Towards Unifying Multi-Lingual and Cross-Lingual Summarization

Jiaan Wang[1]*, **Fandong Meng**[2], **Duo Zheng**[3], **Yunlong Liang**[2]
Zhixu Li[4]†, **Jianfeng Qu**[1]† and **Jie Zhou**[2]
[1]School of Computer Science and Technology, Soochow University, Suzhou, China
[2]Pattern Recognition Center, WeChat AI, Tencent Inc, China [3]Beijing University of Posts and Telecommunications
[4]Shanghai Key Laboratory of Data Science, School of Computer Science, Fudan University, Shanghai, China
jawang.nlp@gmail.com, fandongmeng@tencent.com
zhixuli@fudan.edu.cn, jfqu@suda.edu.cn

Abstract

To adapt text summarization to the multilingual world, previous work proposes multi-lingual summarization (MLS) and cross-lingual summarization (CLS). However, these two tasks have been studied separately due to the different definitions, which limits the compatible and systematic research on both of them. In this paper, we aim to unify MLS and CLS into a more general setting, *i.e.*, many-to-many summarization (M2MS), where a single model could process documents in any language and generate their summaries also in any language. As the first step towards M2MS, we conduct preliminary studies to show that M2MS can better transfer task knowledge across different languages than MLS and CLS. Furthermore, we propose PISCES, a pre-trained M2MS model that learns language modeling, cross-lingual ability and summarization ability via three-stage pre-training. Experimental results indicate that our PISCES significantly outperforms the state-of-the-art baselines, especially in the zero-shot directions, where there is no training data from the source-language documents to the target-language summaries.[1]

1 Introduction

The world we live in is multi-lingual. With globalization, text resources in various languages flood the Internet, where global users can easily access their desired information. Under this background, the text summarization community presents multi-lingual summarization (MLS) and cross-lingual summarization (CLS), respectively. As shown in Figure 1, MLS aims at building a unified model to process documents in multiple languages and generate summaries in the corresponding language (Giannakopoulos et al., 2015; Cao et al., 2020b; Hasan et al., 2021b; Wang et al., 2021; Varab and Schluter, 2021), while CLS generates a summary in the target language from the given document in a different source language (Leuski et al., 2003a; Wan et al., 2010; Wan, 2011; Yao et al., 2015; Zhu et al., 2019; Ladhak et al., 2020; Perez-Beltrachini and Lapata, 2021; Wang et al., 2022b,d,c, 2023). Despite the close relationship between MLS and CLS (*e.g.*, both tasks involve more than one language and require models to distill the key information from documents), previous work studies each task separately, hindering the systematic exploration for both of them.

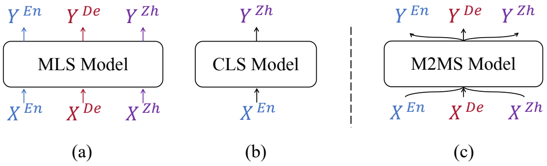


Figure 1: Illustration of (a) multi-lingual summarization, (b) cross-lingual summarization and (c) many-to-many summarization. X^i and Y^i denote the input document and output summary in language i, respectively. En: English; De: German; Zh: Chinese.

In this paper, we aim to unify MLS and CLS into a more general setting named *many-to-many summarization* (M2MS). As its name implies, the goal of M2MS is to build a single summarization model to process a document in any source language and generate the corresponding summary in any given target language. In this manner, one M2MS model could perform more directions than MLS and CLS[2], thus reducing the used parameters. For example, one M2MS model involving n languages could replace one MLS model and $n \times (n-1)$ CLS models. To provide a deeper understanding of M2MS, we also conduct preliminary studies to systematically compare M2MS with MLS and CLS, respectively. In detail, following recent CLS work (Ladhak et al., 2020; Perez-Beltrachini and Lapata, 2021), we use

*Work was done when Jiaan Wang was interning at Pattern Recognition Center, WeChat AI, Tencent Inc, China.

