第 61 卷 第 4 期 厦门大学学报(自然科学版) Vol. 61 No. 4
2022 年 7 月 Journal of Xiamen University (Natural Science) Jul. 2022
doi:10.6043/j.issn.0438-0479.202110023

融合关键词概率映射的汉越低资源跨语言摘要

李笑萌,张亚飞*,郭军军,高盛祥,余正涛
(昆明理工大学信息工程与自动化学院,云南省人工智能重点实验室,云南 昆明 650500)

摘要:在低资源汉越跨语言摘要任务中,由于标注的汉越对齐数据稀缺,较难实现跨语言语义对齐.鉴于此,提出一种融合关键词概率映射的低资源跨语言摘要方法,首先利用源语言关键词实现关键信息的提取,然后基于概率映射对将源语言关键词映射到目标语言,最后基于指针网络将映射的目标语言关键词融入到摘要生成过程中.在构建的汉越跨语言摘要数据集上的实验结果表明,相比于直接的端到端的方法,融入关键词概率映射信息可以有效地提升低资源跨语言摘要的质量.

关键词:低资源跨语言摘要;跨语言语义对齐;关键词;概率映射
中图分类号:TP 391 **文献标志码**:A **文章编号**:0438-0479(2022)04-0630-08

跨语言摘要任务旨在为给定的一篇源语言文本生成另一种语言的摘要.跨语言摘要的传统方法是将源语言文本翻译到目标语言,然后对翻译后的文本进行摘要[1-2];或者先对源语言文本进行摘要,然后将源语言摘要翻译到目标语言[3-5].然而,目前机器翻译(MT)性能仍不能达到预期效果,存在结果错误传播的问题,尤其针对越南语等低资源语言,错误传播的问题更为显著.

近年来,跨语言摘要任务的研究方法主要是基于端到端的跨语言摘要方法.Zhu 等[6]基于端到端的模型,利用大规模跨语言摘要数据来实现跨语言摘要性能的提升,这也是第一个使用大规模的平行语料训练端到端跨语言摘要模型的方法,但是目前缺乏大规模高质量的跨语言摘要数据集.为了在少量的标注数据下使得端到端的跨语言摘要模型获得更好的性能:Yana 等[7]基于训练后的 MT 模型和单语摘要(MS)模型,将两者作为教师模型来教授跨语言摘要的学生模型;后来,Duan 等[8]将 Transformer 模型作为 MS 教师模型和学生模型的主干,进一步提升摘要的质量.但是这种教师-学生模型只能把有限的知识传递给学生模型.因此,为了使模型具备更好的跨语言学习的能力,基于预训练语言模型的跨语言摘要方法被提出,该方法通过预训练语言模型,使模型提前具备跨语言学习的能力,从而使模型获得更好的性能.Xu 等[9]提出了一种混合语言预训练模型,通过掩码语言模型(MLM)、MS 等进行预训练,为了提高模型跨语言学习的能力,预先训练的模型基于 MT 模型等提前从大量 MT 标注数据中学习语言知识.以上跨语言摘要任务主要是在汉英等富资源情况下进行,对于越南语等低资源语言情况并不适用.

目前,也有少量的研究是基于知识增强的方法来获得较好的低资源跨语言摘要效果.Li 等[10]提出自动摘要的正确性问题,通过联合学习摘要生成和文本隐含知识,提出了隐含感知解码器,通过用隐含信息丰富的编码器和解码器,来提高摘要的准确性.该研究结果表明:通过文本隐含知识增强模型的表征可提高摘要的准确性,这也说明将基于知识的学习融入摘要模型对于摘要模型性能的提升非常重要.传统的基于知识增强的跨语言摘要方法是通过构建双语词典,将作为输入的源语言文本和目标语言的参考摘要通过对齐的双语词典映射至同一语义空间,实现跨语言摘要.但是对于越南语等低资源来讲,获取对齐的双语

收稿日期:2021-10-19 **录用日期**:2021-11-20
基金项目:国家自然科学基金(61762056,61972186,61732005,61761026);国家重点研发计划(2018YFC0830105,2018YFC0830101,2018YFC0830100);云南省高新技术产业专项(201606);云南省重大科技专项计划(202002AD080001-5);云南省基础研究计划(202001AS070014,2018FB104)
*通信作者:zyfeimail@163.com
引文格式:李笑萌,张亚飞,郭军军,等.融合关键词概率映射的汉越低资源跨语言摘要[J].厦门大学学报(自然科学版),2022,61(4):630-637.
Citation:LI X M,ZHANG Y F,GUO J J,et al. Low resource cross-language summarization of Chinese-Vietnamese combined with keyword probability mapping[J]. J Xiamen Univ Nat Sci,2022,61(4):630-637.(in Chinese)

