低资源场景下基于联合训练与自训练的跨语言摘要方法

程绍欢，唐煜佳，刘 峤，陈文宇*

（电子科技大学 计算机科学与工程学院，成都 611731）

摘要 随着全球化的不断发展，跨语言摘要任务已成为自然语言处理领域的一项重要课题。在低资源场景下，现有方法存在表征转换受限和数据利用不充分等问题。为此，该文提出了一种基于联合训练与自训练的跨语言摘要方法。该方法使用两个模型分别建模翻译任务和跨语言摘要任务，以统一输出端的语言向量空间，从而避免模型间表征转换受限的问题。此外，通过对齐平行训练对的输出特征和概率进行联合训练，增强模型间的语义共享。同时，在联合训练的基础上引入自训练技术，利用额外的单语摘要数据生成合成数据，有效缓解了低资源场景下数据稀缺的问题。实验结果表明，该方法在多个低资源场景下均优于现有对比方法，实现了 ROUGE 分数的显著提升。

关键词 跨语言摘要；联合训练；低资源场景；机器翻译；自训练

中图分类号 TP391.1 **文献标志码** A DOI 10.12178/1001-0548.2024173

Cross-Lingual Summarization Method Based on Joint Training and Self-Training in Low-Resource Scenarios

CHENG Shaohuan, TANG Yujia, LIU Qiao, and CHEN Wenyu*

(School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China)

Abstract As globalization continues to develop, cross-lingual summarization has become an important topic in natural language processing. In low-resource scenarios, existing methods face challenges such as limited representation transfer and insufficient data utilization. To address these issues, this paper proposes a novel method based on joint training and self-training. Specifically, two models are used to handle the translation and cross-lingual summarization tasks, respectively, which unify the language vector space of the output and avoid the issue of limited representation transfer. Additionally, joint training is performed by aligning the output features and probabilities of parallel training pairs, thereby enhancing semantic sharing between the models. Furthermore, based on joint training, a self-training technique is introduced to generate synthetic data from additional monolingual summary data, effectively mitigating the data scarcity issue of low-resource scenarios. Experimental results demonstrate that this method outperforms existing approaches in multiple low-resource scenarios, achieving significant improvements in ROUGE scores.

Key words cross-lingual summarization; joint training; low-resource scenarios; machine translation; self-training

跨语言摘要（Cross-lingual Summarization, CLS）任务，作为自然语言处理领域的一个重要研究方向，旨在解决从一种语言文档自动生成另一种语言摘要的问题。在全球化快速发展的背景下，语言障碍成为国际交流和信息传播的主要瓶颈。跨语言摘要作为解决多语言信息获取障碍的重要技术，通过在不同语言间生成简洁明了的摘要，帮助用户快速获取并理解外语信息，在多语种社区和国际新闻报道中发挥着至关重要的作用[1]。

目前的跨语言摘要研究可以分为高资源场景[2-4]和低资源场景[5-7]两类。在高资源场景下，模型可以利用大量的文档和摘要对进行训练。然而，在实际应用中，由于跨语言摘要任务需要标注专家同时精通两种语言，往往难以获取大量高质量的训练数据，尤其是对于较少使用的语言。因此，本文专注于低资源场景下的跨语言摘要研究，旨在探索在跨

收稿日期：2024−07−12；修回日期：2024−08−08
基金项目：国家自然科学基金企业联合基金重点项目（U22B2061）
作者简介：程绍欢，博士生，主要从事自然语言处理方面的研究。
*通信作者 E-mail: cwy@uestc.edu.cn

语言摘要对有限的情况下提高跨语言摘要模型（CLS 模型）性能的方法。

跨语言摘要数据集通常基于已有的单语摘要数据集，将单语摘要翻译成另一种语言来获得目标摘要，跨语言摘要数据通常为包含单语摘要的三元组。现有的针对低资源场景的跨语言摘要研究[5-8]基于这一特性，使用多任务学习来利用其中的单语摘要。

文献 [5] 让解码器依次生成单语摘要和跨语言摘要，通过自注意力和交叉注意力机制同时获取源文档和单语摘要的信息，从而提升跨语言摘要的质量。然而，由于不同语言之间可能存在形态和结构差异，如从中文到英文，使用一个解码器难以捕捉这种差异，因此容易导致模型表现不佳[9]。文献 [6] 从跨语言摘要三元组中提取出单语摘要对和跨语言摘要对，使用教师模型建模单语摘要任务，学生模型建模跨语言摘要任务。通过知识蒸馏的方式，让学生模型学习教师模型对平行的单语摘要的表征，从而提升自身的表征能力。然而，该设置下两个模型的输出端语言不一致，导致在不同语言向量空间中转换句子级表征时会损失大量语义信息，影响学生模型的学习效果。此外，这些方法在低资源设置下借助额外的单语摘要数据对整个或部分模型预训练，并未充分利用这部分数据，对 CLS 模型的提升效果不显著。

针对上述问题，本文设计了基于联合训练和自训练的跨语言摘要方法，实现对有限的跨语言摘要数据和额外单语摘要数据的高效利用，从而显著提升 CLS 模型的性能。首先，从跨语言摘要三元组数据中提取翻译对（即单语摘要和跨语言摘要）和跨语言摘要对（即源文档和跨语言摘要）作为平行训练对，使用两个模型来分别建模翻译（Machine Translation, MT）任务和跨语言摘要任务。其次，在输出端的特征层和概率层引入两个损失项，对齐两个模型在每个时间步上的输出特征和概率分布，从而在无须转化表征的情况下实现模型间知识共享。最后，使用训练后的翻译模型（MT 模型）对额外的单语摘要数据进行翻译，生成跨语言摘要，以扩充原有限的跨语言摘要数据，并继续使用双模型联合训练的方式学习扩充后的数据。

