北京大学
PEKING UNIVERSITY

Multilingual Generation in Abstractive Summarization: A Comparative Study

Jinpeng Li[1], Jiaze Chen[3], Huadong Chen[3], Dongyan Zhao[1,2*], Rui Yan[4*]

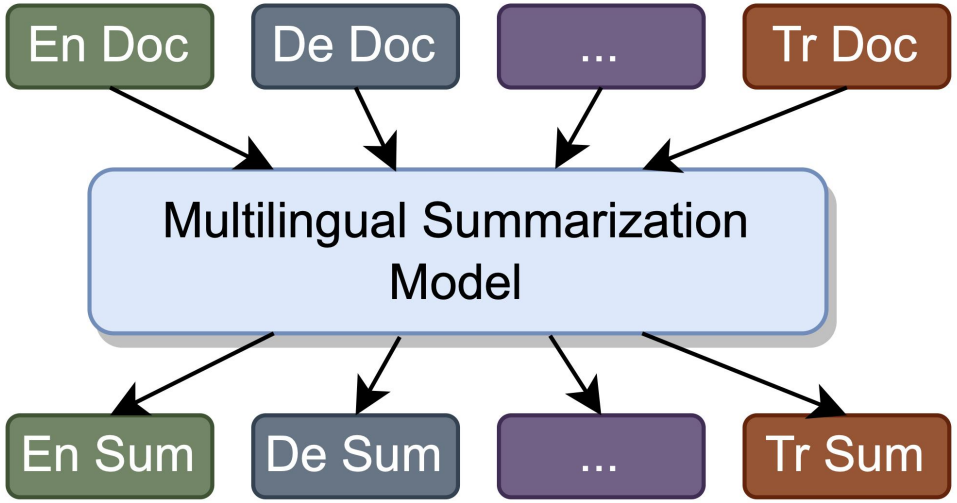
[1]Wangxuan Institute of Computer Technology, Peking University

[2]State Key Laboratory of Media Convergence Production Technology and Systems

[3]Bytedance [4]Gaoling School of Artifical Intelligence, Renmin University of China

Background

- Multilingual Generation 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.
- We specifically focuses on multilingual summarization due to its extensive research and practical significance in natural language generation.



