Multilingual Generation in Abstractive Summarization: A Comparative Study

Jinpeng Li[1], **Jiaze Chen**[3], **Huadong Chen**[3], **Dongyan Zhao**[1,2]*, **Rui Yan**[4]*
[1] Wangxuan Institute of Computer Technology, Peking University
[2] State Key Laboratory of Media Convergence Production Technology and Systems
[3] Bytedance [4] Gaoling School of Artifical Intelligence, Renmin University of China
lijinpeng@stu.pku.edu.cn, teoyde@gmail.com, chenhuadong.howard@bytedance.com
zhaody@pku.edu.cn, ruiyan@ruc.edu.cn

Abstract

The emergence of pre-trained models marks a significant juncture for the multilingual generation, offering unprecedented capabilities to comprehend and produce text across multiple languages. These models display commendable efficiency in high-resource languages. However, their performance notably falters in low-resource languages due to the extensive linguistic diversity encountered. Moreover, the existing works lack thorough analysis impairs the discovery of effective multilingual strategies, further complicating the advancement of current multilingual generation systems. This paper aims to appraise the efficacy of multilingual generation tasks, with a focus on summarization, through three resource availability scenarios: high-resource, low-resource, and zero-shot. We classify multilingual generation methodologies into three foundational categories based on their underlying modeling principles: Fine-tuning, Parameter-isolation, and Constraint-based approaches. Following this classification, we conduct a comprehensive comparative study of these methodologies across different resource contexts using two datasets that span six languages. This analysis provides insights into the unique advantages and limitations of each method. In addition, we introduce an innovative yet simple automatic metric LANGM designed to mitigate the prevalent problem of spurious correlations associated with language mixing. LANGM accurately measures the degree of code-mixing at the language level. Finally, we highlight several challenges and suggest potential avenues for future inquiry, aiming to spur further advancements within the field of multilingual text generation.

Keywords: Multilingual Generation, Abstractive Summarization, Code-mixing

1. Introduction

Multilingual Generation (MLG) represents a significant research area within the domain of natural language generation, focusing on the automated production of coherent text derived from sources in multiple languages. This field engages with a variety of tasks, such as multilingual machine translation (Sun et al., 2022; Cheng et al., 2023), multilingual summarization (Wang et al., 2021; Scirè et al., 2023), and multilingual machine reading comprehension (Wu et al., 2022; Fan and Gardent, 2020). This paper specifically focuses on multilingual summarization due to its extensive research and practical significance in natural language generation (Hasan et al., 2021; Wang et al., 2021; Ladhak et al., 2020; Scialom et al., 2020). Multilingual summarization models aim to produce summaries in the target language while preserving input information, fluency, and linguistic characteristics of lengthy documents. Thereby presenting additional challenges for current summarization systems.

The imbalance in multilingual data poses significant challenges for multilingual generation. While models such as mBERT (Kenton and Toutanova, 2019) and XLM-R (Conneau et al., 2020) demonstrate remarkable transfer capabilities across languages, their performance diminishes notably when generating content in low-resource languages. The generation process is further influenced by various factors including modeling methods, language families, and the availability of annotated data. These factors contribute to problems such as catastrophic forgetting and spurious correlation. Catastrophic forgetting occurs when a multilingual model, initially trained on data from a target language, experiences a decline in performance upon retraining for a new target language, thereby degrading the performance of the previously learned language model. Spurious correlation is predominantly manifested in the common phenomenon of code-mixing in multilingual generation. Experimental observations indicate that multilingual transfer often results in mixed languages, leading to instances of code-mixing. In zero-shot scenarios, target language words are often underrepresented, and target semantics may remain in different languages. Efforts have been made to mitigate catastrophic forgetting through the development of pre-trained language models (PLMs) (Nguyen and Daumé III, 2019; Cao et al., 2020; Wang et al., 2021). However, the absence of comprehensive studies and standardized benchmarks presents challenges in determin-

* Corresponding authors: Dongyan Zhao and Rui Yan.

LREC-COLING 2024, pages 11827–11837

ing the most effective methods for specific languages. Additionally, there is a notable absence of research focusing on automatic metrics for evaluating code-mixing phenomena, which complicates efforts to quantify and assess such occurrences.

To address the aforementioned problems, this paper aims to conduct a comprehensive comparative analysis of existing works on multilingual summarization across various settings, including high-resource, low-resource, and zero-shot scenarios. Multilingual summarization presents unique challenges due to the inherent complexity of information compression, rendering it an area of significant research importance. We categorize multilingual generation methods into three groups based on their underlying modeling principles: the fine-tuning method, the parameter-isolation method, and the constraint-based method. The fine-tuning method involves refining a pre-trained language model (Liu et al., 2020; Stickland et al., 2021) using downstream task data, thereby leveraging abundant data resources available for each language. The parameter-isolation method entails the utilization of language-specific modules with learnable parameters while maintaining fixed parameters of the pre-trained language model (Bapna and Firat, 2019; Li and Liang, 2021; Lee et al., 2019; Li et al., 2020). Adapters and prefixes facilitate training a small number of parameters to facilitate multilingual content generation. The constraint-based method employs advanced training strategies to update parameters of the pre-trained language model, such as contrastive learning (Wang et al., 2021) and XMAML (Nooralahzadeh et al., 2020). These methods are designed to mitigate the problems of catastrophic forgetting and spurious correlation in multilingual summarization.

