生成模型均采用Bert摘要模型,经过单语预训练得到的摘要模型;

Step2.2、利用维基百科提供的汉越双语词向量,进行半监督对抗学习方法如下:

(1)、分别得到维基百科提供的越南语和中文的词向量集:

$S = \{S_1^d, S_2^d, \cdots, S_m^d\}, T = \{T_1^d, T_2^d, \cdots, T_n^d\}$,其中m和n表示词表大小,d表示嵌入向量的维度,S为越南语词向量集,T为中文词向量集;

(2)、预训练阶段,首先利用包含L=30,000的$\{s_i, t_i\}_{i\in(1,2,\cdots,L)}$汉越双语种子词典,训练映射矩阵W:

$\Omega = ||Ws-t||^2$

其中,s为双语词典的源语言,t为对应的目标语言,Ω表示正则器,用于强制表达式两边的相等性,使用随机梯度下降学习W,然后通过最小化经过W转换的源语言单词s_i的向量表征与双语词典中的目标语言t_i之间的平方欧氏距离,来实现双语词向量在同一语义空间下的映射对齐;假设源语言词s为越南语,目标语言t为中文,则由越南语映射到中文的映射矩阵为$W^{s\to t}$,同理可得中文映射到越南文的矩阵为$W^{t\to s}$;

(3)联合训练阶段,用Bert编码器得到越南语和中文的词向量s_i^{vi}和t_i^{zh}来训练学习的映射矩阵W_s^*,W_T^*:

$$W_s^* = \underset{W \in M_d(R)}{argmin} \left\| W^{s\to t} s_i^{vi} - t_i^{zh} \right\|^2$$

$$W_T^* = \underset{W \in M_d(R)}{argmin} \left\| W^{t\to s} t_i^{zh} - s_i^{vi} \right\|^2$$

其中d是嵌入向量的维度,M_d(尺)是一个值为实数的d*d矩阵,$W^{s\to t}$和$W^{t\to s}$表示映射方向,目标是找到最佳映射矩阵W_s^*,以使映射源嵌入$W^{s\to t} s_i^{vi}$和目标语言嵌入t_i^{zh}之间的平方欧几里得距离最小,矩阵W_T^*同理;

(4)最后,将生成器生成的映射后的越南语向量和中文向量同时提交给判别器来预测每个单词的来源,来优化判别器Dis_s和Gen_s:

$$\underset{Gen_s}{min}\ \underset{Dis_s}{max} V(D,G) = \underset{G}{min}\ \underset{D}{max} \left[\log(D_s(t_i^{zh}))\right] + \left[\log\left(1 - D_s\left(W^{s\to t} s_i^{vi}\right)\right)\right]$$

对于判别器Dis_s而言，看作是一个二分类问题，即一个形式为D_s:s→{0,1}的函数，真实样本对应为1，映射样本对应为0，V(D,G)为二分类问题中常见的交叉熵损失，对于生成器Gen_s而言，为了尽可能欺骗判别器，所以需要最大化目标语言向量的判别概率$D_s(t_i^{z\hbar})$，即最小化$\log\left(1-D_s\left(W^{s\to t}s_i^{vi}\right)\right)$；

判别器Dist和生成器Gen_t同理可得：

$$\min_{Gen_t}\max_{Dis_t}V(D,G)=\min_{G}\max_{D}\left[\log(D_t(s_i^{vi}))\right]+\left[\log\left(1-D_t\left(W^{t\to s}t_i^{z\hbar}\right)\right)\right]$$

训练时，生成器和判别器采取交替训练，即先训练Dis_s和Dis_t，然后训练Gen_s和Gen_t，不断往复，直到判别器无法识别词的来源则表示双语词向量位于一个公共语义空间。

2.根据权利要求1所述的基于词对齐的半监督对抗学习跨语言摘要生成方法，其特征在于：所述Step1中，使用Scrapy作为爬取工具，模仿用户操作，为汉语、越南语新闻网站定制不同的模板，根据页面数据元素的XPath路径制定模板获取详细数据，获取新闻标题、新闻时间、新闻正文数据，以及从维基百科获取已经训练好的汉越双语词向量。

3.根据权利要求1所述的基于词对齐的半监督对抗学习跨语言摘要生成方法，其特征在于：所述步骤Step2中：利用汉语和越南语新闻文本进行单语摘要模型训练会得到汉语和越南语两种摘要生成模型；利用汉越双语词向量和双语词典进行对抗学习得到实现将源语言映射到目标语言空间的映射矩阵W。