词典数据十分困难,较难实现跨语言语义对齐,而汉越概率映射对是一种针对语料库级别的全局知识,能够尽可能地反映双语之间的对应关系.基于数据驱动的端到端的跨语言摘要模型在低资源情况下由于没有先验知识指导,不能有效关注到源语言文本中的核心内容,导致生成的摘要出现内容偏差的问题,而关键词包含了源文的重要内容,是对源文的有效增强,因此从源语言文本中挖掘关键词信息并映射至目标语言生成相关的上下文,对于生成简洁、语义正确的跨语言摘要尤为重要.2017 年,See 等[11]提出指针生成器网络,实现了从源文本复制单词.受 See 等[11]的启发,本文认为通过指针生成器网络实现关键词的概率映射作为先验知识,可以增强模型跨语言表征的能力,指导摘要的生成.因此,针对标注数据稀缺导致的跨语言对齐困难等问题,本文提出了关键词概率映射,不仅关注了文本中的重要信息,且在一定程度上解决了跨语言对齐困难的问题.总体来说,本文的主要贡献包括以下两个方面:

1) 提出了融合关键词概率映射的汉越低资源跨语言摘要方法(low resource cross-language summarization of Chinese-Vietnamese combined with keyword probability mapping,C-Vcls),通过获取关键词的概率映射信息来改善汉越低资源跨语言摘要较难实现跨语言语义对齐,摘要质量差的问题;

2) 在构建的 10 万汉越低资源跨语言摘要数据集上进行对比实验,结果证明本文所提模型在汉越低资源跨语言摘要任务上的有效性和优越性.

1　C-Vcls 模型

本文 C-Vcls 模型基于 Transformer 框架,由融合关键词概率映射的文本表征和融合关键词概率映射的解码端构成,模型框架如图 1 所示.首先,获取源语言文本的关键词 G,与源语言文本 X 的隐状态表征 Z 通过编解码注意力机制获取第 s 个关键词与源语言文本的联合表征 ∂_s,查询得到每一个关键词对应汉越概率映射对中的源语言词,通过注意力机制获得第 s 个关键词(词向量表示为 w_s)对应到目标语言词(词向量表示为 w)的概率 $P(\boldsymbol{\omega}_s \Rightarrow \boldsymbol{\omega})$,最后通过指针网络,结合解码端生成的目标词的概率大小 $P_{\mathrm{N}}(\boldsymbol{\omega})$ 获得最后的分布 $P(\boldsymbol{\omega})$.

1.1　融合关键词概率映射的文本表征

给定一组跨语言摘要数据集 D:$D=\{X,Y\}$,其中 X 为源语言文本输入序列,即 $X=\{x_1,x_2,\cdots,x_n\}$,Y 为目标语言参考摘要输入序列,即 $Y=\{y_1,y_2,\cdots,y_m\}$.n,m 跟随源序列长度变化,$n>m$.

编码端输入的源语言文本通过编码器得到输出的隐状态表征 $\boldsymbol{Z}$,如式(1)所示.

$$\boldsymbol{Z}=[z_1,z_2,\cdots,z_l,\cdots,z_n]. \tag{1}$$

本文使用的关键词抽取方法为关键词提取算法 TextRank[12].基于此算法,对每篇源语言文本提取 q 个最重要的关键词,即关键词集合 G 由式(2)所示:

$$G=\{g_1,g_2,\cdots,g_s,\cdots,g_q\}=\mathrm{TextRank}(x_1,x_2,\cdots,x_n). \tag{2}$$

然后,通过编解码注意力机制对关键词和源语言文本进行联合表征,构建关键词到源语言文本的注意力,计算第 s 个关键词 g_s 对于源语言文本的注意力得分,如式(3)所示.

$$\partial_s=\frac{1}{n}\sum_{n=0}^{n}\mathrm{Attention}(g_s,z_l,z_l). \tag{3}$$

为了对关键词信息进行跨语言对齐,映射至目标语言,本文采用汉越概率映射对进行实现.汉越概率映射对的构建在本文构建的汉越跨语言摘要数据集 D 上进行.设 $C=\{c_1,c_2,\cdots,c_i,\cdots,c_j\}$ 为 D 中源语言词的集合,$V=\{v_1,v_2,\cdots,v_k,\cdots,v_r\}$ 为 D 中相对应的目标语言的映射候选词的集合,j,r 表示集合的大小.如图 1 中概率映射对构建模块可知,在对应关系中,可以映射为源语言词“我”的映射候选词有“Chúng”“tôi”,其中“我→tôi”表示一个映射对,利用统计的思想,即“我”映射为“Chúng”的概率为 1/3,映射为“tôi”的概率为 2/3.为了实现这一功能,本文利用 Dyer 等[13]提出的快速对齐方法和统计的思想,通过快速对齐方法得到每一个源语言词 c_i 映射为目标语言 v_k 的概率为:

$$P_{ik}^{\mathrm{MP}}=\frac{\|\{v_k \mid c_i \rightarrow v_k\}\|}{\|\{c_i \mid c_i \rightarrow v_k\}\|}, \tag{4}$$

其中,$c_i \rightarrow v_k$ 表示一个映射对,$\|\{c_i \mid c_i \rightarrow v_k\}\|$ 表示数据集 D 中源语言集合中所有满足映射关系 $c_i \rightarrow v_k$ 的词 c_i 的个数,$\|\{v_k \mid c_i \rightarrow v_k\}\|$ 表示满足映射关系 $c_i \rightarrow v_k$ 的映射候选词 v_k 的个数,P_{MP} 表示汉越概率映射对在数据集 D 中源语言词可以映射为满足映射关系的映射候选词的概率.