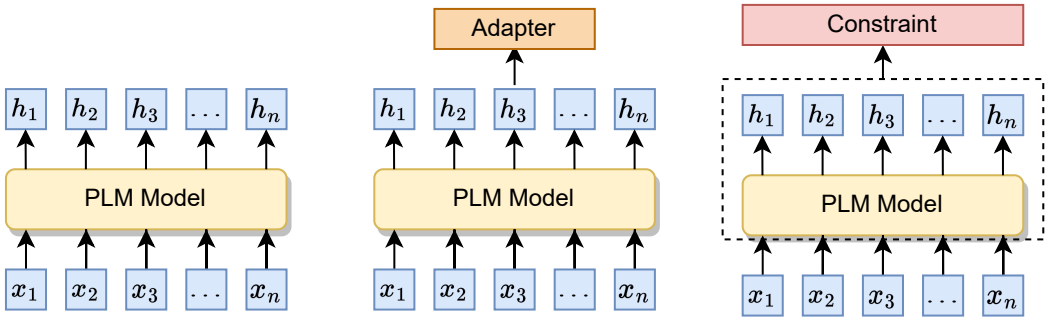
This paper specifically focuses on multilingual summarization across six languages: English (En), German (De), Spanish (Es), French (Fr), Russian (Ru), and Turkish (Tr), utilizing two benchmark datasets, namely WikiLingua and MLSUM. Our objective is to conduct an experimental comparative analysis utilizing varying amounts of data in both high-resource and low-resource scenarios, aiming to aid researchers in selecting the most suitable method and optimization strategy based on specific language contexts. In multilingual generation, the performance of models in the zero-shot setting holds significant importance due to the data imbalance problem. To address this concern, we employ the pre-trained language model mBART (Lewis et al., 2020), fine-tuned with task-specific supervised data, and directly evaluate its performance with low-resource languages in the zero-shot setting. To tackle the common language code-mixing problem in multilingual generation, we propose an automatic metric named **LANGM**. We demonstrate that LANGM exhibits a strong correlation with manual judgments of summary quality. To the best of our knowledge, we are the first to explore automated evaluation metrics for code-mixing, providing a language-level evaluation for multilingual generation, thereby significantly contributing to research in this field. Furthermore, we discuss the challenges and potential research directions in the advancement of multilingual generation to facilitate its progress. In summary, our contributions can be summarized as follows:

- We provide a comprehensive survey of multilingual summarization and categorize existing works based on different modeling principles.
- We introduce the novel and effective automatic metric, LANGM, to address the language code-mixing problem in multilingual generation.
- Experimental comparative analyses with varying amounts of data are designed for three scenarios to inspire future research.

2. Related Work

Multilingual Pre-trained Generation. The impact of pre-trained models on monolingual text has been extensively documented (Kenton and Toutanova, 2019; Liu et al., 2019; Lan et al., 2019), which has led to efforts to extend the success of unsupervised pre-training from English to multiple languages for multilingual comprehension and generation (Conneau and Lample, 2019; Xue et al., 2021; Liu et al., 2020). mBART (Liu et al., 2020) tackles this challenge by denoising full texts in multiple languages and pre-training the entire encoder-decoder model, achieving robust performance in both sentence-level and document-level machine translation. mT5 (Xue et al., 2021) represents a multilingual iteration of T5 specifically tailored for text-to-text tasks. However, multilingual pre-trained models encounter the problem of catastrophic forgetting, a phenomenon absent in monolingual models due to the unified vocabulary. Multilingual models tend to distribute their attention across various languages, resulting in suboptimal performance in downstream tasks. Consequently, effective training strategies become imperative, taking into account the specific language context and the number of languages involved.

Multilingual Summarization. Multilingual summarization aims to generate concise renditions of source documents that encapsulate essential information in multiple languages. It has gained prominence alongside the rapid development of English abstractive summarization techniques (See et al., 2017; Maynez et al., 2020; Zhang et al., 2020; Gehrmann et al., 2018; Chen et al., 2022), owing to its broad applicability. Previous works have explored various approaches in this domain.Nguyen



(a) Fine-tuning method (b) Parameter-isolation method (c) Constraint-based method

Figure 1: The three different methods of multilingual summarization. The fine-tuning aims to update all PLM parameters. The parameter-isolation trains only the parameters of the adapter. The constraint-based uses specific constraint strategies to optimize the model.

and Daumé III (2019) construct a small cross-lingual dataset with English summaries for non-English articles. Cao et al. (2020) utilize a Transformer-based model with six layers of encoder and decoder to integrate auto-encoder training, translation, and summarization. Wang et al. (2021) focus on document-level multilingual summarization with contrastive learning. In addition, the availability of a large number of multilingual datasets has facilitated research in this area. However, the absence of standardized methodologies and benchmarks has hindered equitable comparisons among various approaches. To establish a solid foundation for future research, this paper conducts a comparative analysis based on prevalent model methodologies and datasets.

