融合关键词概率映射的汉越低资源跨语言摘要

李笑萌 [1,2]，张亚飞 [1,2*]，郭军军 [1,2]，高盛祥 [1,2]，余正涛 [1,2]

（1. 昆明理工大学信息工程与自动化学院，云南 昆明 650500；2. 云南省人工智能重点实验室，云南 昆明 650500）

摘要：低资源场景下的跨语言摘要任务标注数据稀缺，基于小规模对齐数据实现跨语言语义对齐较为困难。鉴于此，针对汉越跨语言摘要任务，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重要信息的提取，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 NCLS 等基于序列到序列的方法，融入关键词概率映射信息可以有效的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

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，目前机器翻译并不完善，导致了结果错误传播。近年来，跨语言摘要的研究方法主要集中在获得大规模跨语言摘要数据集来训练模型[6]；或者利用现有的机器翻译模型获得伪摘要句来训练跨语言摘要模型[7][8]；或者基于跨语言任务和单语任务模型，利用大量单语数据增强其跨语言模型的建模能力[12]。但是，这些方法严重依赖于大规模标注数据来改善摘要质量，很难迁移到低资源跨语言摘要上，致使低资源环境下跨语言对齐困难。那么，如何建立两种语言之间的关系增强跨语言的表征来解决跨语言对齐困难的问题。

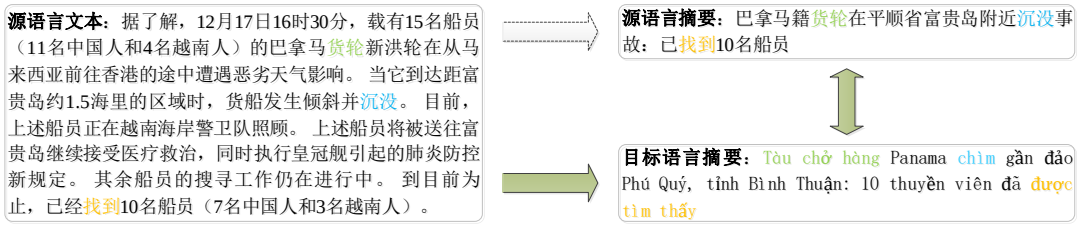

图 1 汉越跨语言摘要示例

Fig.1 Example of Chinese-Vietnamese Cross-Language Summary

图 1 中给出了基于汉越的跨语言摘要示例，其源语言文本讲述了“货轮沉没，正在搜救且已救出 10 名船员”的事实。其摘要中“货轮”“沉没”“找到”等已包括了文本中的核心内容。如果挖掘文本这些核心内容到摘要中，就可以用来辅助摘要的生成。在汉越低资源跨语言摘要中，“tàu chở hang-货轮”、“chìm-沉没”、“được tìm thấy-找到”等为文本核心内容，且为双语对齐的词对。如果挖掘文本中这些核心词并获得其跨语言对齐词，就可以用来指导跨语言摘要的生成。那么，针对源语言文本中核心词提取以及跨语言对齐困难的问题，受启发于 Li 等人[9]的方法，本文提出了关键词概率映射（Keyword probability mapping，Kp-mapping），不仅关注了文本中的重要信息，且在一定程度上解决了其跨语言对齐困难的问题。

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

因此，本文提出了一种融合关键词概率映射的汉越低资源跨语言摘要生成方法。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重要信息提取和跨语言对齐困难，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了对比实验，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 相关工作

针对跨语言摘要任务中数据资源稀缺，跨语言对齐困难的问题，目前解决的主要方法有基于机器翻译的方法和基于知识增强的方法。本文研究基于 Transformer 的序列到序列模型，以此为基础，融合关键词概率映射信息，概率映射为低资源下跨语言对齐做出贡献，关键词信息也为摘要的生成提供了重要线索。因此，本文融合关键词概率映射来解决低资源环境下标注数据稀缺引起的跨语言对齐问题。