为了将关键词映射到目标语言,本文使用了编解码注意力机制查询得到每一个关键词对应汉越映射对中的源语言词,进而得到该关键词的映射候选词.如图 1 所示,为了将关键词“我”映射至目标语言,首先查询得到“我”在汉越概率映射对中对应的源语言词的位置,进而得到对应的映射候选词“Chúng”“tôi”

等,然后利用编解码注意力机制,构建关键词到映射候选词的注意力,即计算出每一个关键词对应其汉越映射概率对的映射概率,取其映射概率最大的映射候选词"tôi"作为目标语言关键词. 具体如公式(5)所示.

$$P(\boldsymbol{\omega}_s^{\mathrm{rc}} \Rightarrow \boldsymbol{\omega}) = \mathrm{Attention}(\boldsymbol{\omega}_s, \boldsymbol{\omega}, \boldsymbol{\omega}), \tag{5}$$

其中:$\boldsymbol{\omega}_s^{\mathrm{rc}}$ 是源语言关键词 g_s 的词向量表示,作为查询;$\boldsymbol{\omega}$ 是根据概率 P^{MP} 查询得到的第 s 个关键词对应的映射候选词的词向量表示,作为键和值. $P(\boldsymbol{\omega}_s \Rightarrow \boldsymbol{\omega})$ 是源语言关键词通过汉越映射对映射为目标语言的概率.

1.2 融合关键词概率映射信息的解码端

在得到关键词的概率映射信息后,融合模块的功能是将关键词概率映射信息融合到跨语言摘要生成过程中,指导摘要的生成. 本文利用 See 等[11]提出的指针网络,通过生成概率 P_{gen} 来确定最后生成的摘要词汇分布.

使用 $\boldsymbol{O}$ 作为解码器在时刻 t 的隐状态,计算时刻 t 的生成概率 P_{gen},$P_{\mathrm{gen}} \in (0,1)$,如式(6)所示.

$$P_{\mathrm{gen}} = \delta(\boldsymbol{W}_2(\boldsymbol{W}_1\boldsymbol{O} + \boldsymbol{b}_1) + b_2), \tag{6}$$

其中,$\boldsymbol{W}_1 \in \mathbf{R}^{d_{\mathrm{model}} \times d_{\mathrm{model}}}$、$\mathbf{W}_2 \in \mathbf{R}^{1 \times d_{\mathrm{model}}}$ 是学习矩阵,$b_1 \in \mathbf{R}^{d_{\mathrm{model}}}$、$b_2 \in \mathbf{R}$ 是偏置向量,d_{model} 表示此时隐状态的维度,δ 是 sigmoid 函数. 本文中,P_{gen} 被用作一个软开关,用于选择从解码端生成一个单词,或者选择从关键词中复制一个单词. 那么,生成一个单词的概率 $P(\boldsymbol{\omega})$ 如式(7)所示.

$$P(\boldsymbol{\omega}) = P_{\mathrm{gen}} \sum_s \partial_s P(\boldsymbol{\omega}_s \Rightarrow \boldsymbol{\omega}) + (1 - P_{\mathrm{gen}}) P_{\mathrm{N}}(\boldsymbol{\omega}), \tag{7}$$

其中,$P(\boldsymbol{\omega}_s \Rightarrow \boldsymbol{\omega})$ 表示关键词 $\boldsymbol{\omega}_s$ 映射到词 $\boldsymbol{\omega}$ 的概率大小,$P_{\mathrm{N}}(\boldsymbol{\omega})$ 表示本模型的解码端生成的词 $\boldsymbol{\omega}$ 的概率大小,$P(\boldsymbol{\omega})$ 是通过生成概率 P_{gen} 决定的最终生成摘要的词汇分布.