†Corresponding authors.

[1] https://hf.co/Krystalan/PISCES

[2] We use "direction" to denote the summarization direction from the source to the target languages, e.g., English (documents) ⇒ Chinese (summaries).

*Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers*, pages 15127–15143
July 9-14, 2023 ©2023 Association for Computational Linguistics

mBART-50 (Tang et al., 2021) as the summarization model, and train the model in the settings of MLS, CLS and M2MS, respectively. After comparing the model performances, we find that the model trained in M2MS setting can better transfer task knowledge across different languages and combine the advantages of those trained in MLS and CLS settings. Therefore, we argue that it is promising to unify MLS and CLS into M2MS.

Furthermore, we propose PISCES[3], a pre-trained M2MS model that learns language modeling, cross-lingual ability and summarization ability via three pre-training stages: (1) *meta pre-training* learns the general language modeling knowledge from multi-lingual unlabeled corpora; (2) *cross-lingual pre-training* makes the model aware of the transformation between different languages based on parallel corpora; (3) *task-specific pre-training* utilizes M2MS objective to simultaneously improve the cross-lingual ability and the summarization abilities of the model. Considering the high-quality M2MS samples are non-trivial to collect, we leverage a simple strategy to construct pseudo M2MS samples from multi-lingual unlabeled corpora. During the three-stage pre-training, PISCES gradually shifts from learning language modeling to the abilities required by M2MS. Among them, the learned cross-lingual ability plays a key role in enhancing the knowledge transferability of the downstream task (*i.e.*, summarization) from high-resource languages to low/zero-resource languages. Lastly, the pre-trained PISCES could be simply fine-tuned on M2MS with input source-language documents and output target-language summaries.

We evaluate PISCES on the WikiLingua (Ladhak et al., 2020) and CrossSum (Hasan et al., 2021a) datasets. Experimental results show that PISCES achieves promising results compared with the state-of-the-art baselines (*i.e.*, mBART-50 and mT5), especially in the zero-shot directions. Moreover, we find that PISCES is even able to generate summaries for documents whose language never occurs in the fine-tuning stage.

Our contributions are concluded as follows:

- To our knowledge, we are the first to unify MLS and CLS into a more general setting (M2MS). We also conduct preliminary studies to provide deeper analyses among MLS, CLS and M2MS.
- We propose PISCES, a pre-trained M2MS model that learns language modeling, cross-lingual ability and summarization ability through a carefully designed three-stage pre-training.
- We conduct extensive experiments and show that our PISCES achieves new state-of-the-art performance on the large-scale benchmark datasets. Besides, the effectiveness of PISCES in low/zero-resource languages is also demonstrated.

2 Related Work

Multi-Lingual Summarization. Multi-lingual summarization (MLS) aims to process documents in multiple languages and generate their summaries in the corresponding language. Giannakopoulos et al. (2015) present MultiLing-2015 dataset. Later, this task receives increasing attention (Vanetik and Litvak, 2015; Litvak et al., 2016). Recently, large-scale MLS datasets (Scialom et al., 2020; Varab and Schluter, 2021; Hasan et al., 2021b; Feng et al., 2022; Liang et al., 2022a) together with sophisticated methods (Cao et al., 2020b; Chi et al., 2020; Wang et al., 2021; Li et al., 2023) are proposed one after another. Considering the close relation between MLS and CLS, Cao et al. (2020b); Feng et al. (2022) also evaluate the MLS models on CLS to show their zero-shot CLS ability.