本文所提方法使得两个模型的输出语言匹配，且目标文本完全一致，便于通过平行训练对实现无障碍的信息交互。此外，双模型的设计使得能利用 MT 模型生成伪标签文本，从而以自训练的方式进一步提升模型性能。在两个基准数据集上的实验结果证明了本文方法的有效性。

1　相关工作

本节对所提算法的相关工作进行简要介绍，主要包括跨语言摘要和自训练技术两部分。

1.1　跨语言摘要

跨语言摘要任务需要同时进行文本翻译[10]和摘要生成[11]。早期的方法通常采用流水线策略，按照“先翻译后摘要”或“先摘要后翻译”的顺序执行[12-13]。由于这两个阶段通常是独立开发和优化的，导致信息无法在不同阶段间传递，因此存在误差传播的问题。近年来，文献 [14] 为跨语言摘要任务创建了两个基准数据集，并使用端到端的方法训练模型，相比流水线方法在性能上取得了显著提升。此外，还提出了利用多任务学习结合单语摘要或翻译数据的两种方法，启发了后续的研究工作[1]。

这些研究主要集中在高资源场景中[2-4,15]，少数工作关注低资源场景[5-8,16-17]。由于跨语言摘要任务的复杂性，这些研究通常利用额外数据或外部知识来提升模型鲁棒性。如，文献 [5,7] 利用跨语言摘要三元组中的单语摘要增强模型能力；文献 [4] 基于变分自编码器，使用额外的摘要和翻译数据，提升模型的层级归纳能力；文献 [15] 利用图网络建模实体关系，提升模型处理不常见实体的准确性。本文针对跨语言摘要三元组数据的特点，设计了基于双模型协调的联合训练方法，并结合自训练技术充分利用额外的单语摘要数据，以应对跨语言摘要数据的固有稀缺性。

1.2　自训练技术

自训练技术通过生成伪标签[18]的方式，充分利用大量未标注数据扩展训练数据集，从而提升模型的泛化能力。近年来，自训练技术在包括自然语言处理的多个领域得到了广泛应用[19-22]。文献 [19] 将自训练应用于多标签分类问题，以缓解缺失标签带来的误导影响。文献 [20] 探讨了如何在文本生成任务中有效应用自训练技术，并通过引入噪声训练进一步提升文本生成质量。文献 [21] 结合预训练和自训练，并改进文本相似度函数以筛选出高质量的无标签法律文本数据来生成伪标签，从而提高模型准确抽取法律文本的能力。然而，自训练技术在跨语言摘要中的应用尚未得到充分探索。本文在有限的跨语言摘要数据的基础上，同时训练两个模型，利用较为准确的 MT 模型生成伪标签文本扩充

原有数据，从而显著提升 CLS 模型的鲁棒性和准确性。

2 模型与方法

本节对具体方法进行详细描述，方法框架图如图 1 所示。整个训练过程遵循自训练的方式。首先，使用跨语言摘要三元组数据对 MT 模型和 CLS 模型进行联合训练。接着，使用 MT 模型对额外的单语摘要数据生成伪标签文本，合成新的伪标签数据。最后，将这些伪标签数据与原有数据混合，再次进行两个模型的联合训练。

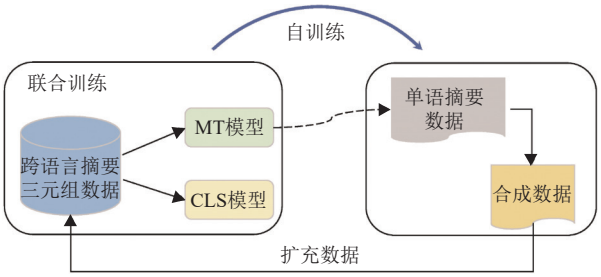

图 1 方法框架示意图

2.1 模型架构

由于本文研究低资源场景下的跨语言摘要任务，因此采用预训练的多语言模型 mBART[23] 进行模型构建和参数初始化。如图 2 所示，mBART 构建于标准的 Transformer 架构之上[24]，包含一个双向编码器和一个单向解码器。其中，编码器部分负责深入理解输入文本的内容，通过双向特征编码捕捉文本的全局信息；而解码器则利用交叉多头注意力机制，动态整合编码器最后一层的隐藏状态信息，以自回归的方式逐步解码生成摘要文本。mBART 遵循 BART 训练范式[25]，添加语言标识符（如“<En>”）使得模型能在大规模多语言语料上预训练。因此，mBART 不仅继承了 BART 在文本理解和生成方面的能力，还具备了理解多种语言的能力，为低资源条件下的跨语言摘要提供了良好的初始性能。