2 实 验

2.1 实验数据

本文数据来自互联网爬取,基于 Zhu 等[6]提出的往返翻译的策略,获得了质量较高的 10 万汉越、汉英

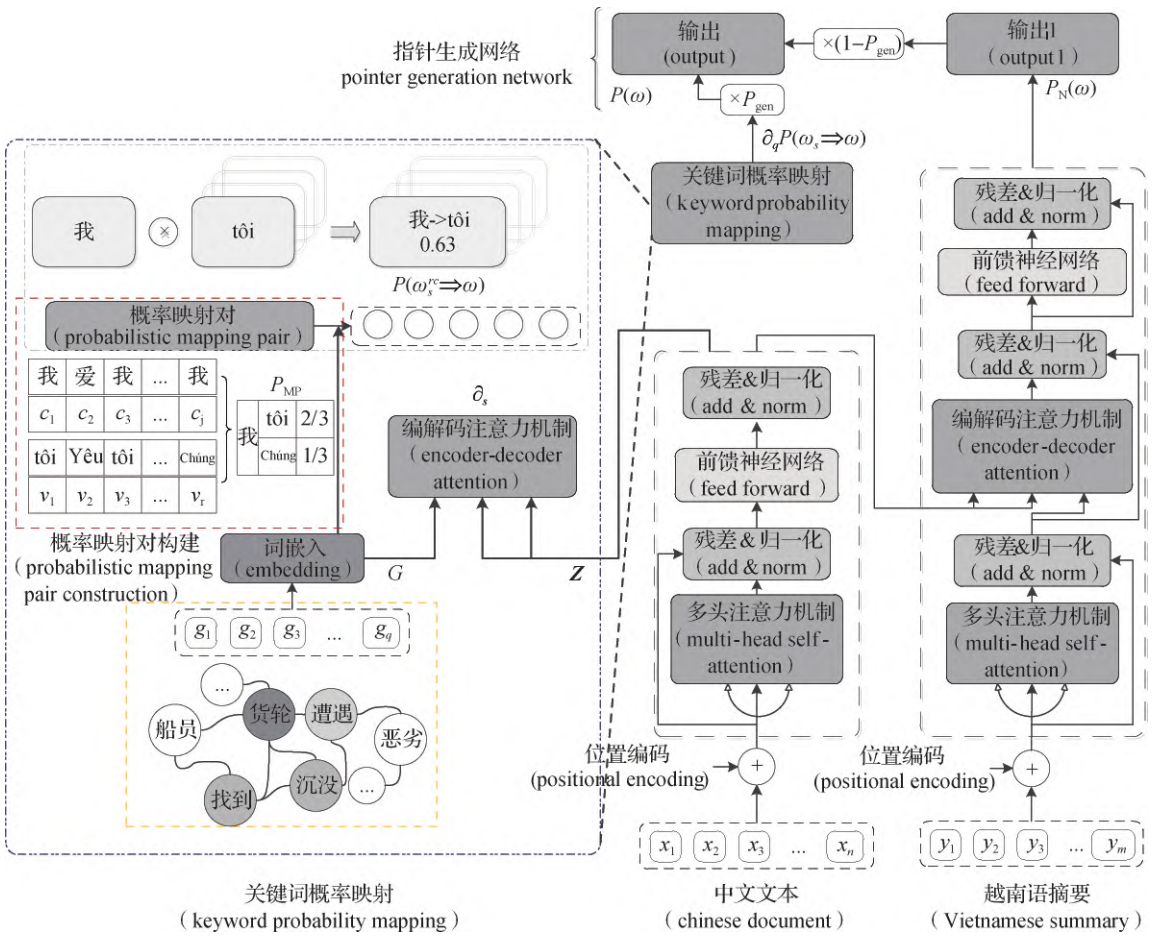

图 1 融合关键词概率映射的汉越低资源跨语言摘要方法框架

Fig. 1 Low resource cross-language summarization of Chinese-Vietnamese combined with Kp-mapping

跨语言摘要数据集(https://github.com/Lxmllx/C-Vcls-dataset/tree/master),其中有效词数为数据集文本分词去重后的剩余词数.表 1 中列出了本文数据集的统计信息.分词处理过程中,汉语使用结巴分词,越南语使用 Vu 等[14]提出的 VnCoreNLP 进行分词,英语采用其本身的词级结构.

表 1　数据集统计结果

Tab. 1　Statistic result of dataset

语料库	训练集	验证集	测试集
汉越文本摘要对	95 000	3 000	2 000
汉英文本摘要对	95 000	3 000	2 000
汉语有效词数	203 200	30 200	23 600
越南语有效词数	53 400	9 800	7 600
英语有效词数	73 000	10 000	8 400

2.2　评价指标

本文利用自动摘要中常用的 ROUGE(recall-oriented understudy for gisting evaluation)值作为评价指标[15],它通过比较候选摘要与参考摘要中共现的 n 元词组(n-gram)来评价候选摘要的质量,其计算方法为:

$$\text{ROUGE-}N = \frac{\sum_{s\in A_R}\sum_{\text{n-gram}\in As} N_{\text{match}}(n\text{-gram})}{\sum_{s\in A_R}\sum_{\text{n-gram}\in As} N(n\text{-gram})}, \tag{8}$$

其中,n-gram 表示 n 元词组,A_R 表示标准的参考摘要,As 表示生成摘要句,$N(n\text{-gram})$表示参考摘要中 n 元词组的个数,$N_{\text{match}}(n\text{-gram})$表示生成摘要句与参考摘要句共同包含的 n 元词组的个数.根据 n-gram 的不同,本文采用 ROUGE-1(一元组、RG-1),ROUGE-2(二元组、RG-2),ROUGE-L(最长子序列、RG-L)来评价参考摘要的好坏.