3. Methods

3.1. Task Formulation

The multilingual summarization model is designed to generate summaries for source documents in multiple languages. Formally, each source document $d^{l_k} = (w_1, w_2, ..., w_m)$ is associated with a standard reference summary $y^{l_k} = (y_1, y_2, ..., y_n)$ in language l_k, where m and n denote the number of words in the original sequences ($m >> n$) and l_k belongs to the set of languages L. The objective of the model is to produce a hypothetical summary $\hat{y}^{l_k}$ in the corresponding language of input. To achieve this, the model is trained on a multilingual dataset using the maximum-likelihood, and the training objective can be defined as follows:

$$\mathcal{L}^{l_k} = -\sum_{t=1}^{n_k} \log P(y_i^{l_k} | d_i^{l_k}), \quad (1)$$

where $d_i^{l_k}$ and $y_i^{l_k}$ represent the i-th sample for language l_k, and n_k is the number of examples in language l_k. This approach enables the multilingual model to generate summaries in the same language as the input document.

3.2. Fine-tuning method

The performance of natural language processing has seen significant enhancements with the advent of pre-trained language models (PLMs) (Kenton and Toutanova, 2019; Radford et al., 2019; Lewis et al., 2020; Liu et al., 2020; Raffel et al., 2020). These models benefit from extensive data and learn powerful language model parameters f_θ. Fine-tuning a PLM f_θ with supervised training dataset for the natural language generation task is the most common training strategy, which has achieved strong performance on many benchmarks(Maurya et al., 2021). However, most fine-tuning strategies rely on supervised data, thereby limiting their effectiveness in low-resource scenarios. Consequently, this strategy is more suitable for scenarios with adequate and balanced data. As shown in Figure 1(a), we utilize a multilingual summarization dataset to fine-tune all parameters of the pre-trained model mBART, which stands as the first sequence-to-sequence denoising auto-encoder pre-trained on large-scale monolingual corpora using BART (Lewis et al., 2020):

$$f_{\hat{\theta}} = \sum_{l_k=1}^{K} f_\theta(d^{l_k}, y^{l_k}). \quad (2)$$

3.3. Parameter-isolation method

The fine-tuning method necessitates balanced representation of each language in the training data. Furthermore, to attain significant performance, it is required to adapt and maintain a separate model for each target language, which can be both costly

and inefficient. To address these challenges, certain methods introduce external parameters to the PLM model, such as Adapter (Bapna and Firat, 2019; Ansell et al., 2021) and Prefix (Li and Liang, 2021), which append specific parameters to the existing pre-trained model for each language. This approach enables the utilization of a small amount of data to train these new parameters while keeping the PLM parameters frozen, thereby effectively enhancing performance in low-resource scenarios. The adapter module typically comprises a single hidden-layer feed-forward network formulation with a nonlinear activation function between the two projection layers. In our experiments, we incorporate the adapter module into the mBART framework to assess its effectiveness across various languages, as shown in Figure 1(b). Formally, the output representation of the i-th layer in the PLM is denoted as $h_i \in \mathbb{R}^b$, then:

$$\text{Adapter}(h_i) = W_{db}(\text{relu}(W_{bd}(LN(h_i)))) + h_i, \quad (3)$$

where LN represents layer-normalization (Ba et al., 2016), and W_{bd} and W_{db} are learnable parameters of the projection layers. Finally, the hidden state is combined with a residual connection (He et al., 2016). In particular, each language corresponds to a distinct set of adapter parameters in this method, while sharing the same PLM parameters.

3.4. Constraint-based method

In addition to fine-tuning and parameter-efficient methods, there exist more intricate training strategies to optimize the model parameters based on the pre-trained model (Chang et al., 2022), as shown in Figure 1 (c). Nooralahzadeh et al. utilize meta-learning based on PLMs (X-MAML) to effectively utilize training data from an auxiliary language for zero-shot and few-shot cross-lingual transfer across a total of 15 languages. Wang et al. employ a contrastive learning strategy to train a multilingual summarization system (CALMS), comprising two training objectives: contrastive sentence ranking and sentence-aligned substitution. These objectives aim to share the ability to extract salient information and align sentence-level representations across languages. Overall, these methods concentrate on diverse constraint strategies for a PLM without introducing additional parameter overhead, while achieving competitive performance. This method can yield more effective constraints based on the data distribution, enabling the model to optimize parameters in the target direction and explore new optimization spaces.

4. Experiment

4.1. Datasets

To evaluate the performance of various multilingual summarization models, we utilize two datasets: WikiLingua (Ladhak et al., 2020) and MLSUM (Scialom et al., 2020). WikiLingua is a large-scale multilingual summarization dataset that consists of article-summary pairs in 18 languages, extracted from WikiHow [1]. It includes a substantial number of English articles along with aligned articles in 17 other languages, enabling the evaluation of multilingual and cross-language summarization tasks. MLSUM is a comprehensive multilingual summarization dataset comprising 1.57 million article-summary pairs in six languages, collected from online newspapers. In order to ensure fairness and comparability in our experiments, we select six languages common to both datasets. Table 1 provides a comprehensive overview of the data statistics used in our study, revealing that Turkish is a low-resource language in WikiLingua dataset.

4.2. Baselines

In our comparative analysis, we consider three prominent methodologies within the field of multilingual summarization: fine-tuning using mBART, parameter-isolation method via Adapter, and constraint-based modeling through CALMS. Specifically, we examine monolingual mBART$_{mon}$, multilingual mBART$_{mul}$, and various mBART variants. For Adapter, we explore its embedding in different network layers to elucidate its significance in multilingual generation. CALMS utilizes mBART as an initialization and incorporates a contrastive learning approach, employing contrastive sentence ranking and sentence-aligned substitution to enhance information extraction proficiency and align sentence-level representations across different languages.