4.根据权利要求1所述的基于词对齐的半监督对抗学习跨语言摘要生成方法，其特征在于：所述步骤Step3中：将汉越新闻伪平行语料作为摘要模型的输入，分别得到由Bert编码器生成的上下文文本向量。

5.根据权利要求1所述的基于词对齐的半监督对抗学习跨语言摘要生成方法，其特征在于：所述步骤Step4的具体步骤：

Step4.1、从汉越伪平行语料的参考摘要数据源构建一个汉越种子词典；

Step4.2、对Bert编码器生成的上下文文本向量表征应用映射矩阵W和种子词典进行半监督对抗学习得到汉越双语在同一语义空间下的对齐向量。

6.根据权利要求1所述的基于词对齐的半监督对抗学习跨语言摘要生成方法，其特征在于：所述步骤Step5的具体步骤为：

Step5.1、将经过对抗性学习训练后在同一语义空间下的对齐向量和参考摘要作为transformer解码器的输入；

Step5.2、解码端根据映射后的对齐向量与参考摘要之间的对数似然率来选取候选摘要；

Step5.3、解码器解码输出目标语言摘要。

基于词对齐的半监督对抗学习跨语言摘要生成方法

技术领域

[0001] 本发明涉及基于词对齐的半监督对抗学习跨语言摘要生成方法，属于自然语言处理技术领域。

背景技术

[0002] 跨语言摘要生成是当前自然语言处理研究的热点问题。中越两国共同关注的问题日益增多，相关新闻报道也随之增多，利用跨语言摘要方法获取越南语新闻的文本摘要信息，对及时的了解两国针对重要事件发表的看法，促进两国共同发展具有重要意义。目前针对小语种的翻译技术尚未成熟，并且不同语言文本很难表示在同一特征空间下，获取跨语言新闻文本的摘要比较困难。因此，利用人工智能技术自动生成汉越双语新闻文本的摘要具有重要意义。

发明内容

[0003] 本发明提供了基于词对齐的半监督对抗学习跨语言摘要生成方法，以用于解决同语言文本很难表示在同一特征空间下，如何利用同一空间下的文本表征进行跨语言摘要任务等问题，以及解决了实现跨语言摘要必须借助翻译，而翻译效果欠佳的问题。

[0004] 本发明的技术方案是：基于词对齐的半监督对抗学习跨语言摘要生成方法，所述方法包括：

[0005] Step1、收集用于训练汉越跨语言摘要生成的新闻文本，以及获取已有的汉越双语词向量；

[0006] Step2、利用汉越新闻文本和汉越双语词向量分别进行单语摘要模型和半监督对抗学习的预训练；

[0007] Step3、利用Bert编码器分别对输入的汉越伪平行语料进行向量表征；

[0008] Step4、利用编码器获得的向量结合汉越双语种子词典进行半监督对抗学习，获取映射到同一语义空间的向量；

[0009] Step5、把映射在同一语义空间下的上下文文本向量和参考摘要作为transformer解码器的输入，解码输出目标语言摘要。

[0010] 作为本发明的进一步方案，所述Step1中，使用Scrapy作为爬取工具，模仿用户操作，为汉语、越南语新闻网站定制不同的模板，根据页面数据元素的XPath路径制定模板获取详细数据，获取新闻标题、新闻时间、新闻正文数据，以及从维基百科获取已经训练好的汉越双语词向量。

[0011] 作为本发明的进一步方案，所述步骤Step2的具体步骤为：

[0012] Step2.1、首先使用一定数量的越南语和中文新闻文本语料，分别进行两种语言的单语摘要模型训练，摘要生成模型均采用Bert摘要模型，经过单语预训练得到的摘要模型；

[0013] Step2.2、利用维基百科提供的汉越双语词向量，进行半监督对抗学习方法如下：

[0014] （1）、分别得到维基百科提供的越南语和中文的词向量集：

$S = \{S_1^d, S_2^d, \cdots, S_m^d\}, T = \{T_1^d, T_2^d, \cdots, T_n^d\}$,其中m和n表示词表大小,d表示嵌入向量的维度,S为越南语词向量集,T为中文词向量集;

[0015] (2)、预训练阶段,首先利用包含L=30,000的$\{s_i, t_i\}_{i\in(1,2,\cdots,L)}$汉越双语种子词典,训练映射矩阵W:

[0016] $\Omega = \|Ws - t\|^2$

[0017] 其中,s为双语词典的源语言,t为对应的目标语言,Ω表示正则器,用于强制表达式两边的相等性,使用随机梯度下降学习W,然后通过最小化经过W转换的源语言单词s_i的向量表征与双语词典中的目标语言t_i之间的平方欧氏距离,来实现双语词向量在同一语义空间下的映射对齐;假设源语言词s为越南语,目标语言t为中文,则由越南语映射到中文的映射矩阵为$W^{s\to t}$,同理可得中文映射到越南文的矩阵为$W^{t\to s}$;

[0018] (3)联合训练阶段,用Bert编码器得到越南语和中文的词向量s_i^{vi}和t_i^{zh}来训练学习的映射矩阵W_s^*, W_T^*:

[0019] $$W_s^* = \underset{W \in M_d(R)}{argmin} \left\|W^{s\to t} s_i^{vi} - t_i^{zh}\right\|^2$$

[0020] $$W_T^* = \underset{W \in M_d(R)}{argmin} \left\|W^{t\to s} t_i^{zh} - s_i^{vi}\right\|^2$$

[0021] 其中d是嵌入向量的维度,$M_d(R)$是一个值为实数的d*d矩阵,$W^{s\to t}$和$W^{t\to s}$表示映射方向,目标是找到最佳映射矩阵W_s^*,以使映射源嵌入$W^{s\to t} s_i^{vi}$和目标语言嵌入t_i^{zh}之间的平方欧几里得距离最小,矩阵W_T^*同理;

[0022] (4)最后,将生成器生成的映射后的越南语向量和中文向量同时提交给判别器来预测每个单词的来源,来优化判别器Dis_s和Gen_s:

[0023] $$\underset{Gen_s}{min}\ \underset{Dis_s}{max} V(D,G) = \underset{G}{min}\ \underset{D}{max} \left[\log(D_s(t_i^{zh}))\right] + \left[\log\left(1 - D_s\left(W^{s\to t} s_i^{vi}\right)\right)\right]$$

[0024] 对于判别器Dis_s而言,看作是一个二分类问题,即一个形式为$D_s: s\to\{0,1\}$的函数,真实样本对应为1,映射样本对应为0,V(D,G)为二分类问题中常见的交叉熵损失,对于生成器Gen_s而言,为了尽可能欺骗判别器,所以需要最大化目标语言向量的判别概率$D_s(t_i^{zh})$,即最小化$\log\left(1 - D_s\left(W^{s\to t} s_i^{vi}\right)\right)$;

[0025] 判别器Dis_t和生成器Gen_t同理可得:

[0026] $$\underset{Gen_t}{min}\ \underset{Dis_t}{max} V(D,G) = \underset{G}{min}\ \underset{D}{max} \left[\log(D_t(s_i^{vi}))\right] + \left[\log\left(1 - D_t\left(W^{t\to s} t_i^{zh}\right)\right)\right]$$

[0027] 训练时,生成器和判别器采取交替训练,即先训练Dis_s和Dis_t,然后训练Gen_s和Gen_t,不断往复,直到判别器无法识别词的来源则表示双语词向量位于一个公共语义空间。

[0028] 作为本发明的进一步方案,所述步骤Step2中:利用汉语和越南语新闻文本进行单语摘要模型训练会得到汉语和越南语两种摘要生成模型;利用汉越双语词向量和双语词典进行对抗学习得到实现将源语言映射到目标语言空间的映射矩阵W。

[0029] 作为本发明的进一步方案,所述步骤Step3中:将汉越新闻伪平行语料作为摘要模

型的输入,分别得到由Bert编码器生成的上下文文本向量。

[0030] 作为本发明的进一步方案,所述步骤Step4的具体步骤:

[0031] Step4.1、从汉越伪平行语料的参考摘要数据源构建一个汉越种子词典;

[0032] Step4.2、对Bert编码器生成的上下文文本向量表征应用映射矩阵W和种子词典进行半监督对抗学习得到汉越双语在同一语义空间下的对齐向量。

[0033] 作为本发明的进一步方案,所述步骤Step5的具体步骤为:

[0034] Step5.1、将经过对抗性学习训练后在同一语义空间下的对齐向量和参考摘要作为transformer解码器的输入;

[0035] Step5.2、解码端根据映射后的对齐向量与参考摘要之间的对数似然率来选取候选摘要;