2.3　实验模型参数设置

本文所有实验均基于 Transformer 架构,采用 Adam 优化器,其中,$\beta_1 = 0.9$,$\beta_2 = 0.998$,$\varepsilon = 1\times10^{-9}$.在训练过程中使用的标签平滑率 $e_{ls} = 0.1$.在验证时使用波束大小为 4 且长度罚分 $\alpha = 0.6$ 的波束搜索.本文采用的学习率 $l_r = 0.1$,批次大小设为 2 048,dropout 为 0.1,编码器和解码器层数、模型隐层大小、前馈隐层大小和头数分别为 6,1 024,2 048 和 8.本文设置编解码器词表大小为:汉语 10 万,英语和越南语均为 1 万,未登录词使用<unk>来代替.本文实验中概率映射词典的大小根据词频设置为 39 311,关键词个数根据摘要平均长度取为 5.本文所有实验均在单个 NVIDIA RTX 2070 SUPER GPU 上进行.

2.4　基准模型

本文选择 TETran、TLTran、NCLS 模型作为基准模型,所有基准模型的训练集、验证集和测试集划分均与本文模型相同.

1) TETran 模型和 TLTran 模型为传统的跨语言摘要模型,其中 TETran 模型表示先利用 MT 模型将源语言文本翻译到目标语言,然后使用 LexRank[16]模型对翻译后的源文档进行摘要.TLTran 模型表示先利用 MS 模型对源语言文本进行摘要,然后利用翻译模型将生成的源语言摘要翻译至目标语言.

2) NCLS[6]模型是一种基于 Transformer 的端到端的跨语言摘要模型.

3) C-Vcls 模型是本文实现的基于 Transformer 的序列到序列模型.此模型引入关键词的概率映射信息作为先验知识.

2.5　实验结果分析

2.5.1　实验结果

为了证明本文融合关键词概率映射方法在汉越低资源跨语言摘要任务上的优势,将本文模型与现有基准模型在汉越跨语言摘要数据集上进行实验对比,表 2 给出了本文模型与基准模型在汉越跨语言摘要测试集上的 RG-1,RG-2 和 RG-L 的对比结果.

表 2　汉越跨语言摘要模型的实验结果对比

Tab. 2　The experimental comparison results of the Chinese-Vietnamese cross-language summarization models

模型	RG-1/%	RG-2/%	RG-L/%
TETran	16.98	7.15	15.44
TLTran	21.91	9.38	18.64
NCLS[6]	19.16	8.56	17.32
C-Vcls	**23.01**	**9.45**	**20.15**

由表 2 可知:TLTran 优于 TETran,说明先翻译后摘要的方法更容易受 MT 性能的影响,发生错误传播.C-Vcls 模型与传统的 TLTran、TETran 模型相比,在 RG-1,RG-2 和 RG-L 上分别取得了1.10,0.07,1.51 和 6.03,2.30,4.71 个百分点的提升,这也说明了不仅关键词概率映射的策略可以有效缓解越南语

MT 性能不佳引起的摘要质量差的问题,通过获得源文关键词的联合表征还可以获得更好的上下文表示,使生成的摘要更精准. 另外,C-Vcls 模型与端到端的 NCLS 模型相比,在 RG-1,RG-2 和 RG-L 上取得了 3.85,0.89,2.83 个百分点的提升,相较于需要大规模语料的端到端的 NCLS 模型,本文构建源文关键词的联合表征并融入关键词的概率映射,通过先验知识增强模型的跨语言表征能力,降低了模型对语料规模的要求,从而取得了更优的性能. 因此,通过以上分析,可以得出明确结论:本文提出的关键词概率映射方法是一种有效的方法,可以有效提高端到端模型的性能.

2.5.2 融合关键词概率映射方法的有效性分析

在 2.5.1 节中,融合关键词概率映射方法能有效提高端到端模型的性能. 为了进一步证明本文融合关键词概率映射模块在汉越低资源跨语言摘要任务上的合理性,本文设置了多组实验进行验证.

1) 关键词融入的有效性

表 3 中给出了关键词个数 q 不同时,C-Vcls 模型在汉越跨语言摘要测试集上的 RG-1,RG-2,RG-L 的比对结果.