Cross-Lingual Summarization. Given documents in one language, cross-lingual summarization (CLS) generates summaries in another language. Early work typically focuses on pipeline methods (Leuski et al., 2003b; Orăsan and Chiorean, 2008; Wan et al., 2010; Wan, 2011; Yao et al., 2015), *i.e.*, translation and then summarization or summarization and then translation. Recently, with the availability of large-scale CLS datasets (Zhu et al., 2019; Ladhak et al., 2020; Perez-Beltrachini and Lapata, 2021; Wang et al., 2022b; Chen et al., 2022; Zheng et al., 2023), many researchers shift the research attention to end-to-end CLS models, including multi-task learning (Cao et al., 2020a; Bai et al., 2021; Liang et al., 2022b), knowledge distillation (Nguyen and Tuan, 2022), resource-enhanced (Zhu et al., 2020) and pre-training (Xu et al., 2020; Chi et al., 2021) approaches. Among them, most CLS work separately builds CLS models in each cross-lingual direction except for Hasan et al. (2021a), who jointly train mT5 (Xue et al., 2021) in multiple directions.

Different from previous MLS and CLS, we unify them into a more general setting (M2MS) starting from the training stage. Besides, we are the first to

[3] PISCES: **P**re-tra**I**ning with gap-**S**entences and **C**ross-lingual d**E**noi**S**ing for many-to-many summarization.