Wan 等人[10]基于源语言和目标语言双方的双语信息进行跨语言文档摘要；Yao 等人[11]提出一种基于机器翻译获得对齐的双语短语来计算句子分数，并通过删除冗余或翻译不良的短语来进行摘要的方法；Ayana 等人[7]首次提出一种基于大规模语料库的神经跨语言摘要系统，基于现有的翻译系统将文本从源语言翻译成目标语言，结合源语言标题构成平行语料库训练跨语言摘要模型；Duan 等人[8]对“教师”-“学生”框架做了进一步的改进，结合了单语摘要系统以及机器翻译来训练跨语言摘要系统；Zhu 等人[6]首次提出使用往返翻译的策略来获取大规模跨语言摘要数据集，将机器翻译和单语摘要结合到跨语言摘要任务中，使用多任务学习获得了很好的效果，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；Xu 等人[12]提出基于 Transformer 的混合语言预训练方法，该方法基于跨语言任务（如翻译）和单语任务模型（如 mask 语言模型），利用大量单语数据增强其语言模型的建模能力。以上跨语言摘要的研究主要集中在如何获得大规模的跨语言摘要数据集，或者通过大规模的单语数据集，利用多任务学习等方式获得良好的跨语言摘要结果，在低资源跨语言摘要上并不适用。目前，有很少的研究是基于建立两种语言之间的关系增强跨语言对齐来获得较好的跨语言摘要效果。

近年来，基于知识增强的摘要方法是研究的热点，See 等人[13]于 2017 年提出指针生成器网络，实现了从源文本复制单词；Li 等人[14]在 2018 年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； Li 等人[9]于 2020 年提出一种将关键词的引导信号应用于序列-序列模型的编码器和解码器的生成式摘要方法，该方法采用多任务学习框架，利用关键词引导的选择性编码，及双注意力机制和双复制机制对指针生成网络进行扩展等方式，在生成式摘要上取得了很好的效果。这种基于知识增强的方法也逐渐过渡到跨语言摘要研究中，传统的方法是构建双语词典，将作为输入的源语言文本和目标语言的参考摘要通过双语词典映射至同一语义空间，然后在进行摘要。Cao 等人[15]基于 Transformer 引入对抗网络，提出了一种共同学习跨语言对齐并实现跨语言摘要的方法，该方法分别设计了有监督和无监督的跨语言摘要方法，通过增强语言间同构和跨语言迁移，在跨语言摘要任务上取得了优秀的效果。

以上研究证明了知识增强方法的有效性，也说明关键词信息在摘要中的关键引导作用。因此，对于关键词概率映射融入的研究是有必要的，且有较少的研究集中在跨语言摘要任务中。综上所述，本文认为，融合关键词概率映射信息的汉越低资源跨语言摘要研究是有意义的。

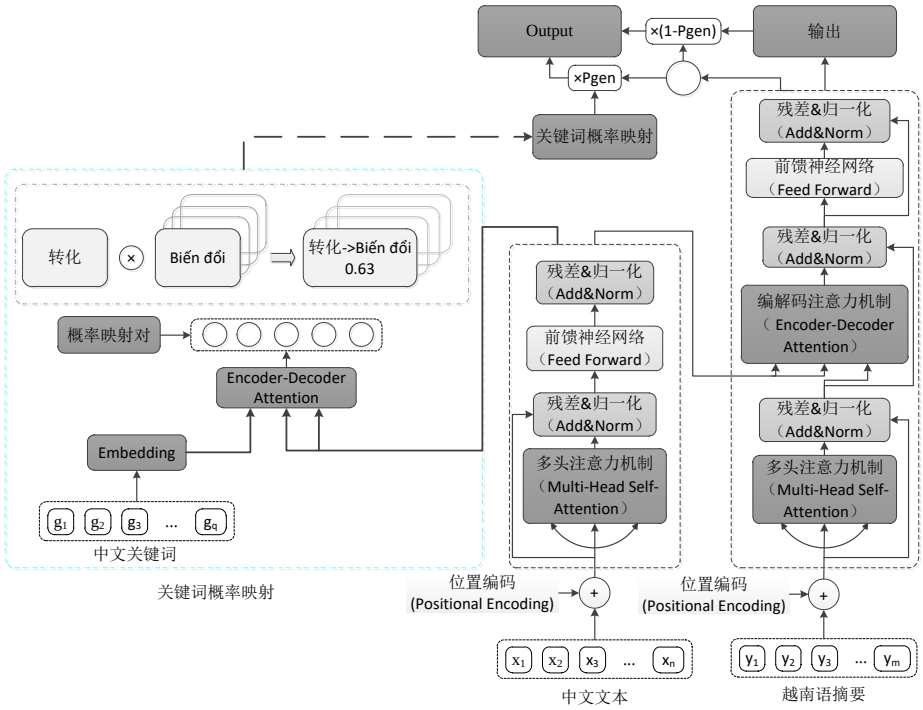

图 2 融合关键词概率映射的汉越低资源跨语言摘要方法框架

Fig. 2 Low Resource Cross-language Summarization of Chinese-Vietnamese combined with Keyword Probability Mapping