表 3 关键词个数对 C-Vcls 模型的影响

Tab. 3 The influence of the number of keywords on the C-Vcls model

q	RG-1/%	RG-2/%	RG-L/%
0	19.16	8.56	17.32
1	19.87	8.73	18.04
2	21.32	8.93	19.87
5	**23.01**	**9.45**	**20.15**

分析表 3 可知,$q=5$ 时,模型取得了更优的性能. 随着 q 从 0 增加到 5,C-Vcls 模型在汉越跨语言摘要测试集上指标 RG-1、RG-2 和 RG-L 不断增加. 与 $q=0$ 相比,$q=5$ 时在指标 RG-1、RG-2 和 RG-L 上分别获得了 3.85、0.89、2.83 个百分点的性能提升. 原因可能是随着关键词个数的增多,获得的文本关键信息越多,对摘要的指导性越强,获得的摘要越可靠. 综上,表明了关键词等先验知识对摘要模型的指导可以有效提升低资源摘要模型的性能.

2) 概率映射策略的有效性

为验证概率映射策略的有效性,本文在概率映射词典的大小上进行相关实验. 根据词频设置概率映射词典大小为 25 087,36 368,39 311,42 399,表 4 中给出了本文模型在汉越跨语言摘要数据集上的 RG-1、RG-2、RG-L 的比对结果,其中覆盖率为概率映射词典相对于关键词词数的占比(此处由 TextRank 得到的关键词未进行去重,故覆盖率的分母不一样).

表 4 概率映射词典对模型的影响

Tab. 4 The impact of the probability mapping dictionary on the model

概率映射词典大小	覆盖率/%	RG-1/%	RG-2/%	RG-L/%
25 087	52.37	16.92	6.67	15.68
36 368	71.27	19.28	7.44	18.24
39 311	78.19	**23.01**	9.45	**20.15**
42 339	80.32	22.98	**9.46**	20.11

分析表 4 可知,概率映射词典大小为 39 311 是性能最好的,在指标 RG-1、RG-2 和 RG-L 上,相较概率映射词典大小为 25 087,36 368,42 339 时分别有 6.09,2.27,4.46,3.73,2.01,1.91;0.03,−0.01,0.04个百分点的提升. 概率映射词典大小为 25 087 时的汉越跨语言摘要效果较差,主要原因可能是覆盖率仅有 52.37%,此时词典的噪声较大,覆盖率较低,在进行映射时不能对关键词进行有效映射,导致部分关键词不起作用,相对于其它模型的结果(表 2),甚至会降低摘要的效果;但是在概率映射词典为 39 311 和 42 339 时,摘要效果相对于其它模型的结果(表 2),仍有提升,但是两者相差不大,这是由于最终生成摘要的单词分布由概率映射词典、翻译概率、神经网络模型生成单词的分布共同决定,可能产生的不确定性较大. 综上,说明了概率映射词典这一策略在汉越跨语言摘要任务上的有效性,但是概率映射词典对于关键词的覆盖率在一定程度上影响了模型的性能.

3) 概率映射以及指针网络对于 C-Vcls 模型的有效性

为验证本文所结合的概率映射以及指针网络策略的作用,本文在汉越低资源跨语言摘要数据集上进行相关实验. 其中,C-Vcls-MP 模型是在 C-Vcls 模型的基础上减少概率映射模块,C-Vcls-PN 模型是在 C-Vcls模型的基础上减少指针网络模块而选择直接拼接 $P_{\text{gen}}\sum_s \partial_s p(w_s \Rightarrow w)$ 与 $(1-P_{\text{gen}})p_{\text{N}}(w)$ 的方式进行关键词的融合.

分析表 5 可知,C-Vcls 模型取得了更好的效果. C-Vcls 模型较 C-Vcls-MP 模型在指标 RG-1、RG-2 和 RG-L 上取得了 4.77,4.52 和 3.21 个百分点的提升,该结果表明当关键词不进行概率映射时,摘要结果下降最为严重且摘要性能低于 NCLS 模型,可能是由于关键词不进行映射时,会给模型引入更多的噪声,说明关键词概率映射模块在模型中起着至关重要的作用,能够建模关键词映射到目标语言作为先验知识指导跨语言摘要的生成. 而 C-Vcls-PN 模型相对于 C-Vcls 模型,在指标 RG-1、RG-2 和 RG-L 上的性能分别下降了 2.45,2.74 和 2.26 个百分点;但是相较 NCLS 模型,RG-L 指标上仍然取得了 0.57 个百分点的增幅,这也说明,尽管融合的方式不同,但是融入关键词概率映射信息到端到端的模型中确实对模型性能的提升是有帮助的,而且指针网络的融合方式优于直接拼接的融合方式. 综合以上分析,本文所提概率映射以及指针网络进行融合的方式对模型的性能提升是有益的.