Trg Src	Setting	En	Fr	Hi	Zh	Th	Tr
En	ONE	41.2 / 17.5 / 34.6 / 74.2	35.2 / 14.8 / 29.2 / 73.0	28.2 / 8.3 / 22.6 / 67.7	34.9 / 11.8 / 30.4 / 69.8	34.3 / 14.3 / 30.0 / 66.1	NA
	U-CLS	39.7 / 16.0 / 32.7 / 73.6	36.8 / 15.3 / 29.9 / 73.6	31.2 / 9.2 / 23.9 / 69.0	**37.9 / 13.9 / 32.7 / 71.5**	38.9 / 17.9 / 33.4 / 68.9	**3.2 / 0.3 / 3.0** / 48.9
	MLS	41.6 / 17.9 / 34.7 / 74.4	5.3 / 0.8 / 4.8 / 63.8	3.3 / 0.7 / 3.1 / 53.7	14.6 / 0.9 / 14.5 / 60.1	20.8 / 5.7 / 20.0 / 54.1	2.5 / 0.2 / 2.4 / 47.3
	M2MS	**41.9 / 18.2 / 34.9 / 74.6**	**37.2 / 15.8 / 30.3 / 73.9**	**31.7 / 9.6 / 24.5 / 69.3**	**37.9 / 13.9 / 32.7 / 71.5**	**39.5 / 18.5 / 34.0 / 69.1**	**3.2** / 0.2 / **3.0 / 49.0**
Fr	ONE	35.6 / 13.6 / 29.8 / 72.1	37.8 / 17.4 / 31.2 / 73.9	NA	32.6 / 10.0 / 28.4 / 68.6	31.4 / 11.8 / 27.6 / 64.9	NA
	U-CLS	37.5 / 14.4 / 30.7 / 72.9	37.6 / 16.1 / 30.5 / 74.0	28.2 / 7.6 / 22.0 / **68.1**	36.7 / **12.8** / 31.3 / **70.9**	37.3 / 16.2 / 32.1 / 68.1	**3.3 / 0.3 / 3.1 / 49.4**
	MLS	8.8 / 2.2 / 7.6 / 64.3	**39.5 / 18.2 / 32.5 / 74.9**	2.1 / 0.4 / 1.9 / 53.3	13.5 / 1.0 / 13.2 / 57.5	18.5 / 3.3 / 17.9 / 54.5	2.1 / 0.1 / 2.1 / 46.8
	M2MS	**38.2 / 15.0 / 31.7 / 73.4**	39.2 / 17.9 / 32.0 / 74.7	**28.7 / 7.9 / 22.3 / 68.1**	**36.9 / 12.8 / 31.6 / 70.9**	**37.9 / 16.6 / 32.6 / 68.5**	3.1 / 0.2 / 3.0 / 49.2
Hi	ONE	32.2 / 10.9 / 26.1 / 70.2	NA	32.8 / 11.5 / 25.8 / 69.6	NA	NA	NA
	U-CLS	36.8 / 14.0 / 29.8 / 72.2	31.9 / 11.6 / 24.7 / 71.4	32.7 / 10.3 / 25.6 / 70.3	32.6 / 10.2 / 27.3 / 68.6	34.9 / 14.3 / 29.4 / 67.1	3.3 / **0.3 / 3.2 / 50.0**
	MLS	11.1 / 3.3 / 9.3 / 57.7	11.6 / 3.2 / 9.5 / 59.3	**36.0 / 12.7 / 27.8 / 71.3**	14.2 / 2.8 / 12.8 / 57.2	23.1 / 6.0 / 21.3 / 57.9	2.1 / 0.1 / 2.0 / 46.7
	M2MS	**37.9 / 14.6 / 30.8 / 72.8**	**32.8 / 12.2 / 25.9 / 72.1**	35.6 / 12.5 / **27.8** / 71.1	**33.2 / 10.6 / 28.2 / 69.1**	**35.4 / 14.6 / 30.1 / 67.4**	**3.4 / 0.3 / 3.2** / 49.7
Zh	ONE	34.6 / 11.8 / 28.4 / 71.4	31.5 / 11.4 / 25.4 / 71.0	NA	40.8 / 16.9 / 35.4 / 71.9	NA	NA
	U-CLS	37.7 / 14.1 / 30.8 / 72.8	35.4 / 14.1 / 28.4 / 73.0	25.8 / 6.1 / 20.0 / 66.4	39.6 / 15.1 / 34.2 / 72.2	36.6 / **15.3** / 31.0 / 67.3	3.3 / **0.2** / 3.1 / **49.8**
	MLS	10.4 / 3.0 / 8.6 / 61.7	24.9 / 7.3 / 19.7 / 68.0	20.4 / 4.4 / 16.0 / 62.4	**42.8 / 17.9 / 37.0 / 73.1**	30.3 / 9.3 / 26.4 / 63.5	2.8 / **0.2** / 2.6 / 48.4
	M2MS	**39.2 / 15.1 / 32.0 / 73.4**	**36.0 / 14.5 / 29.0 / 73.3**	**27.0 / 6.6 / 20.8 / 66.9**	41.7 / 17.0 / 35.9 / 72.7	**36.8 / 15.3 / 31.4 / 67.6**	**3.4 / 0.2 / 3.2** / 49.6
Th	ONE	32.1 / 11.1 / 26.4 / 70.4	27.9 / 2.7 / 22.7 / 69.4	NA	NA	37.8 / 17.6 / 33.0 / 67.4	NA
	U-CLS	37.2 / 14.4 / 30.7 / 72.6	34.9 / 13.9 / 27.7 / 72.3	27.1 / 6.8 / 20.6 / 66.9	34.1 / 10.9 / 28.3 / 68.9	39.9 / 18.4 / 34.3 / 69.5	**3.4 / 0.3 / 3.2 / 49.4**
	MLS	7.4 / 1.8 / 6.6 / 54.9	10.1 / 2.5 / 8.4 / 58.4	11.8 / 2.1 / 9.6 / 57.6	16.8 / 3.3 / 15.0 / 59.4	**43.3 / 22.3 / 37.1 / 70.3**	2.7 / **0.3** / 2.6 / 47.8
	M2MS	**38.5 / 15.4 / 31.9 / 73.4**	**35.6 / 14.2 / 28.3 / 72.9**	**27.8 / 7.3 / 21.4 / 67.4**	**34.6 / 11.3 / 29.0 / 69.4**	42.2 / 20.8 / 36.2 / 70.1	3.3 / **0.3** / 3.1 / 49.3
Tr	ONE	NA	NA	NA	NA	NA	NA
	U-CLS	**16.9 / 3.3 / 14.4 / 62.9**	**16.7 / 3.3 / 13.5 / 64.6**	**16.2 / 2.6 / 13.7 / 61.0**	**21.7 / 3.8 / 19.1 / 61.2**	**22.8 / 5.7 / 19.9 / 60.4**	**3.4 / 0.3 / 3.3 / 48.8**
	MLS	6.6 / 0.8 / 5.9 / 53.5	9.7 / 1.1 / 8.6 / 58.7	7.8 / 0.7 / 7.0 / 54.1	17.9 / 2.8 / 15.3 / 58.7	17.4 / 2.5 / 16.6 / 54.4	2.3 / 0.1 / 2.2 / 44.7
	M2MS	15.7 / 2.6 / 13.4 / 62.1	16.0 / 3.2 / 13.2 / 64.4	14.9 / 2.3 / 12.6 / 60.1	19.9 / 3.0 / 17.6 / 60.0	21.4 / 4.8 / 19.3 / 59.9	3.1 / 0.2 / 3.0 / 48.4