4.3. Evaluation

For automatic evaluation, we employ the widely used summarization metric **ROUGE** (Lin, 2004) to assess the performance of summarization models in different scenarios. ROUGE is based on the n-grams, and is computed by the pyrouge package[2].

Furthermore, as discussed in the preceding section, code-mixing plays a crucial role in multilingual generation models. However, existing automatic evaluation metrics fail to effectively measure the performance of each model in this context. To address this gap, we propose a novel automatic

[1] https://www.wikihow.com
[2] https://pypi.org/poject/pyrouge

Language	Set	English	Spanish	French	German	Russian	Turkish
WikiLingua	Train	131,457	103,215	53,692	48,375	42,928	2,503
	Valid	5,000	5,000	5,000	5,000	5,000	1,000
	Test	5,000	5,000	5,000	5,000	5,000	1,000
MLSUM	Train	287,227	266,367	392,902	220,887	25,556	249,277
	Valid	13,368	10,358	16,059	11,394	750	11,565
	Test	11,490	13,920	15,828	10,701	757	12,775

Table 1: Data statistics of WikiLingua and MLSUM. For the fairness of experiment, we chose the same six languages for the two datasets.

Methods	WikiLingua						MLSUM					
	En	De	Es	Fr	Ru	Tr	En	De	Es	Fr	Ru	Tr
mBART$_{mon}$	**41.78**	**31.92**	**39.15**	**37.62**	**18.89**	26.93	**41.27**	**43.39**	**25.35**	**23.95**	**12.75**	36.28
mBART$_{mul}$	37.96	27.66	29.39	33.55	16.89	**27.92**	40.64	31.58	22.13	23.50	5.14	35.20
Adapter	24.08	18.36	24.79	22.33	8.62	19.53	35.69	28.53	21.63	22.68	13.58	33.77
CALMS	38.27	27.91	29.71	32.97	17.32	26.90	41.64	31.97	22.32	24.70	6.25	**36.30**

Table 2: The ROUGE-1 scores of different methods in high-resource scenarios, except for Turkish in the WikiLingua dataset. mBART$_{mon}$ is to train a monolingual model for each language. The others are multilingual summarization models.

metric termed **LANGM** to evaluate the degree of code-mixing in multilingual generation. Specifically, LANGM leverages the language detection capability of *langid*[3], a standalone language identification tool (Lui and Baldwin, 2012). While *langid* can classify the language of a given text, it does not provide a measure of the degree of code-mixing within the text. To overcome this limitation, we employ a sliding window approach based on n-grams to capture instances of code-mixing. By subjecting the n-gram sliding window to language recognition, we can precisely identify and evaluate code-mixing in the generated text. Formally, **LANGM** is calculated as follows:

$$\text{LANGM}_n = \frac{\sum_{\text{gram}_n \in S}^{p-q+1} \text{langid}^l(\text{gram}_n)}{p-q+1}, \quad (4)$$

Here, p represents the length of the input sequence S and q denotes the length of the sliding window. The gram$_n$ refers to the sequence of words in a sliding window of size n. If gram$_n$ belongs to the target language l, langid^l(gram$_n$) is assigned a value of 1 predicted by the *langid*. Otherwise, it is assigned a value of 0. To determine the most appropriate sliding window size n for each language, we assemble a valid set for each language and evaluate performance across window sizes ranging from 1 to 6. Experimental findings reveal that excessively small values of n may lead to ambiguity problems, wherein a single word may belong to multiple languages. Conversely, overly large values of n diminish the granularity and accuracy of evaluation. Hence, we set $n = 5$ for subsequent experiments based on the experimental results of valid set.

[3]https://github.com/saffsd/langid.py

5. Results and Analysis

5.1. High-resource scenarios

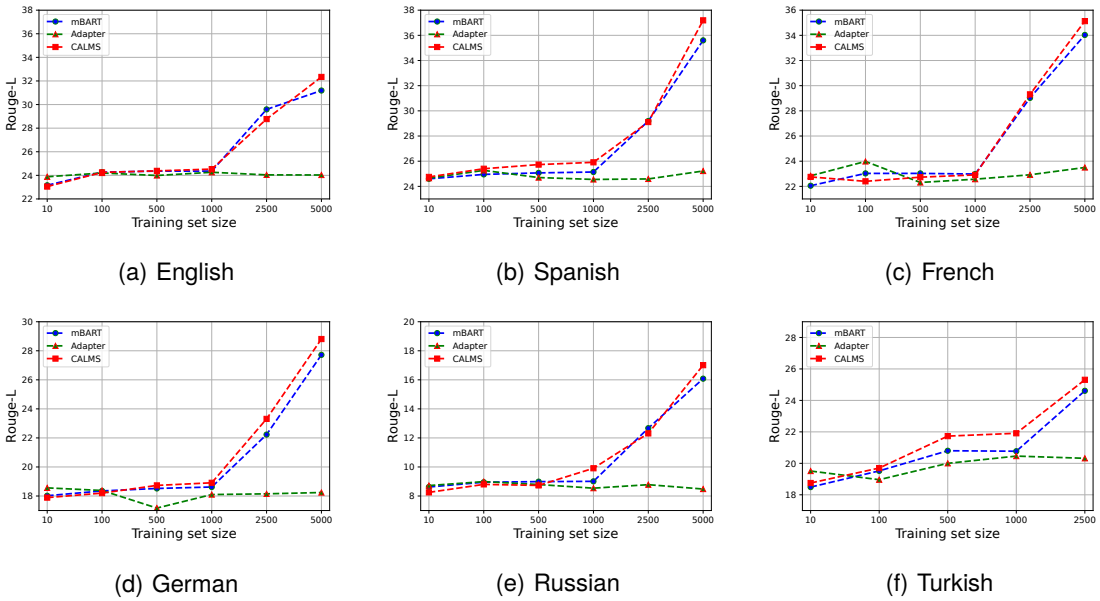
In high-resource scenarios, our investigation focuses on addressing the following inquiries: 1) Does a unified model for all languages outperform individual models for each language? 2) Where does the multilingual summarization model adapter perform better?