[0036] Step5.3、解码器解码输出目标语言摘要。

[0037] 本发明的有益效果是:

[0038] 1、本发明的基于词对齐的半监督对抗学习跨语言摘要生成方法,利用双语词向量来表征汉越双语新闻文本,将汉语越南语的词都映射到同一语义空间中,在这个空间中语义相近的词向量距离相近,语义相关性低的词向量相隔较远;

[0039] 2、本发明的基于词对齐的半监督对抗学习跨语言摘要生成方法,使用预训练的越南Bert模型,能够较好的处理越南语新闻文本;

[0040] 3、本发明的基于词对齐的半监督对抗学习跨语言摘要生成方法,采用将双语映射任务和摘要生成任务联合学习的方法,降低了小语种因翻译效果不佳对跨语言摘要效果的影响;

[0041] 4、本发明的基于词对齐的半监督对抗学习跨语言摘要生成方法,实现了跨语言摘要生成任务,优化了跨语言摘要的效果。

附图说明

[0042] 图1为本发明中的流程图;

[0043] 图2为本发明中的双语词向量对抗训练模型图。

具体实施方式

[0044] 实施例1:如图1-2所示,基于词对齐的半监督对抗学习跨语言摘要生成方法,所述方法包括:

[0045] Step1、收集用于训练汉越跨语言摘要生成的新闻文本,以及获取已有的汉越双语词向量;从新浪微博中抽取的LCSTS数据,该语料主要是从新浪微博上整理的。每条语料均由两部分内容构成:短文本内容以及对应的参考摘要。而越南语语料则通过将以获取的LCSTS语料,然后借助谷歌翻译工具获取伪平行语料。其中训练集大约有20万对伪平行语料,测试及约有1000对伪平行语料。另外,还借助了互联网爬虫技术从中国新闻网、新华网、新浪新闻等国内新闻网站,以及越南每日快讯、越南经济日报,越南通讯社等越南新闻网站收集新闻,收集的数据包含新闻标题、正文详情、发布时间等信息。获得了约2000篇越南语新闻以及对应的10000篇中文可比语料。

[0046] Step2、利用汉越新闻文本和汉越双语词向量分别进行单语摘要模型和半监督对

抗学习的预训练；

[0047] Step3、利用Bert编码器分别对输入的汉越伪平行语料进行向量表征；

[0048] Step4、利用编码器获得的向量结合汉越双语种子词典进行半监督对抗学习，获取映射到同一语义空间的向量；

[0049] Step5、把映射在同一语义空间下的上下文文本向量和参考摘要作为transformer解码器的输入，解码输出目标语言摘要。

[0050] 作为本发明的进一步方案，所述Step1中，使用Scrapy作为爬取工具，模仿用户操作，为汉语、越南语新闻网站定制不同的模板，根据页面数据元素的XPath路径制定模板获取详细数据，获取新闻标题、新闻时间、新闻正文数据，以及从维基百科获取已经训练好的汉越双语词向量。

[0051] 此优选方案设计是本发明的重要组成部分，主要为本发明收集语料过程，为本发明为文本生成跨语言摘要提供了数据支撑。

[0052] 作为本发明的进一步方案，所述步骤Step2的具体步骤为：

[0053] Step2.1、首先使用一定数量的越南语和中文新闻文本语料，分别进行两种语言的单语摘要模型训练，摘要生成模型均采用Bert摘要模型，经过单语预训练得到的摘要模型；

[0054] Step2.2、利用维基百科提供的汉越双语词向量，进行半监督对抗学习方法如下：

[0055] （1）、分别得到维基百科提供的越南语和中文的词向量集：$S=\{S_1^d,S_2^d,\cdots,S_m^d\}$,$T=\{T_1^d,T_2^d,\cdots,T_n^d\}$,其中m和n表示词表大小，d表示嵌入向量的维度，S为越南语词向量集，T为中文词向量集；

[0056] (2)、预训练阶段，首先利用包含L＝30,000的$\{s_i,t_i\}_{i\in(1,2,\cdots,L)}$汉越双语种子词典，训练映射矩阵W：