Table 1: Results on WikiLingua (ROUGE-1 / ROUGE-2 / ROUGE-L / BERTSCORE). Since there is no training data in zero-shot directions, mBART (ONE) cannot be trained and we denote the results as "NA". The **bold** and underline denote the best and the second-best scores, respectively.

systematically investigate the capabilities of models trained with MLS, CLS and M2MS settings.

Pre-Trained Models for Summarization. Pre-trained models have shown their superiority in summarization task, *e.g.*, BART (Lewis et al., 2020) and T5 (Raffel et al., 2020). To enhance the summarization ability during the pre-training stage, PEGASUS (Zhang et al., 2020a) introduces the gap sentence generation (GSG) objective to enable the model to generate key sentences in an article from the remaining ones. Further, PRIMERA (Xiao et al., 2022) extends GSG from single-document to multi-document summarization. In dialogue scenarios, Wang et al. (2022b) present mDIALBART for cross-lingual dialogue summarization.

Among these pre-trained summarization models, PEGASUS and PRIMERA only focus on monolingual summarization. Though mDIALBART aims at CLS, the model is merely built for a single cross-lingual direction (*i.e.*, English ⇒ German/Chinese) and a specific scenario (*i.e.*, dialogue). Our PISCES is the first multi-lingual pre-trained model for general summarization.

3 Does Unifying All Directions in a Single Model Help Each Other?

As discussed previously, M2MS unifies all summarization directions in a single model. Therefore, we wonder *can such a setting help the model better transfer task knowledge across different languages compared with the settings of MLS and CLS?* To answer the question, we conduct preliminary studies to investigate the influence of different settings.

3.1 Setup

Data. The preliminary studies are conducted on WikiLingua (Ladhak et al., 2020), one of the largest CLS datasets. We focus on six languages, *i.e.*, English (En), French (Fr), Hindi (Hi), Chinese (Zh), Thai (Th) and Turkish (Tr). Among them, Tr serves as a zero-resource language, whose documents and summaries only appear in the validation and test sets. More details are given in Section 5.1.

Summarization Model. Following recent CLS literature (Ladhak et al., 2020; Perez-Beltrachini and Lapata, 2021), we use mBART-50 (Tang et al., 2021) as the summarization model, and train the model in the following four settings:

- mBART (ONE): We separately train several models, each of which is built and evaluated in one single direction. When the direction is cross-lingual (or monolingual), the corresonding model is a CLS (or monolingual summarization) model.
- mBART (U-CLS): We train a unified model with all cross-lingual samples, and test the model in all directions.
- mBART (MLS): We train one unified model with monolingual samples in all languages. Then, the trained model is evaluated in all directions.
- mBART (M2MS): It is a new setting introduced by this work, where the model is both trained and evaluated in all directions.

	En⇒Fr	En⇒Hi	En⇒Zh	En⇒Th
mBART (MLS)	5.8	0.2	1.3	1.0
mBART (M2MS)	99.9	99.4	95.4	99.9
	Fr⇒Hi	Fr⇒Zh	Fr⇒Th	Th⇒En
mBART (MLS)	5.3	5.6	9.4	8.2
mBART (M2MS)	99.4	95.8	99.9	99.5

Table 2: Correct language rate (%) of the summaries generated by mBART (MLS) and mBART (M2MS).