2 模型

本模型基于 Transformer 框架提出关键词概率映射（Keyword probability mapping，Kp-mapping）。本文模型包括基于 Transformer 的文本编码、融合关键词概率映射的文本表征、融合关键词概率映射的解码端。本文的整体框架如图 2 所示。

2.1 基于 Transformer 的文本编码

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

本模型基于 Transformer，需对输入序列进行词嵌入，并通过公式(1)及公式(2)进行位置编码。

$$PE_{(pos,2i)}=\sin\left(pos/10000^{2i/d_{model}}\right) \quad (1)$$

$$PE_{(pos,2i+1)}=\cos\left(pos/10000^{2i/d_{model}}\right) \quad (2)$$

其中，pos 表示每个词在输入序列中的位置信息，d_{model} 表示词向量维度，i 表示词向量的位置。

编码器由一个编解码注意力模块以及一个前馈神经网络构成。其中编解码注意力模块采用多头注意力机制，每个头对应一个点积注意力机制，由查询($\boldsymbol{Q}$)，键($\boldsymbol{K}$)和值($\boldsymbol{V}$)组成：

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=softmax\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}}\right)\boldsymbol{V} \quad (3)$$

其中 d_k 是键($\boldsymbol{K}$)的维度。

编解码注意力模块的输出通过前馈网络得到最终值：

$$MultiHead(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=Concat(head_1,head_2,...,head_h)$$

其中，$head_i=Attention(\boldsymbol{Q}\boldsymbol{W}_i^{\boldsymbol{Q}},\boldsymbol{K}\boldsymbol{W}_i^{\boldsymbol{K}},\boldsymbol{V}\boldsymbol{W}_i^{\boldsymbol{V}})$ (4)

其中，$\boldsymbol{W}_i^{\boldsymbol{Q}}$，$\boldsymbol{W}_i^{\boldsymbol{K}}$，$\boldsymbol{W}_i^{\boldsymbol{V}}$ 是学习参数矩阵，h 是头的数量。

编码端输入的源语言文本通过编码器得到输出的隐状态 Z，如式(5)所示。

$$Z=(z_1,z_2,...,z_n) \quad (5)$$

2.2 融合关键词概率映射的文本表征

本文使用的关键词抽取方法为 jieba 开源工具包提供的关键词提取算法 TextRank，TextRank[20]是一种基于图模型的关键词抽取算法。基于此算法，对每篇源语言文本提取 q 个最

重要的关键词，其中源语言输入文本序列为 X，即关键词集合 G 由式(6)所示：

$$G=\left(g_1,g_2,\cdots,g_q\right)=\text{TextRank}\left(x_1,x_2,\cdots,x_n\right) \quad (6)$$

然后，通过编解码注意力机制计算每一个关键词 g_q 对于源语言文本的注意力得分，如式(7)所示。

$$\partial=\frac{\sum_{n=0}^{n}Attention\left(g_q,z_n,z_n\right)}{n} \quad (7)$$

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上进行。其中，源语言输入序列为 $C=\left(c_1,c_2,...,c_j\right)$，其中 c_j 表示源语言输入中的一个词，目标语言输入序列为 $V=\left(v_1,v_2,...,v_r\right)$，其中 v_r 表示目标语言输入序列中的一个词，j，r 表示输入序列的长度，随输入文本的长度变化。