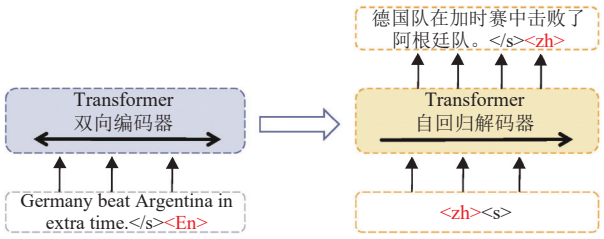

图 2 mBART 应用于翻译任务上的示意图

2.2 双模型协同的联合训练

跨语言摘要任务的数据通常由三元组构成，即（源文档D，单语摘要S，跨语言摘要Y）。本文将其中的单语摘要和跨语言摘要作为翻译对，将源文档和跨语言摘要作为跨语言摘要对，使用两个模型分别建模翻译（MT）任务和跨语言摘要（CLS）任务。通过这种配对方式，两个模型的训练目标完全一致，即同为跨语言摘要Y。通过对齐二者在每个时间步上的输出特征和输出概率来增强该平行训练对之间的语义共享，提升跨语言摘要模型的性能，模型训练示意图如图 3 所示。

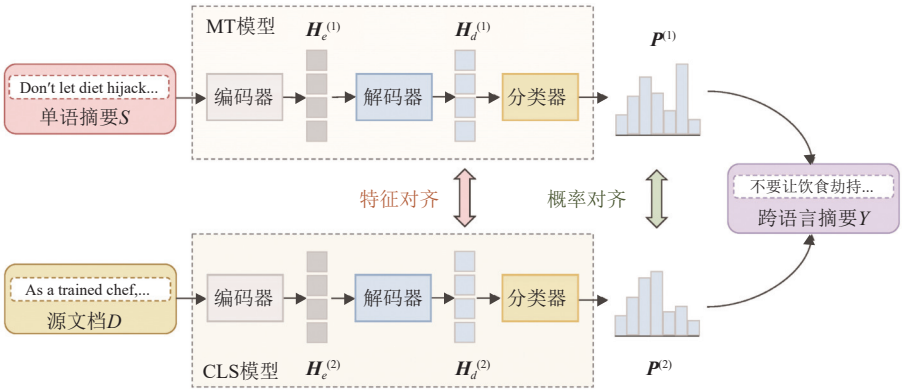

图 3 双模型协同的联合训练示意图

具体来说，MT 模型接收输入S，通过编码器编码后得到句子表征向量$\boldsymbol{H}_e^{(1)}$并传递给解码器，解码器在时刻t接收来自之前时刻的目标输入$Y_{1:t-1}$，并通过交叉注意力机制对$\boldsymbol{H}_e^{(1)}$进行动态组合，以获得当前时刻需要关注和生成的信息。将解码器在时刻t的输出特征表示为$\boldsymbol{H}_{d,t}^{(1)}$，则$\boldsymbol{P}_t^{(1)}$是$\boldsymbol{H}_{d,t}^{(1)}$经过转换后的概率分布。

同样地，CLS 模型接收输入D，编码为$\boldsymbol{H}_e^{(2)}$后传递给解码器，解码器在时刻t接收来自之前时刻的目标输入$Y_{1:t-1}$并通过交叉注意力获取动态的编

码器特征表示，得到每个时刻的输出特征$\boldsymbol{H}_{d,t}^{(2)}$和输出概率$\boldsymbol{P}_t^{(2)}$。其中$t\in[1,T]$，$T$代表目标摘要$Y$的长度。

损失函数包括 3 部分。首先，对两个模型应用交叉熵损失，使其都能从目标Y中学习，具体公式为：

$$\mathcal{L}_{\text{ce}}=\frac{1}{2T}\left(\sum_{t=1}^{T}\text{CE}(\boldsymbol{P}_t^{(1)},\boldsymbol{y}_t)+\sum_{t=1}^{T}\text{CE}(\boldsymbol{P}_t^{(2)},\boldsymbol{y}_t)\right)\tag{1}$$

式中，$\boldsymbol{y}_t$是Y在时刻t的令牌的独热编码。

此外，联合训练通过对齐两个模型的输出特征和输出概率来增强平行的翻译对和跨语言摘要对之间的联系，使得两个模型能从不同的输入表征中获得更丰富和准确的语义信息。其中，概率对齐损失近似两个模型输出概率的双向 KL 散度，以鼓励模型在接收语义近似的输入文本时输出相同的分布，具体为：

$$\mathcal{L}_{\text{kl}}=\frac{1}{2T}\sum_{t=1}^{T}\left(\mathcal{D}_{\text{KL}}(\boldsymbol{P}_t^{(1)}\parallel\boldsymbol{P}_t^{(2)})+\mathcal{D}_{\text{KL}}(\boldsymbol{P}_t^{(2)}\parallel\boldsymbol{P}_t^{(1)})\right)\tag{2}$$

特征对齐损失通过计算两个模型输出特征的均方差得到，使得 CLS 模型通过交叉注意力机制获得的融合注意力表征与 MT 模型获得的注意力表征相似，从而促使 CLS 模型能准确关注冗长的输入文本中的关键信息，同时 MT 模型也能由此捕获更为丰富的上下文信息。具体为：

$$\mathcal{L}_{\text{mse}}=\frac{1}{T}\sum_{t=1}^{T}\left\|\boldsymbol{H}_{d,t}^{(1)}-\boldsymbol{H}_{d,t}^{(2)}\right\|_2^2\tag{3}$$

最后，联合训练损失函数$\mathcal{L}$由以上 3 部分损失构成，具体为：

$$\mathcal{L}=\mathcal{L}_{\text{ce}}+\mathcal{L}_{\text{kl}}+\mathcal{L}_{\text{mse}}\tag{4}$$

通过双模型协同的方式，在输出特征和概率层级上将 MT 和 CLS 模型进行联合训练，使其在建模不同任务的同时，能通过平行训练对来交流各自的内部知识和动态表征，从而获得比单一模型更好的翻译和跨语言摘要能力。