Monolingual v.s Multilingual. Table 2 presents our main results across six languages in high-resource scenarios. The Adapter method entails the incorporation of adapter layers into each layer of the encoder. In tasks with ample supervision data, such as English abstractive summarization, English story generation, and English reading comprehension, researchers have developed numerous monolingual and multilingual approaches that have achieved impressive results. However, when it comes to multilingual generation, the highly imbalanced data distribution leads to inconsistent model performance. We observe that monolingual models outperform multilingual models in most cases, except for Turkish in WikiLingua. The scarcity of training data in Turkish allows mBART to leverage information from other languages, thereby enhancing its summary generation capability. This underscores the advantages of exploring multilingual models, enabling the extraction of valuable information from related domains even with fewer parameters and less data.

Adapter Location. Adapters can be inserted as flexible plug-ins at various locations in pre-trained language models. However, the location of the adapter has varying effects on the original pre-trained language model. Quantifying this performance is crucial for guiding future research on mul-

Adapter	WikiLingua						MLSUM					
	En	De	Es	Fr	Ru	Tr	En	De	Es	Fr	Ru	Tr
Encoder-Start	24.05	15.51	23.05	18.75	7.27	16.37	33.54	26.43	21.52	23.90	14.52	31.53
Encoder-End	23.84	18.24	24.47	22.32	8.64	20.51	35.69	28.53	21.63	22.68	13.58	33.77
Decoder-Start	-	-	-	-	-	-	-	-	-	-	-	-
Decoder-End	24.05	17.70	24.89	22.31	8.78	20.03	31.63	25.56	20.10	21.48	13.58	33.96
Encoder-End$_{ALL}$	**24.29**	**18.89**	**24.99**	**22.43**	**9.01**	**20.85**	**35.73**	**29.53**	**21.71**	**23.16**	**14.61**	**34.58**

Table 3: The ROUGE-1 F_1 scores of different location adapter methods in high-resource scenarios (WikiLingua). '-' indicates that the model cannot converge.



(a) English (b) Spanish (c) French

(d) German (e) Russian (f) Turkish

Figure 2: The trend-lines depicting performance with different training set sizes in WikiLingua.

tilingual models. Encoder-End$_{ALL}$ indicates that an adapter layer is added at the end of each layer of the encoder, whereas other methods involve only one adapter layer. Table 3 presents the results of mBART with different adapters at different locations. We observe that placing the adapter at the end of the encoder (Encoder-End and Encoder-End$_{ALL}$) gives the best performance. Conversely, using an adapter at the beginning of the decoder (Decoder-Start) leads to convergence failure in most languages. The autoregressive generation approach leads to a decoder that is sensitive to the representation of the input sequence. Adding an untrained adapter layer to the beginning of the decoder can lead to increased difficulty in convergence of the multilingual model. The embedding vectors after passing through the untrained adapter layer, the implicit state maps to a new hidden space, which leads to difficulties in convergence of the whole decoder. However, adding the adapter at the beginning of the encoder allows for some mapping relationships to be learned after training, but performance is also decreased. This emphasizes the different sensitivities of the encoder and decoder to parameters and the key role of word embeddings in multilingual modeling.

5.2. Low-resource scenarios

In this subsection, we assess the impact of training set size on the performance of multilingual models in low-resource scenarios. We illustrate the performance of three multilingual models with different languages in Figure 2, from which we draw the following noteworthy conclusions:

1) Generally, the adapter method excels when the sample size is below 100. This indicates that employing different adapters for each language effectively enhances performance in resource-constrained situations, particularly with very small datasets. This benefit stems from the fact that distinct adapters model specific language texts, thereby mitigating spurious correlations.

2) As the dataset size increases to approximately 2,500, the fine-tuning pre-trained method demonstrates significant advantages over the other two methods. This highlights the importance of having a sufficient amount of data to effectively transfer knowledge from the pre-trained language model to downstream tasks. When optimizing model pa-

Methods	En	De	Es	Fr	Ru	Tr	AVG
mBART	23.85	18.04	24.58	22.04	8.62	20.46	19.60
mBART$_{En}$	*41.79*	13.23	22.91	22.31	6.89	10.12	19.54
mBART$_{De}$	27.30	*31.94*	22.57	19.74	9.58	17.10	**21.37**
mBART$_{Es}$	26.12	11.30	*39.16*	16.35	7.05	10.09	18.35
mBART$_{Fr}$	22.76	9.08	20.91	*37.61*	8.94	14.34	18.94
mBART$_{Ru}$	3.31	1.12	1.92	0.66	*18.87*	13.33	6.54
mBART$_{Tr}$	14.01	6.79	18.87	13.66	9.44	*30.98*	15.63

Table 4: The ROUGE-1 F_1 scores of different fine-tuning models in zero-shot scenarios. Note that mBART is the pre-trained model without fine-tuning. The mBART$_*$ indicates that mBART is fine-tuned using the $*$ language data, and the italics are supervised results.