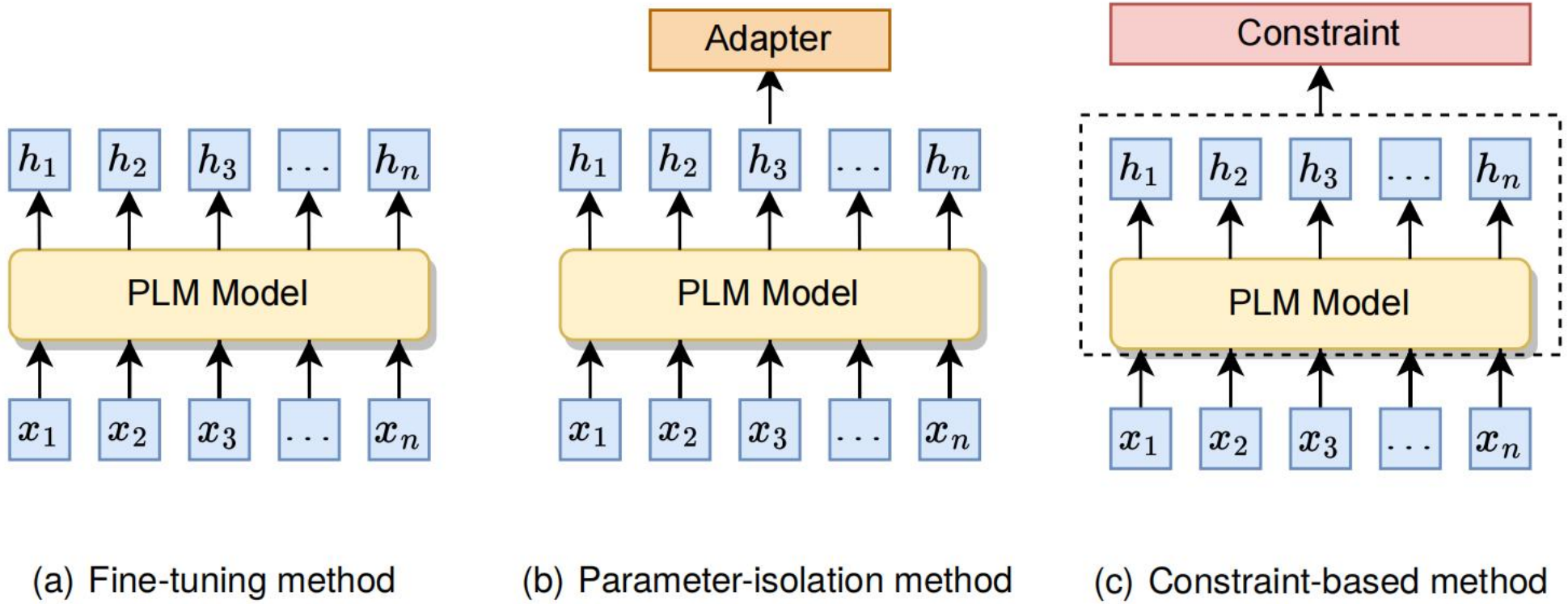
Movitation

Challenge: The imbalance in multilingual data poses significant challenges for multilingual generation.

- The generation process is influenced by various factors. These factors contribute to problems such as catastrophic forgetting and spurious correlation.
  - The modeling methods
  - The language families
  - The annotated data
- 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.
- The absence of comprehensive studies and standardized benchmarks presents challenges in determining the most effective methods for specific languages.

Methods

Mulilingual Generation

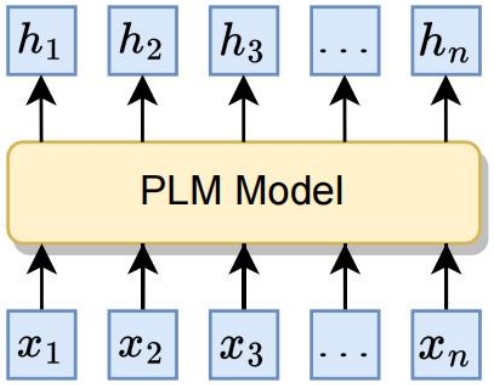


(a) Fine-tuning method (b) Parameter-isolation method (c) Constraint-based method

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.

Methods

Mulilingual Generation



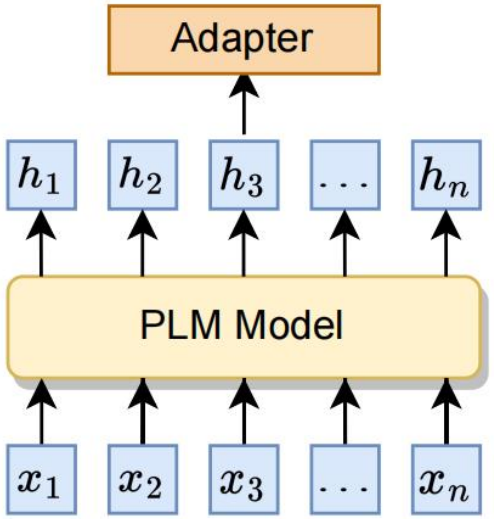
(a) Fine-tuning method

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

- Fine-tuning a PLM with supervised training dataset has achieved strong performance on many benchmarks.
- Mmost fine-tuning strategies rely on supervised data, thereby limiting their effectiveness in low-resource scenarios.
- This strategy is more suitable for scenarios with adequate and balanced data.

Methods

Mulilingual Generation



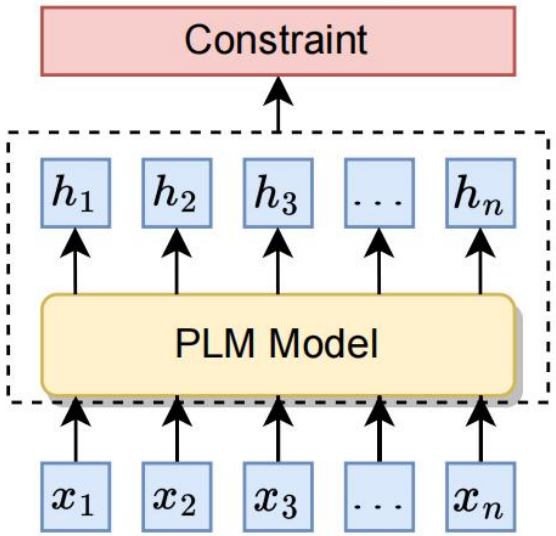
(b) Parameter-isolation method

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

- The fine-tuning method necessitates balanced representation of each language in the training data.
  - Costly and Inefficient
- Introduce external parameters to the PLM model, such as Adapter and Prefix, 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.

Methods

Mulilingual Generation



(c) Constraint-based method

- More intricate training strategies to optimize the model parameters based on the pre-trained model.
- 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.