2.3　融合单语摘要数据的自训练

相较于跨语言摘要数据，单语摘要数据无须标注专家同时精通两种语言，因此更为容易获取。在有限的跨语言摘要数据下，基于本文提出的联合训练方式结合自训练方式可以高效地利用额外的单语摘要数据进一步提升模型性能。如图 4 所示，将自训练过程分为以下 3 个步骤。

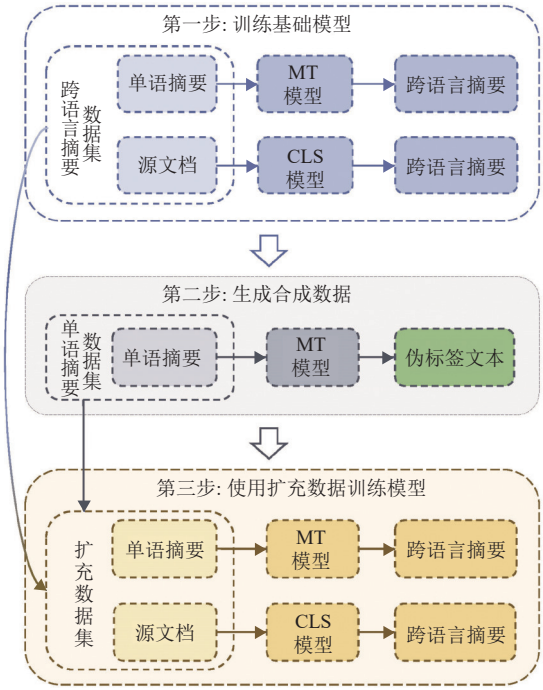

图 4　自训练过程示意图

1）训练基础模型

使用跨语言摘要数据$\mathcal{D}_{\text{cls}}=\{(D_i,S_i,Y_i)\}_{i=1}^{N}$用联合训练方式得到基础 MT 和 CLS 模型。其中，N是跨语言摘要样本的数量。

2）生成合成数据

对于额外的单语摘要数据$\mathcal{D}_{\text{ms}}=\{(D_i',S_i')\}_{i=1}^{M}$，使用基础 MT 模型生成每个样本$S_i'$对应的翻译$Y_i'$，同时$Y_i'$也是$D_i'$的跨语言摘要。因此，可以构建一个合成的跨语言摘要数据集，以此作为对原始跨语言摘要数据集的扩充。其中，M代表额外单语摘要样本的数量。

选择基础 MT 模型来生成伪标签数据，是因为 CLS 模型需要同时掌握摘要生成和翻译两种能力，而 MT 模型专注于翻译任务，能够产生更为准确和流畅的伪标签文本。这一点将在后续的消融实验中得到验证。

3）使用扩充数据训练

将原有的跨语言摘要数据集$\mathcal{D}_{\text{cls}}$与合成的跨语言摘要数据集$\mathcal{D}_{\text{syn}}$合并，形成新的跨语言摘要三元组数据集。随后，基于第一步训练得到的基础模型，继续联合训练直至模型性能不再显著提升。在训练过程中，原始数据确保了模型在训练过程中不会过分偏离真实的跨语言摘要分布，而合成数据引入额外的多样性和复杂性，进一步增强了模型的泛化能力。

本文方法利用跨语言摘要三元组的特点，在联合训练阶段同时训练 MT 模型和 CLS 模型，使二者能够相互促进，并获得质量较高的文本生成能力。在自训练阶段，利用训练得到的 MT 模型扩充跨语言摘要数据，继续进行联合训练，进一步提升模型性能。该设计利用了人工标注的跨语言摘要数据，并在此基础上实现了对额外单语摘要数据的自动标注，减轻了标注成本，同时获得了显著的性能提升。

3 实验与分析

3.1 数据集

本文选取两个基准跨语言摘要数据集 Zh2EnSum 和 En2ZhSum 进行实验。这两个数据集由文献 [14] 基于现有的单语摘要数据集生成，因此每个训练样本包含源文档、单语摘要和跨语言摘要这 3 部分。其中，Zh2EnSum 的训练集、验证集和测试集的样本数分别为 1 693 713、3 000、3 000，En2ZhSum 的样本数分别为 364 687、3 000、3 000。遵循文献 [5-6] 的设置，从两个训练集中随机选取不同比例的样本构成少量、中等、大量 3 种低资源场景，剩余样本去除跨语言摘要后，作为额外的单语摘要数据。每种场景的具体训练样本数和比例如表 1 所示。

表 1 不同低资源场景下的样本数据量 个

场景	Zh2EnSum	En2ZhSum
少量	5 000 (0.3%)	1 500 (0.4%)
中量	25 000 (1.5%)	7 500 (2.0%)
大量	50 000 (3.0%)	15 000 (4.0%)
总数	1 693 713	364 687