3.2 Analytic Results

Table 1 shows the results in terms of ROUGE (Lin, 2004) and BERTSCORE (Zhang et al., 2020b).

mBART (M2MS) **vs. mBART** (CLS)**.** The results in all directions show that mBART (M2MS) outperforms mBART (CLS) in all metrics, illustrating that unifying all directions in a single model could transfer task knowledge across different languages.

mBART (M2MS) **vs. mBART** (MLS)**.** Comparing mBART (M2MS) and mBART (MLS), it is apparent to find that mBART (M2MS) significantly outperforms mBART (MLS) in cross-lingual directions (*e.g.*, 26.9 vs. 11.7 ROUGE-1 in average), while achieving competitive results in monolingual directions (*e.g.*, 33.9 vs. 34.2 ROUGE-1 in average).

To give a deeper understanding of why mBART (MLS) performs poorly in cross-lingual directions, we analyze its generated summaries and find that most of them are not in the language we expected. Table 2 shows the rate of the generated summaries in the correct language.[4] The languages of the generated summaries are detected by *fastlangid*[5]. Compared with mBART (M2MS), mBART (MLS) struggles to generate summaries in the target language. We conjecture this is because that mBART (MLS) is only trained with monolingual data from multiple languages without any cross-lingual signals, resulting in limited cross-lingual ability.

Based on the above analyses, we argue that the summarization signals from cross-lingual directions could help mBART (M2MS) perform CLS and transfer the task knowledge to zero-shot directions, while mBART (MLS) does not own such abilities.

mBART (M2MS) **vs. mBART** (U-CLS)**.** The only difference between mBART (M2MS) and mBART (U-CLS) is that the training data of mBART (M2MS) contains all monolingual samples, while mBART (U-CLS) does not. We find that the performance gap between mBART (M2MS) and mBART (U-CLS) is extremely smaller than that between mBART (M2MS) and mBART (CLS) / mBART (MLS). In detail, mBART (M2MS) outperforms mBART (U-CLS) in most directions when the source and the target languages have been seen during the fine-tuning stage, *i.e.*, the source and the target languages are from {En, Fr, Hi, Zh, Th}. However, when the source or target language is unseen (*i.e.*, Tr), the performance of mBART (M2MS) is slightly worse than mBART (CLS). This is because the monolingual training data used in mBART (M2MS) makes the word embeddings of the unseen language[6] drift away from those of other languages (see details in Appendix A). Additionally, the cross-lingual signal between the unseen language and other languages never occurs in the fine-tuning stage, making it difficult to summarize from or to the unseen language.

3.3 Preliminary Conclusion

The preliminary studies comparing mBART trained in different settings indicate that (1) the multilingual model trained in M2MS setting can better transfer task knowledge across different languages than those trained in the settings of MLS, CLS and unified CLS. (2) Compared with unified CLS, M2MS helps the model achieve better transferability across visible languages, but sacrifices the transferability to unseen languages.

Grounding the above analyses, we argue that it is valuable to unify previous MLS and CLS to M2MS. Meanwhile, *how to improve the transferability to unseen languages* becomes a keypoint in M2MS.