本文利用 Chris 等人[16]提出的 fast-align 方法，及统计的思想。汉越概率映射对的计算方法如下。

根据汉越平行语料 C，V 以及双语对应的编码，得到一个映射对，即 $v_r\rightarrow c_j$ 表示一个映射对。利用统计的思想，得到每一个 v_r 映射为 c_j 的概率 P_{MP}，如式(8)所示。

$$P_{MP}=\frac{\sum_{v_r\rightarrow c_j}v_r}{\sum c_j} \quad (8)$$

其中，c_j 表示源语言输入中的一个词，v_r 表示目标语言输入序列中的一个词。

为了将关键词映射到目标语言，本文使用了概率映射机制。利用概率映射机制，计算出每一个关键词对应其汉越映射概率对的映射概率，如式(9)所示。

$$P_{\left(\omega_1\Rightarrow\omega_2^{tgt}\right)}=Attention\left(\omega_1,\omega_2^{tgt},\omega_2^{tgt}\right) \quad (9)$$

其中，ω_1 是每个关键词的隐状态表示，作为查询，ω_2^{tgt} 是每一个关键词对应的映射候选词，作为键和值，即获得每个关键词映射到目标语言的概率分布。

2.3 融合词粒度概率映射信息的解码端

本文利用 See 等人[14]提出的指针网络，通过指针从源文本复制单词，它允许通过指针复制单词，并从固定的词汇中生成单词。

使用 O 作为解码器在时间步 t 下的隐状态，计算时间步 t 下的生成概率 P_{gen}，其中 $P_{gen}\in\left(0,1\right)$。如式(10)所示。

$$P_{gen}=\delta\left(\boldsymbol{W}_2\left(\boldsymbol{W}_1O+\boldsymbol{b}_1\right)+\boldsymbol{b}_2\right) \quad (10)$$

其中，$\boldsymbol{W}_1\in R^{d_{model}\times d_{model}}$，$\boldsymbol{W}_2\in R^{1\times d_{model}}$ 是学习矩阵，$\boldsymbol{b}_1\in R^{d_{model}}$，$\boldsymbol{b}_2\in R$ 是偏置向量，d_{model} 表示此时隐状态的维度，δ 是 sigmoid 函数。其中，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 $P\left(\omega\right)$ 如式(11)所示。

$$P\left(\omega\right)=P_{gen}\sum_{\omega_q=\omega_{src}}\partial_q P\left(\omega_{src}\rightarrow\omega\right)+\left(1-P_{gen}\right)P_N\left(\omega\right) \quad (11)$$

其中，$P\left(\omega_{src}\rightarrow\omega\right)$ 表示关键词 ω_{src} 映射到词 ω 的概率大小，$P_N\left(\omega\right)$ 表示本模型的解码端生成的词 ω 的概率大小。

本模型结合了指针网络将映射到目标语言的关键词融合到 Transformer 框架中，因此损失函数如式(12)所示。

$$Loss=-\left[\varepsilon\log\left(p_t\right)+\left(1-\varepsilon\right)\log\left(1-p_t\right)\right] \quad (12)$$

其中，p_t 表示在 t 时刻预测结果正确的概率，ε 为超参数，本文设置为 4.32E-6。

3 实验

3.1 实验数据

本文在汉越跨语言摘要数据集以及汉英跨语言摘要数据集上进行实验。本文所有数据，中文使用 jieba 分词进行处理，越南语使用 Vu 等人[17]提出的 VnCoreNLP 进行分词，英文采用其

本身的词级结构。本文的数据来自 CNN/Dailymail 以及互联网爬取，有大约 100 万基于中文的文章-摘要数据集，利用 Zhu 等人[6]提出的往返翻译的策略，分别获得了质量较高的 10 万汉越跨语言摘要数据集以及 10 万汉英跨语言摘要数据集[1]，其中有效词数为数据集文本分词去重后的剩余词数。表 1 中列举出了汉越、汉英数据集的统计信息。

表 1 数据集统计结果

Tab.1 Dataset statistics

语料库	训练集	验证集	训练集
汉越文本摘要对	9 5000	3000	2000
汉英文本摘要对	9 5000	3000	2000
中文有效词数	20 3200	3 0200	2 3600
越南文有效词数	5 3400	9800	7600
英文有效词数	7 3000	1 0000	8400