3.2 评价指标

本文采用常用于自动文本摘要的 ROUGE[26] 指标作为摘要质量的评价指标，并选取 ROUGE-1、ROUGE-2 和 ROUGE-L 进行测评。其中，ROUGE-1 和 ROUGE-2 基于 n 元语法（n-gram）通过计算系统摘要和参考摘要之间的重叠程度来计算分数，ROUGE-L 则通过最长公共子序列的重叠程度来计算，具体计算公式为：

$$\text{ROUGE-N} = \frac{\sum_{\text{gram}_n \in R} \text{Count}_{\text{match}}(\text{gram}_n)}{\sum_{\text{gram}_n \in R} \text{Count}(\text{gram}_n)} \tag{5}$$

$$\text{ROUGE-L} = \frac{(1+\beta^2)R_{\text{lcs}}P_{\text{lcs}}}{R_{\text{lcs}}+\beta^2 P_{\text{lcs}}} \tag{6}$$

$$R_{\text{lcs}} = \frac{\text{LCS}(R,G)}{\text{len}(R)} \tag{7}$$

$$P_{\text{lcs}} = \frac{\text{LCS}(R,G)}{\text{len}(G)} \tag{8}$$

式中，$\text{Count}_{\text{match}}(\text{gram}_n)$表示系统摘要和参考摘要重叠的$n$元语法个数；$\text{Count}(\text{gram}_n)$表示参考摘要中$n$元语法个数；$\text{LCS}(R,G)$表示系统摘要和参考摘要之间最长公共子序列的长度；len表示摘要长度。测评时，使用版本为 1.5.5 的官方 ROUGE 脚本来计算 ROUGE 分数，脚本参数设为“-c 95 -r 1000 -n 2 -a”。

3.3 实验设置

实验环境设置如下：环境搭配采用 Linux 系统，CPU 型号为 AMD EPYC 7282，显卡型号为 Nvidia GeForce A100。编程语言为 python3.10，模型框架为 PyTorch Lightning 1.9.4。

实验参数设置如下：遵循文献 [7] 的设置，采用多语言预训练模型 mBART[23] 初始化模型参数，优化器使用 AdamW[27]，学习率设为 5×10^{-5}，输入和输出的最大文本长度分别设为 768 和 128。根据显卡最大容量，设置数据集 Zh2EnSum 的批大小为 10，而 En2ZhSum 平均句子长度较长，因此批大小设为 2 且累计 5 个批次后更新 1 次梯度。训练阶段，对验证集的 ROUGE-2 分数进行监测，如果连续 3 轮分数都未能继续提升，则停止训练并使用在验证集上分数最好的模型进行测试。推理阶段，采用束大小为 4 的束搜索来生成系统摘要，并将非重复元组大小设为 3 来防止生成重复序列。

3.4 整体评价和对比分析

为了验证本文算法在低资源场景下的跨语言摘要能力，选择 6 种适合低资源场景的算法作为对比模型，列举如下。

1）mBERT-CLS：使用文献 [14] 提出的端到端训练方法直接在跨语言摘要对上训练，模型框架使用预训练 mBERT 模型[28] 作为编码器，初始化的 Transformer 作为解码器，作为其他使用 mBERT 模型的对比方法的基线。

2）MCLAS[5]：模型框架与 mBERT-CLS 相同，训练时让解码器依次生成单语摘要和跨语言摘要从而加强不同语言摘要之间的对齐关系。使用整个数据集上的单语摘要数据进行预训练后，再用给定的

跨语言摘要三元组数据进行训练。

3）KD[6]：模型框架与 mBERT-CLS 相同，但使用两个独立模型分别负责单语摘要和跨语言摘要任务，并用知识蒸馏方式获取单语摘要教师模型的知识。该方法先用整个数据集上的单语摘要数据预训练教师模型，再用给定的跨语言摘要三元组数据进行训练。

4）mBART-CLS：使用文献 [14] 提出的端到端训练方法直接在跨语言摘要对上训练，模型框架使用端到端预训练的 mBART 模型[23]，作为其他使用 mBART 模型的对比方法的基线。

5）mBART+MS：同样基于 mBART 模型，使用由文献 [14] 提出的多任务训练方法。该方法使用两个解码器来分别负责单语摘要任务和跨语言摘要任务，训练数据为跨语言摘要三元组数据。

6）TFLCLS[7]：基于 mBART 模型进行两阶段训练。其中，阶段一使用给定的单语摘要和跨语言摘要构成的翻译对进行训练，阶段二额外引入一个解码器负责跨语言摘要任务并改为使用三元组数据进行多任务训练。

对于所提算法，分别展示了在联合训练后（Ours-1）和继续自训练后（Ours-2）的结果。所有算法的实验结果如表 2 所示，除了两个基线模型 mBERT-CLS 和 mBART-CLS 外，还展示了每个模型相较于其基线模型的平均性能提升，记为 Avg-I，其中，加粗的分数代表最好的结果，Avg-I 代表每个方法较其基线模型的平均提升性能。