Methods	En	De	Es	Fr	Ru	Tr	AVG
mBART$_{Ru}$	3.31	1.12	1.92	0.66	18.87	13.33	6.54
mBART$_{Ru}(10En)$	22.04	4.59	6.30	6.55	19.73	10.69	11.65
mBART$_{Ru}(10De)$	16.61	17.38	9.93	11.05	19.85	10.29	**14.19**
mBART$_{Ru}(10Es)$	4.97	2.39	16.13	1.82	19.48	11.35	9.36
mBART$_{Ru}(10Fr)$	12.60	5.58	10.93	16.36	19.70	11.91	12.84
mBART$_{Ru}(10Tr)$	9.83	4.17	5.16	3.66	19.45	19.04	10.22
mBART$_{Ru}(100En)$	30.40	11.66	14.12	14.47	18.35	13.82	17.14
mBART$_{Ru}(100De)$	24.89	23.86	15.86	18.01	20.84	8.99	18.74
mBART$_{Ru}(100Es)$	27.76	9.96	32.49	23.04	19.42	11.19	**20.64**
mBART$_{Ru}(100Fr)$	21.25	10.81	17.25	30.61	18.04	14.26	18.70
mBART$_{Ru}(100Tr)$	19.94	6.95	11.13	13.21	19.41	26.40	16.17

Table 5: The ROUGE-1 scores for language reproduction in zero-shot scenarios. The superscript indicates the mixed language and number.

rameters, a balance must be struck between the ability to understand multiple languages and the ability to generate summaries. Otherwise, the mixture of multilingual data may cause the pre-trained language model to converge to suboptimal points, resulting in catastrophic forgetting.

3) With continued data augmentation, the knowledge transfer between the pre-trained model and the downstream tasks reaches a bottleneck. The validation set loss plateaus, accompanied by stagnant metric performance. The constraint-based method CALMS begins to exhibit its advantages. The model parameters attain a state of local optimization, and further performance enhancement is restricted by the learning rate. Consequently, it becomes necessary to impose suitable constraints based on the characteristics of the downstream tasks to achieve better performance.

5.3. Zero-shot scenarios

The language family on multilingual summarization. In order to better understand the influence of languages on the summarization task, we select mBART as the baseline and fine-tune it using the summarization data from six languages individually. The results are presented in Table 4. After fine-tuning mBART on language-specific training sets, its ROUGE-1 score exhibits a significant improvement. However, upon direct testing on other languages, the performance decreases in most cases. This phenomenon can be primarily attributed to catastrophic forgetting. Taking English as an example, the original mBART exhibits strong multilingual understanding capabilities. However, during the training process on English data, the model receives no constraints for other languages. As a result, the parameters of mBART are optimized to better understand and generate summaries in English, while the comprehension of other languages is inadvertently forgotten. mBART$_{De}$ achieves the highest average score because it shares the closest language family relationship with other languages. Both English and German belong to the Germanic language family, which contributes to a higher ROUGE-1 score for English (a high-resource language) and boosts the overall average score. On the other hand, Russian exhibits the lowest performance due to its distinct vocabulary compared to the other languages. It should be noted that there is also a significant gap between Turkish tokens and other languages. However, the average score for Turkish is much higher than that of Russian, mainly due to the different sizes of the training data (49,928 for Russian and 2,503 for Turkish). The large amount of Russian data exacerbates the catastrophic forgetting problem in the original mBART, resulting in the lowest overall performance. This situation is commonly observed in low-resource scenarios, necessitating the adoption of specific training strategies to mitigate this problem.

In lifelong learning, researchers often employ reproduction methods to alleviate catastrophic forgetting. Therefore, during the fine-tuning process on German data, we experiment with mixing a small amount of data from other languages. Subsequently, we evaluate the performance as shown in Table 5. Russian is chosen because it has the lowest performance in Table 4, to explore how the overall performance of the model on multilingual summaries can be effectively improved with less cost. The experiments reveal that by incorporating only 10 pieces of English summarization data, the average ROUGE-1 score across the six languages improves by 5.11 points (line 2). This finding holds significant research implications for handling multilingual summarization tasks. Mixing 10 German data samples into Russian yields the most substantial improvement. This result aligns with the observations from Table 4, indicating that German contributes positively to the learning of the other five languages when the data is extremely sparse. Additionally, mBART$_{Ru}(100Es)$ achieves the best performance when mixed with 100 data samples. This can be attributed to Spanish and French belonging to the Romance language family.