注：有效词数数据集文本分词去重后的剩余词数

3.2 评价指标

自动摘要中常用的 ROUGE 值作为评价指标，ROUGE 是一种基于召回率的相似性度量方法[18]，它通过比较候选摘要与参考摘要中共现的 n 元组 $n-gram$ 来评价候选摘要的质量。ROUGE 值的质量越高说明候选摘要的质量越好，其计算方法为：

$$\text{ROUGE-n}=\frac{\sum_{s\in R}\sum_{n-gram\in s}Count_{match}(n-gram)}{\sum_{s\in R}\sum_{n-gram\in s}Count(n-gram)} \quad (13)$$

其中，n 表示 n 元组的长度，R 表示构成的参考摘要的句子的集合，s 表示参考摘要的句子，$Count(n-gram)$ 表示句子 s 中 n 元组的数目，$Count_{match}(n-gram)$ 表示候选摘要句与参考摘要句 s 共同包含的 n 元组的数目。通过公式可以发现 ROUGE-n 反映的是参考摘要句的 n 元组的召回率。实验中本文使用 ROUGE-1，ROUGE-2,ROUGE-L 来评价参考摘要的好坏。

3.3 实验模型参数设置

本文所有实验均基于 Transformer 架构，并根据 Zhu 等人[6]采用 Adam 优化器，其中，β_1=0.9，β_2=0.998，ε=1e−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 $lr=0.1$，批次大小 $batch\_size=2048$，$dropout=0.1$，编码器和解码器层数、模型隐藏大小、前馈隐藏大小和头数分别为 6、1024、2048 和 8。本文设置编解码器词表大小分别为：中文 10 万，英文、越南语均为 1 万，未登录词使用<unk>来代替。本文实验所构概率映射词典的大小根据词频设置为 39311。本文所有实验均在单个 Nvidia RTX 2070 SUPER GPU 上进行。

3.4 基准模型

本文选择 GETran、GLTran、NCLS 3 个模型作为基准模型，所有基准模型的训练集、验证集和测试集划分均与本文相同。本文提出的 C-Vcls 模型将分别与 3 个基准模型进行比较。

1) TETran[19]：首先通过基于 Transformer 的机器翻译模型翻译源文档到目标语言，然后使用 LexRank 模型对翻译后的源文档进行摘要；

2) TLTran：首先通过基于 transformer 的单语摘要模型对源语言文本进行摘要，然后利用摘要模型将源语言的摘要翻译至目标语言；

3) NCLS[6]：基于 Transformer 框架的神经跨语言摘要模型。

3.5 实验结果分析

3.5.1 实验结果

为了证明本文融合关键词概率映射的方法在汉越低资源跨语言摘要任务上的优势，将本文模型 C-Vcls 与现有基准模型在汉越跨语言摘要数据集上进行了实验对比，表 2 列举了本文模型与基准模型在汉越跨语言摘要测试集上的 ROUGE-1（RG-1）、ROUGE-2（RG-2）、ROUGE-L（RG-L）的 F1 值比对结果。

分析表 2 可知，本文模型 C-Vcls 的三项指标均超过所有基准模型。C-Vcls 模型与基准模型

[1] https://github.com/Lxmllx/C-Vcls-dataset/tree/master

表 2 融合关键词概率映射的模型实验结果

Tab.2 Model experiment results combined with keyword probability mapping

模型	RG-1	RG-2	RG-L
TETran	16.98	7.15	15.44
TLTran	21.91	9.38	18.64
NCLS	19.16	8.56	17.32
C-Vcls（本文模型）	**23.01**	**9.45**	**20.15**

注：基准模型 TETran 为先翻译后摘要的，TLTran 为先摘要后翻译，NCLS 为端到端的跨语言摘要模型在汉越跨语言摘要数据集上的 F1 值(%)

相比，本文提出的 C-Vcls 模型较传统模型 TLTran 模型、TETran 模型在指标 RG-1、RG-2 和 RG-L 上分别取得了 1.1、0.07、1.51 和 6.03、2.3、4.71 的提升；与基于 Transformer 的 NCLS 模型相比，取得了 3.85、0.89、2.83 的提升。此外，从表 2 中可以看出，NCLS 模型，传统的 TLTran 模型与 TETran 模型均取得了不错的效果且 TLTran 模型取得了比 TETran 模型与 NCLS 模型都好的效果，原因在于模型 TETran 首先将源语言文本翻译到目标语言在进行摘要，更容易从机器翻译中带来错误。TLTran 首先进行摘要，且摘要模型 LexRank 为抽取式摘要，取得的摘要结果会更加出色。NCLS 模型是直接的端到端的模型，较 TLTran 模型效果较差，从 Zhu 等人[6]的分析可以了解到 NCLS 需要较多的语料来训练获得更佳的效果。另外，C-Vcls 模型相较端到端的 NCLS 模型在汉越跨语言摘要数据集上的结果有明显的优势，证明了融合关键词概率映射方法的有效性。