表 2　不同低资源场景下的实验结果

%

场景	方法	Zh2EnSum				En2ZhSum			
		R-1	R-2	R-L	Avg-I	R-1	R-2	R-L	Avg-I
少量	mBERT-CLS	20.93	5.88	17.58	–	34.14	12.45	21.20	–
	MCLAS	21.03	6.03	18.16	0.28↑	32.03	13.17	21.17	0.47↓
	KD	22.37	6.50	18.47	0.98↑	35.59	13.77	22.56	1.38↑
	mBART-CLS	24.61	7.24	20.21	–	32.87	13.25	29.28	–
	mBART+MS	24.60	7.39	20.29	0.07↑	32.14	12.66	28.35	0.75↓
	TFLCLS	25.17	7.68	20.72	0.50↑	33.22	13.79	29.67	0.43↑
	Ours-1	25.76	8.08	21.51	1.10↑	35.52	14.14	30.04	0.77↑
	Ours-2	**33.21**	**13.48**	**28.51**	**7.71↑**	**36.04**	**16.63**	**32.13**	**3.13↑**
中等	mBERT-CLS	26.42	8.90	22.05	–	35.98	15.88	23.79	–
	MCLAS	27.84	10.41	24.12	1.67↑	37.28	18.10	25.26	1.66↑
	KD	27.97	11.51	27.16	3.09↑	40.30	20.01	25.79	3.48↑
	mBART-CLS	29.18	10.40	24.36	–	36.46	17.22	32.46	–
	mBART+MS	28.84	10.12	24.05	0.31↓	36.35	17.10	32.29	0.13↓
	TFLCLS	29.76	10.83	25.29	0.65↑	37.51	18.13	33.45	0.98↑
	Ours-1	30.86	11.93	25.92	1.59↑	38.46	19.17	34.57	2.02↑
	Ours-2	**42.15**	**23.81**	**38.24**	**13.42↑**	**43.78**	**25.76**	**39.96**	**7.79↑**
大量	mBERT-CLS	29.05	10.88	24.32	–	40.18	19.86	26.52	–
	MCLAS	30.73	12.26	26.51	1.75↑	38.35	19.75	26.41	0.68↓
	KD	31.08	12.70	27.16	2.23↑	41.24	20.01	27.06	0.58↑
	mBART-CLS	31.43	12.25	26.59	–	37.05	18.45	33.18	–
	mBART+MS	31.17	11.91	25.97	0.41↓	37.59	18.42	33.62	0.32↑
	TFLCLS	32.14	12.81	27.22	0.63↑	38.28	19.54	34.26	1.13↑
	Ours-1	32.77	13.55	27.56	1.20↑	39.39	20.50	35.26	2.16↑
	Ours-2	**43.59**	**24.13**	**39.62**	**12.36↑**	**44.52**	**26.38**	**40.25**	**7.49↑**

通过表 2 可以得出以下结果。

1）mBART-CLS 较 mBERT-CLS 在大部分 ROUGE 分数上均有提升，表明使用端到端预训练的模型更有利于跨语言摘要任务。

2）mBART+MS 在 2/3 的场景中的平均得分均低于其基线 mBART-CLS，说明简单地添加模型结构来利用三元组中的单语摘要并不会带来明显的性能收益。

3）在未使用额外单语摘要数据的方法中（mBART+MS、TFLCLS 和 Ours-1），Ours-1 在

所有场景中均获得了最多的平均性能提升。此外，在1/3 的场景中，Ours-1 的平均性能提升甚至优于使用额外单语摘要数据进行预训练的方法（MCLAS 和 KD）。这表明，统一输出端语言并对平行语料对的输出进行近似对齐的训练方式，能够实现对跨语言摘要三元组数据的高效利用，从而极大地提升了 CLS 模型在翻译和摘要生成方面的能力。

4）在所有使用额外单语摘要数据的方法中（MCLAS、KD 和 Ours-2），Ours-2 在所有场景下均获得了最多的平均性能提升，且提升极为显著。如，在 Zh2EnSum 的中等场景中，Ours-2 的平均性能提升比 KD 高出 10.33 个百分点。这表明，在相同数据量的条件下，相较于仅利用额外单语数据进行预训练的方式，结合联合训练和自训练的方式更有利于吸收额外数据带来的知识，从而能够显著提升 CLS 模型的性能。

3.5 消融实验

由于所提算法涉及两个阶段，因此分别对两个阶段进行消融实验，从而更准确地判断每个模块性能。本节消融实验均在 Zh2EnSum 的中等资源场景下进行。

在联合训练阶段，引入 3 个对比模型，分别为移除概率对齐损失（M1）、移除特征对齐损失（M2）以及全部移除（M3，此时等同于 mBART-CLS 方法）。实验结果如表 3 所示。从表中可以看出，移除不同模块后，模型性能均有所下降，表明所有模块均具有正向效用。此外，移除特征对齐损失（M2）导致的性能下降幅度较移除概率对齐损失（M1）更大，说明对齐具有相同输出的平行语料的特征向量能够更好地促进信息共享。

表 3 联合训练阶段的消融实验结果 %

方法	R-1	R-2	R-L
Ours-1	30.86	11.93	25.92
M1	30.21(0.65↓)	11.48(0.45↓)	25.58(0.34↓)
M2	29.32(1.54↓)	10.58(1.35↓)	24.62(1.30↓)
M3	29.18(1.68↓)	10.40(1.53↓)	24.36(1.56↓)

在自训练阶段，引入 3 个对比模型。首先，使用三元组中的跨语言摘要对训练一个 CLS 模型，并对额外的单语摘要生成伪标签，扩充数据后在此基础上继续训练，记为 M4。其次，用三元组中的翻译对训练一个 MT 模型生成伪标签，然后使用扩充数据中的跨语言摘要对训练一个 CLS 模型，记为 M5。最后，使用联合训练的 MT 模型生成伪标签，并使用扩充数据中的跨语言摘要对仅在 CLS 模型上继续训练，记为 M6。实验结果如表 4 所示。

表 4 自训练阶段的消融实验结果 %

方法	R-1	R-2	R-L
Ours-2	42.15	23.81	38.24
M4	28.96	10.63	23.88
M5	37.19	17.15	31.95
M6	38.26	19.93	34.49

分析实验结果可以得出以下结论。

1）M5 较 M4 在 ROUGE-2 上提升了 6.52%，说明使用 MT 模型生成伪标签较使用 CLS 模型的效果更好。这是由于在同等数据量下，MT 模型较 CLS 模型更易训练，生成的文本质量更高。