The code-mixing on multilingual summarization. To further investigate the specific challenges of multilingual summarization in low-resource sce-

Methods	LANGM(n=5)				Manual			
	En	De	Es	Fr	En	De	Es	Fr
mBART	95.51	95.12	81.22	94.91	97.59	99.63	99.60	99.92
mBART$_{mon}$	90.05	94.93	86.21	94.38	95.92	99.59	99.76	99.81
mBART$_{Ru}$	8.72	4.72	10.47	10.89	73.15	68.20	46.45	65.26
mBART$_{Ru}(10En)$	52.73	28.08	32.08	18.65	83.06	71.41	65.88	74.15
mBART$_{Ru}(100En)$	90.76	81.83	49.20	55.81	96.99	91.39	76.42	80.71

Table 6: The LANGM and manual results (%) of models on WikiLingua with language reproduction.

narios, we introduce the LANGM (from 0 to 1) metric for the first time to evaluate the performance of code-mixing. The Russian and Turkish show a negative correlation so we exclude them. We attribute this discrepancy to the lower accuracy of the *langid* tool in detecting this language family, which also highlights the complexity and diversity of evaluating multilingual generation. The results of the LANGM and manual evaluations across four languages are presented in Table 6. We observe that the consistency between these metrics for four languages (English, Spanish, French, and German) aligns well with the manual judgments of language consistency. In other words, higher LANGM scores and positive manual evaluations indicate better language consistency. Table 6 reveals a counterintuitive result for mBART$_{mon}$. The fine-tuning model on specific language summarization data shows a lower degree of code-mixing compared to the original mBART. Upon investigating the original WikiLingua data, we find that the quality of the training corpus at the language level is poor. For instance, the English training data contains a significant amount of Chinese symbols, resulting in the fine-tuned model easily outputting Chinese tokens when generating English text, bringing about the code-mixing problem. We also observe that by incorporating a small amount of data from other languages, the model can effectively mitigate code-mixing. This demonstrates that reasonable multilingual data reproduction can effectively improve the robustness of the model. As a result, the model develops a better understanding of unseen languages during testing, reducing code-mixing problem.

To assess the correlation between our proposed **LANGM** metric and manual judgments, we randomly select 100 examples generated by the mBART, mBART$_{mon}$, mBART$_{Ru}$, mBART$_{Ru}(10En)$, and mBART$_{Ru}(100En)$ models on the WikiLingua dataset. We invite three well-educated annotators to determine the language of the summaries. Language is considered the primary quality criterion, and the annotators assess the percentage of sentences that belong to the target language. To analyze consistency, we calculate the Pearson's correlation coefficients between the annotators and the **LANGM** scores.

Methods	En	De	Es	Fr
mBART	0.69	0.60	0.58	0.65
mBART$_{mon}$	0.66	0.58	0.57	0.63
mBART$_{Ru}$	0.75	0.63	0.69	0.78
mBART$_{Ru}(10En)$	0.61	0.68	0.68	0.74
mBART$_{Ru}(100En)$	0.71	0.46	0.58	0.53

Table 7: The Pearson's correlation coefficient between LANGM and Manual scores in zero-shot scenarios with language reproduction, and the p-value < 0.01.

Table 7, showcases the results of Pearson's correlation analysis between the automatic metric LANGM and manual evaluations. This table demonstrates the correlation between the proposed metric and the subjective judgments made by human evaluators regarding the code-mixing problem in multilingual summarization.

6. Challenges and Future Directions

Evaluation. Existing automatic evaluation metrics predominantly cater to monolingual tasks, potentially lacking suitability for multilingual generation. Despite some research addressing code-mixing, the field lacks clear automatic metrics for quantifying the problem (Pratapa et al., 2018). Although we introduce the LANGM metric to evaluate language-level consistency, it falls short in effectively monitoring semantic aspects. Evaluating multilingual generation poses significant challenges but holds substantial research importance.
Dataset. Creating high-quality datasets is resource-intensive and time-consuming. Furthermore, certain minority languages face a scarcity of annotators, necessitating the exploration of more efficient algorithms in addition to data annotation. Although numerous multilingual datasets have emerged, they often exhibit imbalanced language distribution and demand more comprehensive and balanced data. Future multilingual research should encompass a broader array of languages and language families. The inclusion of multilingual multimodal data can enhance multilingual generation by providing additional opportunities (Li et al., 2023).
Future Directions. In the future, the development

of multilingual datasets and evaluation methodologies will remain pivotal areas of research. The selection of appropriate models should be guided by data resources and the diversity of language families. While this work offers model recommendations for six languages in varying scenarios, future research should encompass a more extensive set of languages for a comprehensive comparative study. Additionally, we aim to expand the automatic evaluation LANGM metric to encompass the semantics of multiple languages in the future, enhancing its utility for evaluating multilingual generation. Exploring the integration of multimodal and multilingual features also holds great potential. The inclusion of visual content, videos, speech, and other modalities can supply additional contextual information to enhance the fidelity of multilingual generation. These collective efforts will contribute to the broader adoption of multilingual generation techniques within the research community.

7. Conclusion

In this paper, we present a comprehensive study of multilingual generation models in abstractive summarization. We categorize various approaches based on the core modeling principles. Our comparative analysis of these methods using datasets encompassing six languages reveals the following key findings: (1) In high-resource scenarios, monolingual models generally outperform multilingual models, with exceptions in cases of scarce training data, such as Turkish. (2) The placement of adapters within model architecture significantly affects performance, with adapters positioned at the end of the encoder layers showing promising results. (3) Multilingual model performance varies significantly with the size of the training dataset, with fine-tuning exhibiting advantages with larger data volumes. By comparing the strengths and weaknesses of these methods, we address the need for automatic metrics like LANGM to assess code-mixing problem. We also discuss the challenges and potential directions for the development of multilingual generation, including evaluation, dataset, and the integration of multilingual multimodal features. Overall, this study contributes to the advancement of multilingual generation research and offers valuable insights for future developments in the field.