表 5　概率映射、指针网络对 C-Vcls 模型的影响

Tab. 5　The influence of probability mapping and pointer network on C-Vcls model

模型	RG-1/%	RG-2/%	RG-L/%
C-Vcls-MP	18.24	4.93	16.94
C-Vcls-PN	20.56	6.71	17.89
NCLS[6]	19.16	8.56	17.32
C-Vcls	**23.01**	**9.45**	**20.15**

4) C-Vcls 模型与基准模型在汉英跨语言摘要测试集上的对比

为了验证本文所提模型的泛化性,本文在汉英跨语言摘要数据集上进行实验. 表 6 给出了本文模型与基准模型在汉英跨语言摘要数据集上的 RG-1、RG-2 和 RG-L 的比对结果.

表 6　汉英跨语言摘要模型的实验结果对比

Tab. 6　The experimental comparison results of the Chinese-English cross-language summarization models

模型	RG-1/%	RG-2/%	RG-L/%
TETran	15.15	4.18	14.10
TLTran	19.81	7.34	16.37
NCLS[6]	16.07	4.34	15.69
C-Vcls	**21.37**	**8.01**	**18.67**

分析表 6 可知,本文模型的指标均优于基准模型. C-Vcls 模型较 TLTran 模型和 TETran 模型在指标 RG-1、RG-2、RG-L 上分别有 1.56,0.67,2.30 和 6.22,3.83,4.57 个百分点的提升;较 NCLS 模型有 5.30,3.67,2.98 个百分点的提升. 根据表 2 和 6 可以看出,同样数量级的数据在同样的基准模型上,不同的数据集取得的结果有所差异,且在汉英跨语言摘要数据集上的实验结果低于汉越跨语言摘要数据集. 主要原因是因为虽然越南语和英文构造的词典均为 1 万,但根据越南语和英文文本构造特点及本文数据集的有效词数来看,越南语词典对于测试集文本的覆盖率高于英文词典对于测试集文本的覆盖率,即汉越跨语言摘要的实验结果没有大量未登录词<unk>的出现,进而提高了摘要的准确性. 但是,从实验结果依然可以看出,本文提出的引入具有引导性的关键词概率映射的方法对于汉英跨语言摘要任务同样有效,也证明了本文所提模型的泛化性.

2.6　实例分析

为了进一步验证算法的有效性,本文列举了不同模型的摘要结果. 具体如表 7 所示,源语言文本与标准摘要都来自汉越跨语言摘要数据集. 本文列举出了所有基准模型的输出结果作为对比,为了便于理解,本文给出了对应汉语的翻译结果.

分析表 7 可知,源语言文本主要讲述 19 名前往张家口的驴友被困海坨山,其中 15 名驴友失去联系的事实. 由于模型限制,传统模型 TETran 模型表达出了 19 名来自河北石家庄的朋友,但是并没有表述出 15 名前往张家口的朋友在河北失去联系的关键信息;TLTran 模型表现相对较好,但是仍然没有表输出“张家口”的关键事实. 而对于端到端的 C-Vcls 模型和 NCLS 模型均能表达出“15 名驴友”的主要信息,但是 NCLS 模型,并没有体现出其“失去联系”的关键信息,且内容过于冗杂,而本文提出的融合关键词概率映射的策略,获取源文中的关键词“北京”“失去”“联系”等映射至目标语言,通过有关键词概率映射信息等具有引导性信息的融入增强了模型的跨语言表征能力,提高了摘要的信息覆盖度以及事实性,生成质量更高的文本摘要.

3　结　论

针对汉越低资源跨语言摘要,本文在 Transformer 框架下,提出关键词概率映射方法. 通过实验证明,在低资源情况下,通过获取源语言文本的关键词信息映射至目标语言指导摘要生成的方式,对汉越低资源跨

表 7 不同模型生成摘要样例

Tab. 7 Sample summary of different models

名称	内 容
源语言文本	2 日晚,河北张家口赤城县公安机关接到北京转来的报警,称在有 3 名驴友和 1 名向导被困海坨山,还有15 名驴友失去联系. 当地有关部门已立即赶赴现场搜救. 据了解,这 19 名驴友并非专业登山爱好者,大多为散客,目前被困具体人数、位置正在进一步核实中.
参考摘要	15 người bạn đi du lịch Bắc Kinh mất liên lạc ở Trương Gia Khẩu(15 位前往北京的朋友在张家口失去联系)
TETran	Vào tối ngày 2,19 người bạn đã nhận được một cảnh sát ở huyện <unk>, tỉnh Hà Bắc Kinh và 1 hướng dẫn viên du lịch bị mắc kẹt(第 2 天晚上,有 19 个朋友从河北省接待了一名警察,导游被困)
TLTran	15 người bạn đi du <unk> lịch Bắc Kinh mất liên lạc ở Hà Bắc(15 位前往北京的朋友在河北失去联系)
NCLS[6]	15 người bạn đi du lịch Bắc Kinh bị mắc kẹt ở Trương Gia Khâu trong những ngày ở Hà Bắc. Bạn có thể đi du lịch Bắc ?(在河北张家口期间,有 15 位前往北京的朋友被困在张家口. 你可以北行吗?)
C-Vcls	15 người bạn đi du lịch Bắc Kinh mất liên lạc ở Hà Bắc(15 位前往北京的朋友在河北失去联系)

语言摘要任务存在一定的提升,通过实验也可以证明,利用关键词概率映射信息可以为跨语言摘要模型提供更丰富的指导信息,也证明本文提出的方法对低资源跨语言摘要任务可能是更加有效的. 多模态等多源信息是对文本内容的高度概括,可以很好的对文本内容进行信息补充. 因此,如何利用多模态信息对跨语言摘要进行指导是下一步研究的重点.