[0057] $\Omega=\|Ws-t\|^2$

[0058] 其中，s为双语词典的源语言，t为对应的目标语言，Ω表示正则器，用于强制表达式两边的相等性，使用随机梯度下降学习W，然后通过最小化经过W转换的源语言单词s_i的向量表征与双语词典中的目标语言t_i之间的平方欧氏距离，来实现双语词向量在同一语义空间下的映射对齐；假设源语言词s为越南语，目标语言t为中文，则由越南语映射到中文的映射矩阵为$W^{s\to t}$，同理可得中文映射到越南文的矩阵为$W^{t\to s}$；

[0059] (3)联合训练阶段，用Bert编码器得到越南语和中文的词向量s_i^{vi}和t_i^{zh}来训练学习的映射矩阵W_s^*，W_T^*:

[0060] $$W_s^*=\underset{W\in M_d(R)}{argmin}\left\|W^{s\to t}s_i^{vi}-t_i^{zh}\right\|^2$$

[0061] $$W_T^*=\underset{W\in M_d(R)}{argmin}\left\|W^{t\to s}t_i^{zh}-s_i^{vi}\right\|^2$$

[0062] 其中d是嵌入向量的维度，$M_d(R)$是一个值为实数的d*d矩阵，$W^{s\to t}$和$W^{t\to s}$表示映射方向，目标是找到最佳映射矩阵W_s^*，以使映射源嵌入$W^{s\to t}s_i^{vi}$和目标语言嵌入t_i^{zh}之间的平方欧几里得距离最小，矩阵W_T^*同理；

[0063] (4)最后，将生成器生成的映射后的越南语向量和中文向量同时提交给判别器来

预测每个单词的来源,来优化判别器Dis_s和Gen_s:

[0064] $$\min_{Gen_s}\max_{Dis_s} V(D,G)=\min_{G}\max_{D}\left[\log(D_s(t_i^{zh}))\right]+\left[\log\left(1-D_s\left(W^{s\to t}s_i^{vi}\right)\right)\right]$$

[0065] 对于判别器Dis_s而言,看作是一个二分类问题,即一个形式为$D_s:s\to\{0,1\}$的函数,真实样本对应为1,映射样本对应为0,V(D,G)为二分类问题中常见的交叉熵损失,对于生成器Gen_s而言,为了尽可能欺骗判别器,所以需要最大化目标语言向量的判别概率$D_s(t_i^{zh})$,即最小化$\log\left(1-D_s\left(W^{s\to t}s_i^{vi}\right)\right)$;

[0066] 判别器Dis_t和生成器Gen_t同理可得:

[0067] $$\min_{Gen_t}\max_{Dis_t} V(D,G)=\min_{G}\max_{D}\left[\log(D_t(s_i^{vi}))\right]+\left[\log\left(1-D_t\left(W^{t\to s}t_i^{zh}\right)\right)\right]$$

[0068] 训练时,生成器和判别器采取交替训练,即先训练Dis_s和Dis_t,然后训练Gen_s和Gen_t,不断往复,直到判别器无法识别词的来源则表示双语词向量位于一个公共语义空间。

[0069] 此优选方案设计是本发明的重要组成部分,主要为本发明提模型与训练过程过程,为后续工作提供模型训练时所需摘要生成模型和双语词向量映射矩阵。并且为本发明识别事件时序关系提供了支撑和挖掘的对象。(结合其它步骤,它是一个数据输入,后面都会用得到);

[0070] 作为本发明的进一步方案,所述步骤Step2中:利用汉语和越南语新闻文本进行单语摘要模型训练会得到汉语和越南语两种摘要生成模型;利用汉越双语词向量和双语词典进行对抗学习得到实现将源语言映射到目标语言空间的映射矩阵W。

[0071] 作为本发明的进一步方案,所述步骤Step3中:将汉越新闻伪平行语料作为摘要模型的输入,分别得到由Bert编码器生成的上下文文本向量。

[0072] 作为本发明的优选方案,所述步骤Step3的具体步骤:

[0073] Step3.1、词在句中的不同位置有不同的语义信息,将每个位置编号,每个编号对应一个向量,通过位置向量和汉越双语词向量的结合,为每个词引入一定的位置信息,注意力机制即可以分辨出不同位置的词;

[0074] Step3.2、将汉越双语词向量和位置向量的拼接作为Bert编码器的输入。

[0075] 此优选方案设计是本发明的重要组成部分,主要为本发明提供向量编码的过程,结合双语词向量,并对每个词的位置进行编码有助于语义结构信息的获取,进而提升模型的性能。

[0076] 作为本发明的进一步方案,所述步骤Step4的具体步骤:

[0077] Step4.1、从汉越伪平行语料的参考摘要数据源构建一个汉越种子词典;

[0078] Step4.2、对Bert编码器生成的上下文文本向量表征应用映射矩阵W和种子词典进行半监督对抗学习得到汉越双语在同一语义空间下的对齐向量。

[0079] 作为本发明的进一步方案,所述步骤Step5的具体步骤为:

[0080] Step5.1、将经过对抗性学习训练后在同一语义空间下的对齐向量和参考摘要作为transformer解码器的输入;

[0081] Step5.2、解码端根据映射后的对齐向量与参考摘要之间的对数似然率来选取候

选摘要；

[0082] Step5.3、解码器解码输出目标语言摘要。

[0083] 本发明跨语言摘要生成网络包括由编码器和解码器构成的seq2seq摘要模型，以及实现汉越双语在同一个语义空间对齐的映射器和判别器。其中，左右两边的编码器分别为越南语和中文Bert编码器，经过Bert编码器得到两种语言文档向量表征V_s和V_t；mapping过程分别对编码器生成的向量进行线性映射处理，交由Dis_t和Dis_s鉴别向量是由编码器生成还是由映射得到；鉴别器无法判别的向量即为同一语义空间下的对齐向量，作为解码器的输入进行解码得到跨语言摘要。

[0084] 所述Bert编码器：

[0085] 为表示单个句子，文本由[CLS]这个标记作为开头，在每个句子末尾插入[SEP]标记，作为句子边界的表示。然后将预处理后的文本表示为一系列令牌$X=[w_1,w_2,\cdots w_n]$。另外，为了区分句子所在位置，本发明为每个句子分配E_A或E_B，这个取决于句子是奇数还是偶数。例如，对于文档$X=[w_1,w_2,\cdots w_n]$，将分配分句嵌入由$[E_A,E_B,\cdots,E_A]$表示。这样就实现了，分层进行文档学习，较低层的输出表示相邻句子的信息，最高层的输出则包含整个文本的重要信息；

[0086] 所述映射器和判别器：

[0087] 假定源语言为越南语，目标语言为中文。本发明使用映射后的源语言嵌入借助双语词典寻找与之对应的和目标语言嵌入，计算二者之间的点积作为相似性度量，相当于余弦相似性。例如，中文中的「足球」和越南语中的「bóng đá」在嵌入空间中距离非常近，因为它们在不同语言中代表着相同的意思。

[0088] 1)分别将给定的源语言和目标语言使用word2vec进行单语词向量训练，分别得到越南语和中文的词向量：

[0089] $$S=\{S_1^d,S_2^d,\cdots S_m^d\},T=\{T_1^d,T_2^d,\cdots T_n^d\} \quad (1)$$

[0090] 其中m和n表示词表大小，d表示向量维度。

[0091] 2)利用矩阵将嵌入投影到共同空间。即越南语词嵌入集合通过该映射函数转换后与目标语种的词嵌入很接近或者说重合。如果用一个有n＝20000的双语词典由$\{S_i,T_i\}_{i\in(1,n)}$对构成，本发明需要选择投影矩阵W:

[0092] $$W^*=\underset{W\in M_d(R)}{argmin}D_{ij}\|WS-T\|^2 \quad (2)$$

[0093] 其中d是嵌入向量的维度，Md(R)是一个值为实数的d×d矩阵。S和T是需要对齐大小为d×n的词嵌入矩阵。假设源语言第i个单词与目标语言第j个单词对应，那么本发明的目标是寻找最佳的映射矩阵W，实现S和T之间的欧氏距离最短。

[0094] 3)对来自越南语训练得到的向量集S应用映射函数f:

[0095] $S'=f(S)=SW$ (3)

[0096] 其中S′是映射后的向量集，W是映射矩阵。

[0097] 4)将映射后的越南语向量和中文向量同时提交给对抗神经网络的判别器来预测每个单词的来源。直到判别器无法识别词的来源则表示双语词向量位于一个公共语义空间。

[0098] 所述解码器：

[0099] 使用了一个基于transformer的解码器作为摘要层。本发明在预训练阶段分别使用了一定量的单语语料训练摘要模型，经过单语预训练之后的模型，为进一步解码映射后的向量解码降低了难度。编码器部分输出的向量经过预训练得到的对抗神经网络处理实现由源语言和目标语言映射到共享语义空间下，由生成器生成的映射向量的词对齐，得到映射后的向量作为解码器的输入，解码输出得到最终的摘要。