4 PISCES

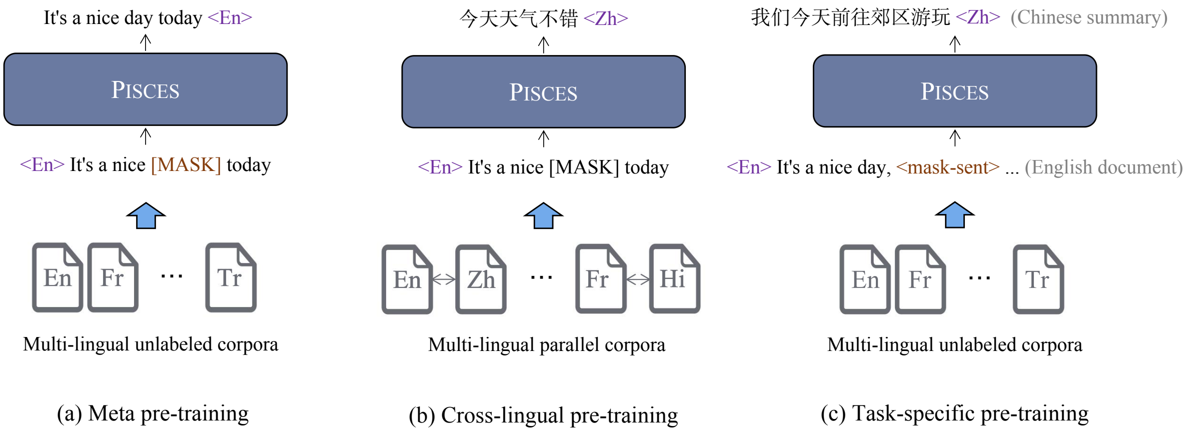
In this section, we propose PISCES, a pre-trained multi-lingual model for M2MS with the backbone of transformer (Vaswani et al., 2017).

Figure 2 shows the overview of PISCES, which contains three pre-training stages. Specifically, the meta pre-training (§ 4.1) lets the pre-trained model learn general language modeling via monolingual denoising objective in multiple languages. Then, to improve the transferability across different languages, the cross-lingual pre-training (§ 4.2) adds noises to the source-language sentences, and encourages the model to translate them into parallel sentences in the target language. Note that the parallel sentences used in this stage might involve the languages which are not seen in downstream tasks,

[4] Other directions also show similar situations.

[5] https://pypi.org/project/fastlangid/

[6] We use "unseen language" to indicate the language does not occur in the *fine-tuning* stage.



It's a nice day today <En>
PISCES
<En> It's a nice [MASK] today
En
Fr
···
Tr
Multi-lingual unlabeled corpora
(a) Meta pre-training
今天天气不错 <Zh>
PISCES
<En> It's a nice [MASK] today
En
Zh
···
Fr
Hi
Multi-lingual parallel corpora
(b) Cross-lingual pre-training
我们今天前往郊区游玩 <Zh> (Chinese summary)
PISCES
<En> It's a nice day, <mask-sent> ... (English document)
En
Fr
···
Tr
Multi-lingual unlabeled corpora
(c) Task-specific pre-training

Figure 2: Overview of the three-stage pre-training in PISCES. Specifically, (a) meta pre-training requires the model to generate original sentences based on the noisy counterparts; (b) cross-lingual pre-training generates the sentences in the target language based on the noisy parallel sentences in the source language; (c) task-specific pre-training utilizes pseudo M2MS samples to pre-train the model.

and it is the key to improving the transferability to these languages. Finally, to narrow the gap between the pre-training and fine-tuning stages, the task-specific pre-training (§ 4.3) trains the model with pseudo M2MS samples, which are constructed from the multi-lingual unlabeled corpora via gap sentences selection and machine translation. During the three-stage pre-training process, the model gradually learns the ability of language modeling, then the cross-lingual ability, and finally the adaptation to the specific task.

4.1 Meta Pre-Training

The goal of meta pre-training is to provide good initialization for the subsequent pre-training stages. Here, we directly utilize mBART-50 (Tang et al., 2021) as the meta pre-trained model.

mBART-50 is a multi-lingual BART (Lewis et al., 2020) with the transformer encoder-decoder architecture. The model is pre-trained on large-scale multi-lingual unlabeled corpora to learn the multi-lingual language modeling. Specifically, following BART, the denoising task is used as the pre-training objective, and there are two types of noise: (1) *text infilling* randomly masks text spans in text sequences, and (2) *sentence permutation* randomly shuffles sentences in documents. The model is required to comprehend the noisy text sequences and recover them. To indicate the input and output languages, the language tags (*e.g.*, <En> and <Zh>) are appended at the inputs of encoder and decoder sides, respectively.