3. 5. 2 融合关键词概率映射方法的有效性分析

在 3.5.1 节中，验证了本文提出融合关键词概率映射方法的有效性。为了证明本文融合关键词概率映射模块（见图 2）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 ROUGE-1（RG-1）、ROUGE-2（RG-2）、ROUGE-L（RG-L）上的 F1 值比对结果。其中，q 的取值决定了文本提供信息量的多少，本文根据数据中参考摘要的平均长度取 q 为 5，以及无关键词提供、关键词个数较少时，模型的性能对比。

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

模型	q	RG-1	RG-2	RG-L
C-Vcls (本文模型)	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**

注：在汉越跨语言摘要数据集上的 F1 值(%)

分析表 3 可知，C-Vcls 模型在关键词个数 $q=5$ 时取得了最好的效果。与最低的相比，在汉越低资源跨语言摘要任务上分别获得了 3.85、0.89、2.83 的性能提升。当 $q=5$ 时，较 $q=2$ 时性能上获得了较大的提升，也证明了信息量越大时，对摘要的指导性越强，获得的摘要也就越可靠。也证实了关键词等重要信息对于摘要生成的指导作用。

（2）概率映射词典的有效性

为验证概率映射策略对本文模型的有效性，本文在概率映射词典的大小上进行了相关实验。根据词频的多少将概率映射词典控制在 25087、36368、39311、42399，表 4 中给出了本文模型在汉越跨语言摘要数据集上的 ROUGE-1（RG-1）、ROUGE-2（RG-2）、ROUGE-L（RG-L）的 F1 值。

分析表 4 可知，本文所使用概率映射词典大小为 39311 是性能最好的，相较概率映射词典大小为 25087 时有 6.09、2.27、4.46 的提升，较概率映射词典大小为 36368 时有 3.73、2.01、1.91 的提升，概率映射词典大小为 42339 时模型在

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

模型	概率映射词典大小	覆盖率	RG-1	RG-2	RG-L
C-Vcls (本文模型)	25087	52.37	16.92	6.67	15.68
	36368	71.27	19.28	7.44	18.24
	39311	**78.19**	**23.01**	**9.45**	**20.15**
	42339	80.32	22.98	9.46	20.11

注：在汉越跨语言摘要数据集上的 F1 值(%)，覆盖率(%)为概率映射词典对于关键词的覆盖程度

RG-1 上有 0.03 的提升，在 RG-2 上性能降低了 0.01、在 RG-L 上有 0.04 的提升。概率映射词典大小为 25087 时取得了最差的效果，本文认为原因可能在于概率词典对于关键词的覆盖率仅有 52.37%，此时词典的噪声较大，覆盖率较低，因此在进行映射时不能对关键词进行有效映射，导致一些关键词不起作用，甚至会降低摘要的效果；但是在概率映射词典为 39311、42339 时，结果表明了摘要效果的有效提升，但是两者变化不明显，本文认为原因在于其对关键词的覆盖率相差较小。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了相关实验。其中，C-Vcls 模型是本文所提融合关键词概率映射的汉越低资源跨语言摘要方法，C-Vcls-MP 模型是在 C-Vcls 模型的基础上减少了概率映射模块的模型，C-Vcls-PN 模型是在 C-Vcls 模型的基础上减少了指针网络模块的模型。