[0100] 进行单语训练时，给定一对平行的文本摘要对(x,s)，本发明的目标是最大化decoder生成的摘要：$\tilde{s} = argmax_{\tilde{s}} P(\tilde{s}|x)$。实验过程中进行最大对数似然率计算，其摘要损失值计算公式为：

[0101] $$L_{SUM}(x,s) = -\sum_{t=1}^{T} \log P(s_t|\tilde{s}_{<t}, V_x) \qquad (4)$$

[0102] 其中，T是参考摘要的长度，$\tilde{s}$是解码生成的摘要，V_x是编码器对文本x的编码生成序列。

[0103] 进行跨语言摘要任务训练时，给定一对平行的源语言新闻文本和目标语言参考摘要对(x,y)。则其跨语言摘要的损失函数计算公式为：

[0104] $$L_{SUM}(x,y) = -\sum_{t=1}^{T} \log P(y_t|\tilde{y}_{<t}, V_x) \qquad (5)$$

[0105] 其中，$\tilde{y}$是由映射后的向量解码生成的摘要。

[0106] 为了验证本发明的效果，分别对模型摘要生成、结果双语映射生成跨语言摘要过程进行实验探究，证明模型设置的合理性与高效性，又将该模型与现有模型进行对比，证明本方法在汉越双语跨语言摘要生成上具有较好效果。

[0107] 本文采用摘要任务中广泛使用的ROUGE分值作为评估指标，其工具包已被DUC和TAC等国际会议作为摘要体系的标准评价工具，用于预测生成文本和标准文本之间的接近程度。具体地说，摘要质量将依据模型预测生成的摘要与标准摘要的重叠单元进行量化计算，公式如下：

[0108] $$ROUGH-N = \frac{\sum_{S\in\{Ref Sum\}} \sum_{n-gram\in s} count_{match}(n-gram)}{\sum_{S\in\{Ref Sum\}} \sum_{n-gram\in s} count(n-gram)}$$

[0109] 其中n代表n-gram的长度，$Count_{match}$(n-gram)是模型生成摘要和人工书写的标准摘要中共同出现的n-gram的数量，公式旨在通过计算与参考摘要重叠的系统生成摘要中的n-gram的百分比来衡量系统生成摘要与参考摘要的匹配程度。本文将采用ROUGH评价指标N元共现统计ROUGH-1，ROUGH-2以及句子中最长公共子序列共现统计ROUGH-L，前者预定义n-gram的长度，后者使用最长公共子序列直接进行匹配，因此它自动包括最长的顺序共现，在一定程度上反映了句子结构信息。

[0110] 为验证本专利提出的摘要方法的可行性，如表1所示，Vi/Zh_BertSum表示模型预训练阶段汉越单语下的摘要结果。

[0111] 表1单语预训练Bert摘要实验结果

[0112]

方法	ROUGE-1	ROUGE-2	ROUGE-3
Vi_BertSum	25.3	16.1	23.6
Zh_BertSum	31.2	19.8	30.4

[0113] 为了验证翻译效果对于稀缺资源文本摘要生成的影响，如表2所示，本发明设置两

组基于管道翻译在同一数据集上的对比实验。VI-ZH CLS和ZH-VI CLS分别表示源语言为越南语或中文条件下的跨语言摘要对比实验。其中,Pipe_TS方法表示的是先进行原文本翻译,再进行单语摘要任务;Pipe_ST方法表示的是先进行单语摘要,再将生成摘要翻译的目标语言的结果;Ours即为本发明提出的基于词对齐的半监督对抗学习跨语言摘要生成方法。

[0114] 表2不同摘要生成方法对比实验结果

[0115]

方法	VI-ZH CLS			ZH-VI CLS		
	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
Pipe_TS	20.3	14.6	19.6	22.1	16.6	20.6
Pipe_ST	22.3	17.1	20.1	23.5	17.7	21.1
Ours	24.6	18.2	22.3	25.3	19.6	23.4

[0116] 实验数据表明,基于词对齐的半监督对抗学习跨语言摘要生成模型,将双语词向量映射到同一语义空间实现双语对齐的方法应用于跨语言摘要生成任务的有效性,能够有效改善跨语言摘要生成性能,ROUGE值在管道方法的实验结果上都有接近两个百分点的提升。其可能原因如下:(1)当前基于小语种的翻译技术尚未成熟,长篇幅的对文本进行翻译会造成信息损失,而在源语言上使用基于半监督对抗学习得到的映射矩阵能够在一定程度上保存文本信息;(2)半监督的对抗学习训练方式可能有助于获取文本摘要的高阶特征,这些特征可指导摘要生成中对原文中特定内容的选择。