参考文献:

[1] LEUSKI A, LIN C Y, ZHOU L, et al. Cross-lingual C* ST* RD: English access to Hindi information[J]. ACM Transactions on Asian Language Information Processing, 2003, 2(3): 245-269.

[2] OUYANG J, SONG B, MCKEOWN K. A robust abstractive system for cross-lingual summarization[C]// Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: ACL, 2019: 2025-2031.

[3] LIM J M, KANG I S, LEE J H. Multi-document summarization using cross-language texts[EB/OL]. [2021-10-01]. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings4/TSC/NTCIR4-TSC-LimJM.pdf.

[4] ORĂSAN C, CHIOREAN O A. Evaluation of a cross-lingual romanian-english multi-document summariser[C]// International Conference on Language Resources and Evaluation. Marrakech: LREC, 2008: 2114-2119.

[5] WAN X, LI H, XIAO J. Cross-language document summarization based on machine translation quality prediction[C]// Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2010: 917-926.

[6] ZHU J N, WANG Q, WANG Y N, et al. NCLS: neural cross-lingual summarization[C]// Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Stroudsburg: ACL, 2019: 3045-3055.

[7] YANA A, SHEN S Q, CHEN Y, et al. Zero-shot cross-lingual neural headline generation [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, 26(12): 2319-2327.

[8] DUAN X Y, YIN M M, ZHANG M, et al. Zero-shot cross-lingual abstractive sentence summarization through teaching generation and attention[C]// Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2019: 3162-3172.

[9] XU R, ZHU C, SHI Y, et al. Mixed-lingual pre-training for cross-lingual summarization[EB/OL]. [2021-10-01]. https://arxiv.org/abs/2010.08892.

[10] LI H, ZHU J, ZHANG J, et al. Ensure the correctness of the summary: incorporate entailment knowledge into abstractive sentence summarization[C]// International Conference on Computational Linguistics. New Mexico: COLING, 2018: 1430-1441.

[11] SEE A, LIU P J, MANNING C D. Get to the point: summarization with pointer-generator networks[C]// Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2017: 1073-1083.

[12] MIHALCEA R, TARAU P. TextRank: bringing order into texts[C]// Conference on Empirical Methods in Natural Language Processing. Barcelona: UNT Scholarly Works, 2004: 404-411.

[13] DYER C, CHAHUNEAU V, SMITH N A. A simple, fast, and effective reparameterization of IBM model 2 [C]// Conference of the North American Chapter of the Association for Computational Linguistics: Human

Language Technologies. Atlanta: NAACL-HLT, 2013: 644-648.

[14] VU T, NGUYEN D Q, NGUYEN D Q, et al. VnCoreNLP: a Vietnamese natural language processing toolkit[C]// Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations. Stroudsburg: ACL, 2018: 56-60.

[15] LIN C Y. ROUGE: a package for automatic evaluation of summaries[C]// Workshop on Text Summarization Branches Out, Barcelona. Stroudsburg: ACL, 2004: 74-81.

[16] ERKAN G, RADEV D R. LexRank: graph-based lexical centrality as salience in text summarization[J]. Journal of Artificial Intelligence Research, 2004, 22: 457-479.

Low resource cross-language summarization of Chinese-Vietnamese combined with keyword probability mapping

LI Xiaomeng, ZHANG Yafei*, GUO Junjun, GAO Shengxiang, YU Zhengtao

(Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Key Laboratory of Artificial Intelligence of Yunnan Province, Kunming 650500, China)

Abstract: For cross-language summarization tasks, the scarcity of annotated Chinese-Vietramese alignment data leads to difficulties of achieving cross-language semantic alignments. In view of this drawback, for the task of Chinese-Vietnamese cross-language summarization, a low-resource cross-language summarization method that integrates keyword probability mapping is proposed. First, source language keywords are used to extract important information. Second, we map source language keywords to the target language based on the probability mapping pair. Finally, we integrate the mapped target language keywords into the abstract generation process based on the pointer network. Experimental results on the constructed Chinese-Vietnamese cross-language abstract data set show that, in comparison with NCLS and other methods based on sequence to sequence, incorporating keyword probability mapping information can effectively improve the quality of low-resource cross-language abstracts.

Keywords: low-resource cross-language abstract; cross-language semantic alignment; keywords; probability mapping

(责任编辑:任滢滢)