分析表 5 可知，C-Vcls 取得了最好的效果，在汉越低资源跨语言摘要任务上，相较 C-Vcls-MP 模型，提高了 4.77、4.52、3.21；相较 C-Vcls-PN 模型，提高了 2.45、2.74、2.26。在表 5 中可以看到，在关键词概率映射方法中不进行概率映射的摘要结果有明显下降，结果低于 C-Vcls 模型和 NCLS 模型，由此证明本文结合概率映射的方法是有效的。但是分析表 4、表 5，可以了解到，不加入概率映射时，较概率词典大小为 25087 的效果更好，本文认为原因在于，词典大小为 25087 时，概率映射词典对于关键词的覆盖率较低，在进行映射对齐的过程中，有过大的噪声，并不能有效帮助关键词进行映射，导致摘要效果较差。而不加入指针网络时，模型性能相较 C-Vcls 有明显下降，分析其原因是没有指针网络的加入，本文采用直接相加的方式将关键词概率映射与摘要相结合，这样会导致重复词的出现，影响摘要的性能，证明了使用指针网络、概率映射的必要性。因此，本文实验验证了概率映射、指针网络的有效性与必要性。

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	19.16	8.56	17.32
C-Vcls (本文模型)	**23.01**	**9.45**	**20.15**

注：在汉越跨语言摘要数据集上的 F1 值(%)

（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

为了验证本文所提出模型的泛化性，本文在

汉英跨语言摘要数据集上进行了实验。表 6 中给出了本文模型 C-Vcls 与基准模型在汉英跨语言摘要数据集上 ROUGE-1（RG-1）、ROUGE-2（RG-2）、ROUGE-L（RG-L）的 F1 值。

表 6 汉英跨语言摘要数据集实验结果的 F1 值(%)

Tab.6 The F1 value (%) of the experimental results of the Chinese-English cross-language summary dataset set

模型	RG-1	RG-2	RG-L
TETran	15.15	4.18	14.10
TLTran	19.81	7.34	16.37
NCLS	16.07	4.34	15.69
C-Vcls (本文模型)	**21.37**	**8.01**	**18.67**

分析表 6 可知，本文模型 C-Vcls 的三项指标均超过所有基准模型。C-Vcls 模型与基准模型相比，本文提出的 C-Vcls 模型 TLTran 模型、TETran 模型在指标 RG-1、RG-2 和 RG-L 上分别取得了 1.56、0.67、2.3 和 6.22、3.83、4.57 的提升；与 Transformer 直接生成摘要的 NCLS 模型相比，取得了 5.3、3.67、2.98 的提升。根据表 2、表 6，可以看到同样数量级的数据在同样的基准模型上，在不同的数据集上取得了不同的效果。在汉英跨语言摘要数据集上较汉越跨语言摘要数据集上取得的 F1 的分数是较低的。本文认为原因可能在于实验设置时，越南语和英文构造的词典均为 1 万，根据越南语和英文文本构造特点，以及本文在构造数据集时展示的有效词数，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，提高了摘要的准确性。但是，依然可以从实验结果看出，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中-英跨语言摘要任务是有效的，也证明了本文所提模型的泛化性。

3.6 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。原文与标准摘要都来自中文-越南语摘要数据集。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，本文给出了其中文的翻译结果。

表 7 不同模型生成摘要样例

Tab.7 Sample summary of different models

名称	内容
原文	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	15 người bạn đi du lịch Bắc Kinh bị mắc kẹt ở Trương Gia Khẩu 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 (本文模型)	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）

分析表 7 可知，原文主要讲述 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，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练习关键信息；TLTran 模型是表现较好的模型，但是仍然没有表输出“张家口”的关键事实；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”的主要信息，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，因为有引导性信息的融入可以很好的捕捉到 15 名旅游失去联系的这一主要信息，从而提高了摘要的信息覆盖度，生成质量更高的文本摘要。

4 结论

本文针对汉越低资源跨语言摘要，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重要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。因此，在下一步研究中，拟在低资源跨语言摘要任务中利用关键词的特性，通过分析关键词之间的关联关系，将其利用图结构更好的融入到模型中指导摘要生成。

参考文献：

[1] Leuski A , Lin C Y , Liang Z , et al. Cross-lingual C*ST*RD: English access to Hindi information[J]. ACM Transactions on Asian Language Information Processing (TALIP).2003:245–269

[2] Jessica Ouyang, Boya Song, Kathy McKeown. A robust abstractive system for cross-lingual summarization[J]. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), 2019, pages 2025–2031.