Acknowledgements

We would like to thank the anonymous reviewers for their helpful comments. This work is supported by the National Key Research and Development Program of China (No.2022YFC3301900).

Ethics Statement

This research study on multilingual generation in natural language processing follows ethical considerations to ensure the responsible and ethical use of data, models and evaluation metrics. The research follows ethical guidelines to ensure the protection of individual privacy and confidentiality. The datasets used in this research adhere to appropriate data use rights and licenses. While this ethics statement aims to address the primary ethical considerations of the research, it is important to recognize that ethical challenges in AI research are complex and multifaceted. We remain committed to ongoing discussion, collaboration with the community, and adherence to evolving ethical standards to ensure the responsible development and deployment of multilingual generation models.

8. Bibliographical References

Alan Ansell, Edoardo Maria Ponti, Jonas Pfeiffer, Sebastian Ruder, Goran Glavaš, Ivan Vulić, and Anna Korhonen. 2021. Mad-g: Multilingual adapter generation for efficient cross-lingual transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4762–4781.

Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *ArXiv*, abs/1607.06450.

Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics.

Yue Cao, Xiaojun Wan, Jinge Yao, and Dian Yu. 2020. Multisumm: towards a unified model for multi-lingual abstractive summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11–18.

Ernie Chang, Alex Marin, and Vera Demberg. 2022. Improving zero-shot multilingual text generation via iterative distillation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5876–5881.

Yiran Chen, Zhenqiao Song, Xianze Wu, Danqing Wang, Jingjing Xu, Jiaze Chen, Hao Zhou, and Lei Li. 2022. Mtg: A benchmark suite for multilingual text generation. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2508–2527.

Xuxin Cheng, Qianqian Dong, Fengpeng Yue, Tom Ko, Mingxuan Wang, and Yuexian Zou. 2023. M 3 st: Mix at three levels for speech translation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5. IEEE.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

Angela Fan and Claire Gardent. 2020. Multilingual amr-to-text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*.

Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. 2018. Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Tahmid Hasan, Abhik Bhattacharjee, Md Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M Sohel Rahman, and Rifat Shahriyar. 2021. Xl-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Alina Kramchaninova and Arne Defauw. 2022. Synthetic data generation for multilingual domain-adaptable question answering systems. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 151–160.

Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen Mckeown. 2020. Wikilingua: A new benchmark dataset for cross-lingual abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4034–4048.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Jaejun Lee, Raphael Tang, and Jimmy Lin. 2019. What would elsa do? freezing layers during transformer fine-tuning. *arXiv preprint arXiv:1911.03090*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Xian Li, Changhan Wang, Yun Tang, C. Tran, Yuqing Tang, Juan Miguel Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2020. Multilingual speech translation from efficient finetuning of pretrained models. In *Annual Meeting of the Association for Computational Linguistics*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 4582–4597.

Yaoyiran Li, Ching-Yun Chang, Stephen Rawls, Ivan Vulić, and Anna Korhonen. 2023. Translation-enhanced multilingual text-to-image generation. *arXiv preprint arXiv:2305.19216*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *ACL 2004*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations*, pages 25–30.

Kaushal Kumar Maurya, Maunendra Sankar Desarkar, Yoshinobu Kano, and Kumari Deepshikha. 2021. Zmbart: An unsupervised cross-lingual transfer framework for language generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2804–2818.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919.

Khanh Nguyen and Hal Daumé III. 2019. Global voices: Crossing borders in automatic news summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 90–97.

Farhad Nooralahzadeh, Giannis Bekoulis, Johannes Bjerva, and Isabelle Augenstein. 2020. Zero-shot cross-lingual transfer with meta learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 4547–4562.

Adithya Pratapa, Gayatri Bhat, Monojit Choudhury, Sunayana Sitaram, Sandipan Dandapat, and Kalika Bali. 2018. Language modeling for code-mixing: The role of linguistic theory based synthetic data. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 1543–1553.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. Mlsum: The multilingual summarization corpus. In *2020 Conference on Empirical Methods in Natural Language Processing*, pages 8051–8067. Association for Computational Linguistics.

Alessandro Scirè, Simone Conia, Simone Ciciliano, and Roberto Navigli. 2023. Echoes from alexandria: A large resource for multilingual book summarization. *arXiv preprint arXiv:2306.04334*.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1073–1083.

Asa Cooper Stickland, Xian Li, and Marjan Ghazvininejad. 2021. Recipes for adapting pre-trained monolingual and multilingual models to machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3440–3453.

Simeng Sun, Angela Fan, James Cross, Vishrav Chaudhary, Chau Tran, Philipp Koehn, and Francisco Guzmán. 2022. Alternative input signals ease transfer in multilingual machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 5291–5305.

Danqing Wang, Jiaze Chen, Hao Zhou, Xipeng Qiu, and Lei Li. 2021. Contrastive aligned joint learning for multilingual summarization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2739–2750.

Linjuan Wu, Shaojuan Wu, Xiaowang Zhang, Deyi Xiong, Shizhan Chen, Zhiqiang Zhuang, and Zhiyong Feng. 2022. Learning disentangled semantic representations for zero-shot cross-lingual transfer in multilingual machine reading comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 991–1000.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.