[0117] 上面结合附图对本发明的具体实施方式作了详细说明,但是本发明并不限于上述实施方式,在本领域普通技术人员所具备的知识范围内,还可以在不脱离本发明宗旨的前提下作出各种变化。

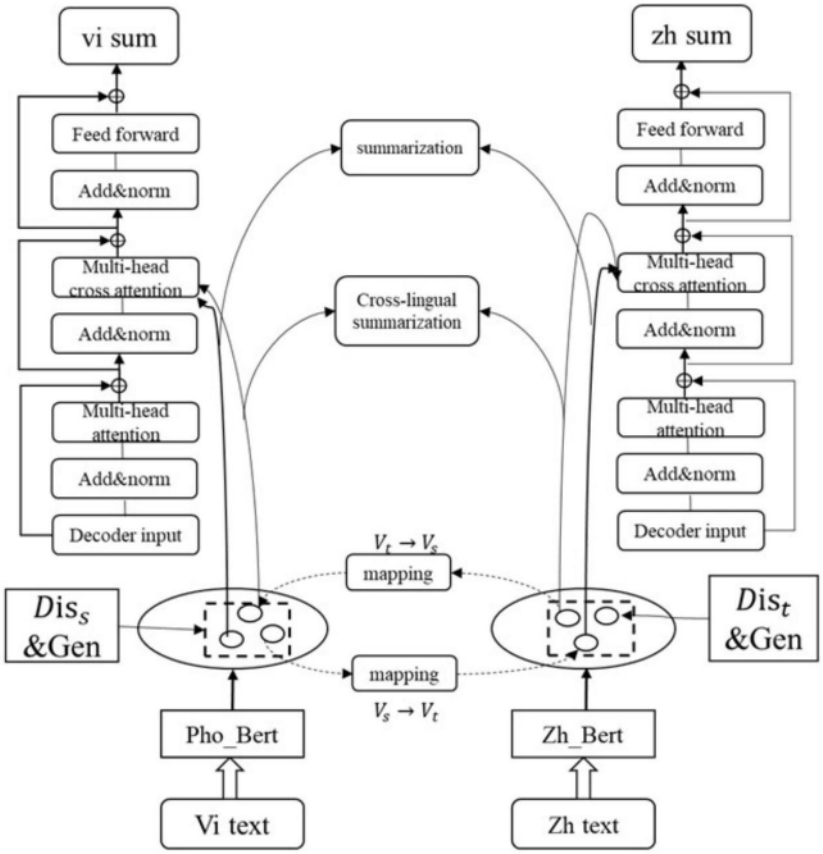

vi sum
Feed forward
Add&norm
Multi-head cross attention
Add&norm
Multi-head attention
Add&norm
Decoder input
summarization
Cross-lingual summarization
$V_t \rightarrow V_s$
mapping
mapping
$V_s \rightarrow V_t$
Dis_s &Gen
Dis_t &Gen
Pho_Bert
Vi text
Zh_Bert
Zh text
zh sum
Feed forward
Add&norm
Multi-head cross attention
Add&norm
Multi-head attention
Add&norm
Decoder input

图1

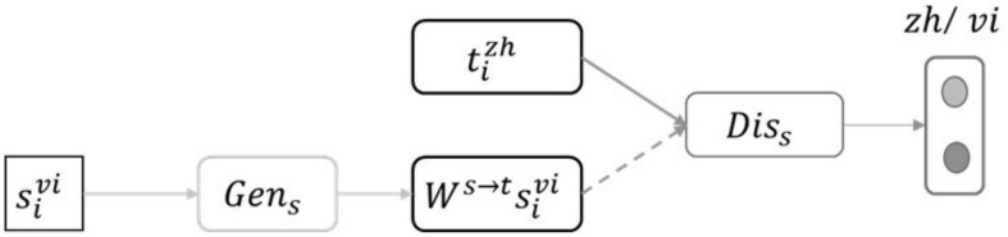

t_i^{zh}
Dis_s
$zh/\ vi$
s_i^{vi}
Gen_s
$W^{s \rightarrow t} s_i^{vi}$

图2