[3] Lim J M, Kang I S, Lee J H. Multi-Document Summarization Using Cross-Language Texts[C]// NTCIR. In Proceedings of the F ourth NTCIR Workshop on Research in Information Access Technologies Information Retrieval, Question Answering and Summarization (NTCIR),2004.

[4] Orasan C, Chiorean O A. Evaluation of a Cross-lingual Romanian-English Multi-document Summariser. [C] // International Conference on Language Resources & Evaluation. 2008.

[5] Wan X, Li H, Xiao J. Cross-Language Document Summarization Based on Machine Translation Quality Prediction[C]// ACL. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, July 11-16, 2010, Uppsala, Sweden. Association for Computational Linguistics.2010:917–926.

[6] Zhu J, Wang Q, Wang Y, et al. NCLS: Neural Cross-Lingual Summarization[C]//EMNLP-IJCNLP. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP).2019:3045–3055.

[7] Ayana, Shen S Q, Yun C, et al. Zero-Shot Cross-Lingual Neural Headline Generation[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, PP(99):1-1.

[8] Duan X Y, Yin M M, Zhang M, et al. Zero-shot cross-lingual abstractive sentence summarization through teaching generation and attention[C]//ACL. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), 2019:3162–3172.

[9] Li H, Zhu J, Zhang J J. Keywords-Guided Abstractive Sentence Summarization[C]//AAAI. 2020: 8196-8203

[10] Wan X J. Using bilingual information for cross-language document summarization[C]//ACL. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL), 2011:1546–1555.

[11] Yao J G, Wan X J, Xiao J G. Phrase-based compressive cross-language summarization[C]//EMNLP. In Proceedings of the 2015 conference on empirical

methods in natural language processing (EMNLP), 2015:118–127.

[12] Xu R, Zhu C, Shi Y, et al. Mixed-Lingual Pre-training for Cross-lingual Summarization[J]. 2020.

[13] See A, Liu P J, CD Manning. Get To The Point: Summarization with Pointer-Generator Networks[C]// ACL. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017.

[14] H Li, Zhu J, Zhang J, et al. Ensure the Correctness of the Summary: Incorporate Entailment Knowledge into Abstractive Sentence Summarization[C]//COLING. In Proceedings of the 27th International Conference on Computational Linguistics (COLING), 2018：pages 1430–1441.

[15] Cao Y, Liu H, Wan X J. Jointly Learning to Align and Summarize for Neural Cross-Lingual Summarization. [C]//ACL. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.2020: 6220–6231.

[16] Dyer C, Chahuneau V, Smith N A. A Simple, Fast, and Effective Reparameterization of IBM Model 2[C]// NAACL. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT),2013:644-648.

[17] Vu T, Nguyen D Q, Dai Q N, et al. VnCoreNLP: A Vietnamese Natural Language Processing Toolkit[J]. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, NAACL 2018, 56-60.

[18] Lin C Y. ROUGE: a package for automatic evaluation of summaries[C]//ACL. Proceedings of the Workshop on Text Summarization Branches Out, Barcelona, Jul 21-26, 2004. Stroudsburg: ACL, 2004: 74-81.

[19] Erkan G, Radev D R. LexRank: Graph-based Lexical Centrality as Salience in Text Summarization[J]. Journal of Artificial Intelligence Research (JAIR), 2004, 22:457–479.

[20] Mihalcea R, Tarau P. TextRank:bringing order into texts[C]//In:Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. Barcelona, Spain:UNT Scholarly Works, 2004: 404-411.

Low Resource Cross-language Summarization of Chinese-Vietnamese combined with Keyword Probability Mapping

Xiaomeng Li[1,2],Yafei Zhang[1,2*],Junjun Guo[1,2],Shengxiang Gao[1,2],Zhengtao Yu[1,2]

(1. Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500,China; 2. Key Laboratory of Artificial Intelligence of Yunnan Province, Kunming 650500,China)

Abstract: Annotation data for cross-language summarization tasks in low-resource scenarios is scarce, and it is difficult to achieve cross-language semantic alignment based on small-scale alignment data. In view of this, 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, and then map the source language keywords to the target language based on the probability mapping pair, finally integrate the mapped target language keywords into the abstract generation process based on the pointer network. The experimental results on the constructed Chinese-Vietnamese cross-language abstract data set show that compared 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