2）M6 较 M5 在 ROUGE-2 上提升了 2.78%，说明使用联合训练同样有利于提升 MT 模型的性能，从而获得更优的伪标签。

3）Ours-2 较 M6 在 ROUGE-2 上提升了 3.88%，说明生成伪标签后继续使用联合训练来同时训练 MT 模型和 CLS 模型比训练单个 CLS 模型的效果更好，即对齐平行语料对在自训练阶段依然能促进模型间的信息共享，从而同时提升两个模型的性能。

3.6 单语摘要数据影响

在前面的实验中，额外的单语摘要数据采用了原始跨语言摘要数据中除去有标签三元组之外的全部数据，其数量相对较大。然而，在实际应用中，单语摘要数据可能没有如此多的数量。因此，本小节对额外单语数据的样本数量进行调整，分别选取三元组样本数量，即 Zh2EnSum 中等场景 25 000 的 1 倍、3 倍、5 倍和 10 倍的样本量进行自训练，以探讨少量单语摘要样本对自训练效果的影响。此外，从与 Zh2EnSum 单语摘要数据不同源的中文单语摘要数据集 RASG[29] 中选取等量样本作为对照，以研究来自不同分布的单语数据对自训练的影响。实验结果如图 5 所示。

从图 5 可以观察到，所有 ROUGE 分数随着额外单语摘要数据的增多而逐渐提升。其中，使用来自同一数据集的单语摘要数据获得的性能提升幅度较高，而不同数据集的单语摘要数据虽然提升幅度较低，但依然呈现随数据量增加而上升的趋势。因此，在跨语言摘要数据有限的情况下，通过增加额外单语摘要数据可以提升模型性能，即使是不同来源的数据也能带来一定的提升。这些结果表明，自

训练方法结合额外单语摘要数据在提升模型性能方面具有显著有效性。

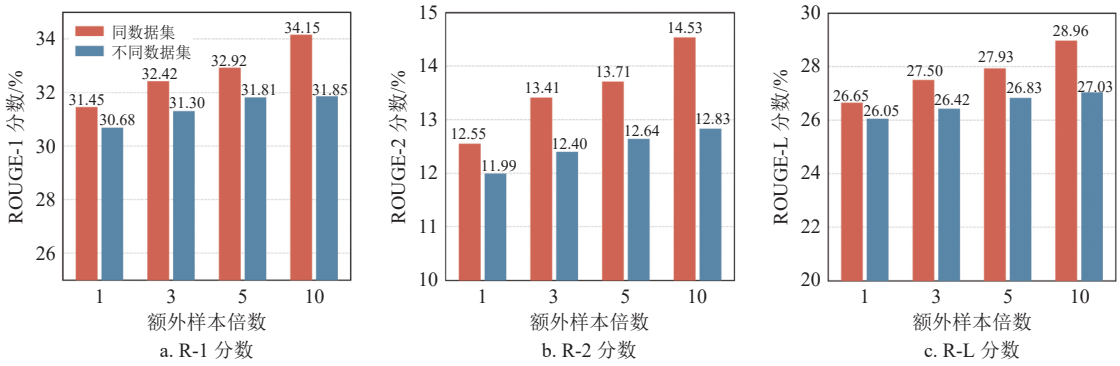

图 5　使用不同额外单语摘要数据的实验结果

4　结束语

本文提出了一种针对低资源场景的跨语言摘要方法，通过联合训练和自训练技术，有效融合了跨语言摘要和单语摘要数据，显著提升了低资源环境下跨语言摘要模型的性能。该方法基于双模型协同的联合训练，将翻译任务和跨语言摘要任务并行训练，通过对齐输出特征和概率，增强了语义共享。此外，结合自训练方法，利用基础模型生成伪标签数据，缓解跨语言摘要数据稀缺的问题。实验结果显示，该方法在多个低资源场景设置下均取得了显著的性能提升，并通过消融实验验证了算法中关键模块的有效性。此外，本文也存在一些局限性：双模型的设计对计算资源的需求较高；在自训练阶段，生成额外样本的伪标签时耗费时间较长，增加了训练成本；目前的实验主要针对特定语言对，其他更复杂的语言场景未能充分验证。因此，未来的研究将继续探索更加简单高效的方法，并在更广泛的语言场景下进行验证，以进一步提升跨语言摘要模型在低资源场景下的性能和适用性。

参 考 文 献

[1] WANG J, MENG F, ZHENG D, et al. A survey on cross-lingual summarization[J]. Transactions of the Association for Computational Linguistics, 2022, 10: 1304-1323.

[2] BAI Y, HUANG H, FAN K, et al. Unifying cross-lingual summarization and machine translation with compression rate[C]//Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2022: 1087-1097.

[3] ZHU J, ZHOU Y, ZHANG J, et al. Attend, translate and summarize: An efficient method for neural cross-lingual summarization[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Virtual: ACL, 2020: 1309-1321.

[4] LIANG Y, MENG F, ZHOU C, et al. A variational hierarchical model for neural cross-lingual summarization [C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Ireland: ACL, 2022: 2088-2099.

[5] BAI Y, GAO Y, HUANG H Y. Cross-lingual abstractive summarization with limited parallel resources[C]// Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Virtual: ACL, 2021: 6910-6924.

[6] NGUYEN T T, LUU A T. Improving neural cross-lingual abstractive summarization via employing optimal transport distance for knowledge distillation[C]//Proceedings of the 36th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI, 2022: 11103-11111.

[7] ZHANG K, ZHANG Y, YU Z, et al. A two-stage fine-tuning method for low-resource cross-lingual summarization [J]. Mathematical Biosciences and Engineering, 2024, 21(1): 1125-1143.