Experiments

Dataset

- WikiLingua and MLSUM

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.

Experiments

Experimental Results --- High-resource scenarios

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.

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.

Experiments

Experimental Results --- Low-resource scenarios

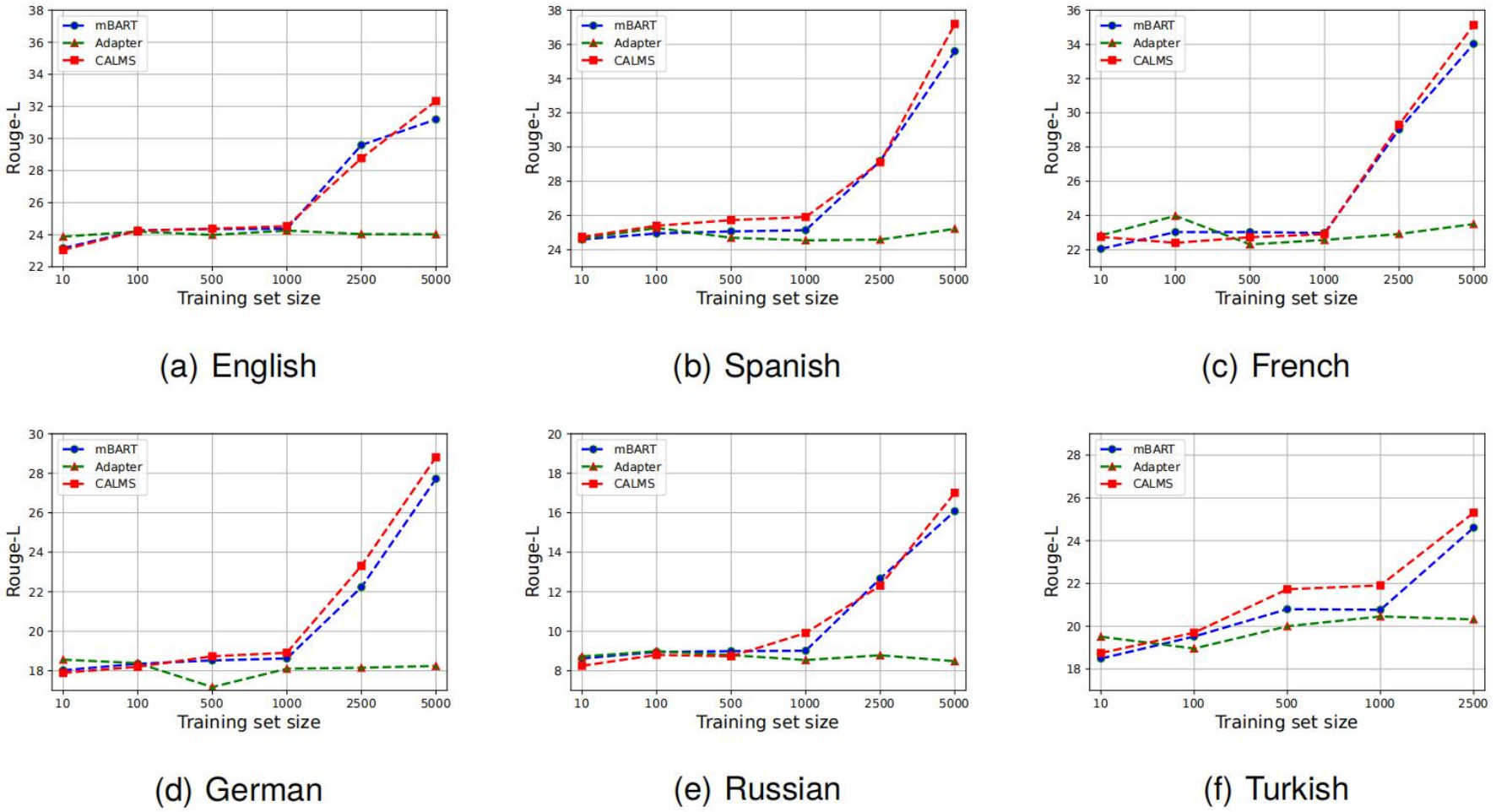


Figure 2: The trend-lines depicting performance with different training set sizes in WikiLingua.

Experiments

Experimental Results --- Zero-shot scenarios

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.

Experiments

LANGM

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

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.

Challenges and Future Directions

Evaluation.

- Existing automatic evaluation metrics predominantly cater to monolingual tasks, potentially lacking suitability for multilingual generation.

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.

Challenges and Future Directions

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.
- 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.

Thanks