[8] YANG X, YUN J, ZHENG B, et al. Oversea cross-lingual summarization service in multilanguage pre-trained model through knowledge distillation[J]. Electronics, 2023, 12(24): 5001.

[9] LUO F, WANG W, LIU J, et al. VECO: Variable and flexible cross-lingual pre-training for language understanding and generation[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Virtual: ACL, 2021: 3980-3994.

[10] 头旦才让, 仁青东主, 尼玛扎西, 等. 基于改进字节对编码的汉藏机器翻译研究[J]. 电子科技大学学报, 2021, 50(2): 249-255.
THUPTEN Tsering, RINCHEN Dhondub, NYIMA Tashi, et al. Research on Chinese-Tibetan machine translation model based on improved byte pair encoding[J]. Journal of University of Electronic Science and Technology of China, 2021, 50(2): 249-255.

[11] 朱颀林, 王羽, 徐建. 基于异构图和关键词的抽取式文本

摘要模型[J]. 电子科技大学学报, 2024, 53(2): 259-270.
ZHU Q L, WANG Y, XU J. Extractive document summarization model based on heterogeneous graph and keywords[J]. Journal of University of Electronic Science and Technology of China, 2024, 53(2): 259-270.

[12] 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 (TALIP), 2003, 2(3): 245-269.

[13] ORĂSAN C, CHIOREAN O A. Evaluation of a cross-lingual romanian-english multi-document summariser[C]// Proceedings of the 6th International Conference on Language Resources and Evaluation ({LREC}'08). Morocco: ELRA, 2008: 2114-2119.

[14] ZHU J, WANG Q, WANG Y, et al. NCLS: Neural cross-lingual summarization[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: ACL, 2019: 3054-3064.

[15] JIANG S, TU D, CHEN X, et al. ClueGraphSum: Let key clues guide the cross-lingual abstractive summarization [EB/OL]. (2022-03-09). https://arxiv.org/abs/2203.02797.

[16] 何志磊, 高盛祥, 朱恩昌, 等. 基于强化语言关联的中缅跨语言摘要研究[EB/OL].[2024-07-10]. https://doi.org/10.19678/j.issn.1000-3428.0069057.
HE Z L, GAO S X, ZHU E C, et al. Research on cross-language summarization in Chinese-Burmese-Vietnamese based on enhanced linguistic relationships[EB/OL].[2024-07-10].https://doi.org/10.19678/j.issn.1000-3428.0069057.

[17] 冯雄波, 黄于欣, 赖华, 等. 基于多策略强化学习的低资源跨语言摘要方法研究[J]. 计算机工程, 2024, 50(2): 68-77.
FENG X B, HUANG Y X, LAI H, et al. Research on low-resource cross-lingual summarization method based on multi-strategy reinforcement learning[J]. Computer Engineering, 2024, 50(2): 68-77.

[18] LEE D H. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks [C]//Workshop on Challenges in Representation Learning, ICML. Atlanta: ICML, 2013: 896-902.

[19] 任俊飞, 朱桐, 陈文亮. 基于部分标注的自训练多标签文本分类框架[J]. 清华大学学报 (自然科学版), 2024, 64 (4): 679-687.
REN J F, ZHU T, CHEN W L. Self-training with partial labeling for multi-label text classification[J]. Journal of Tsinghua University (Science and Technology), 2024, 64 (4): 679-687.

[20] HE J, GU J, SHEN J, et al. Revisiting self-training for neural sequence generation[C]//International Conference on Learning Representations. Vienna Austria: ICLR, 2020: 1-15.

[21] 周裕林, 陈艳平, 黄瑞章, 等. 结合预训练和自训练的法律信息抽取增强式方法[J]. 燕山大学学报, 2023, 47(3): 255-261.
ZHOU Y L, CHEN Y P, HUANG R Z, et al. An enhanced method of legal information extraction based on pre-training and self-training[J]. Journal of Yanshan University, 2023, 47(3): 255-261.

[22] 张笑燕, 逄磊, 杜晓峰, 等. 基于单语优先级采样自训练神经机器翻译的研究[J]. 通信学报, 2024, 45(4): 65-72.
ZHANG X Y, PANG L, DU X F, et al. Research on self-training neural machine translation based on monolingual priority sampling[J]. Journal on Communications, 2024, 45(4): 65-72.

[23] LIU Y, GU J, GOYAL N, et al. Multilingual denoising pre-training for neural machine translation[J]. Transactions of the Association for Computational Linguistics, 2020, 8: 726-742.

[24] VASWANI A, SHAZEER N, PARMER N, et al. Attention is all you need[C]//Proceedings of the 31st Conference on Neural Information Processing Systems. Long Beach, CA: NIPS, 2017: 1-9.

[25] LEWIS M, LIU Y, GOYAL N, et al. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Virtual: ACL, 2020: 7871-7880.

[26] LIN C Y. ROUGE: A package for automatic evaluation of summaries[C]//Proceedings of Workshop on Text Summarization of ACL. Spain: ACL, 2004: 74-81.

[27] LOSHCHILOV I, HUTTER F. Decoupled weight decay regularization[C]//International Conference on Learning Representations. New Orleans: ICLR, 2019: 1-10.

[28] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: ACL, 2019: 4171-4186.

[29] GAO S, CHEN X Y, LI P J, et al. Abstractive text summarization by incorporating reader comments[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI, 2019: 6399-6406.

编 辑 刘飞阳