4.2 Cross-Lingual Pre-Training

Despite the effectiveness of mBART-50, the input and output sequences in its pre-training stage are always in the same language, resulting in the under-explored cross-lingual ability. However, such ability is indispensable for M2MS. Therefore, cross-lingual pre-training is designed to improve the cross-lingual transferability.

In detail, we propose a simple yet effective pre-training task, *i.e.*, cross-lingual denoising, which lets the model generate sentences in the target language based on their noisy parallel sentences in a different source language. The noise used in this stage is *text infilling*. In this way, the pre-trained model is required to not only understand the text in the source language but also learn the transformation between different languages.

4.3 Task-Specific Pre-Training

Task-specific pre-training aims to narrow the gap between the pre-training and fune-tuning stages. We directly adopt M2MS as its pre-training task. Grounding the truth that high-quality M2MS samples are difficult to collect, we construct the pseudo samples from multi-lingual unlabeled corpora.

In detail, for a source-language document $D = \{s_i^{src}\}_{i=1}^{|D|}$, where s_i^{src} denotes the i-th sentence in D. Following previous monolingual pre-trained summarization methods (Zhang et al., 2020a; Xiao et al., 2022), we calculate the importance of each sentence as $\mathcal{S}(s_i^{src}) = \text{ROUGE-1}(s_i^{src}, D/s_i^{src})$, where D/s_i^{src} indicates the rest of the document after s_i^{src} is removed. The sentences with high importance are selected as the gap sentences $S_*^{src} =$

Trg / Src		En	Fr	Hi	Zh	Th	Tr
En	# Samples	124589 / 8351 / 8517	53232 / 5161 / 5258	5707 / 1538 / 2672	13462 / 2697 / 2713	9170 / 2883 / 2697	- / 267 / 2730
	# Avg. Tokens	492.8 / 47.3	521.3 / 55.4	500.6 / 71.8	516.8 / 49.4	524.2 / 48.4	458.3 / 54.3
Fr	# Samples	53232 / 5161 / 5258	53232 / 5161 / 5258	- / 1449 / 2337	10628 / 2605 / 2400	7281 / 2750 / 2386	- / 232 / 2391
	# Avg. Tokens	659.4 / 45.3	659.3 / 55.5	617.3 / 73.1	649.0 / 48.5	673.4 / 47.3	589.9 / 54.4
Hi	# Samples	5707 / 1538 / 2672	- / 1449 / 2337	5707 / 1538 / 2672	- / 1134 / 2000	- / 1266 / 2146	- / 180 / 2091
	# Avg. Tokens	682.1 / 46.2	668.3 / 58.2	684.3 / 72.3	637.9 / 50.5	626.1 / 48.7	627.4 / 53.0
Zh	# Samples	13462 / 2697 / 2713	10628 / 2605 / 2400	- / 1134 / 2000	13462 / 2697 / 2713	- / 2392 / 2218	- / 90 / 2147
	# Avg. Tokens	428.4 / 46.4	432.9 / 58.1	388.7 / 73.6	429.1 / 49.2	371.1 / 49.8	373.2 / 55.5
Th	# Samples	9170 / 2883 / 2697	7281 / 2750 / 2386	- / 1266 / 2146	- / 2392 / 2218	9170 / 2883 / 2697	- / 191 / 2172
	# Avg. Tokens	488.6 / 44.5	504.9 / 56.2	424.6 / 71.8	412.1 / 51.0	490.1 / 48.2	404.1 / 54.2
Tr	# Samples	- / 267 / 2730	- / 232 / 2391	- / 180 / 2091	- / 90 / 2147	- / 191 / 2172	- / 267 / 2730
	# Avg. Tokens	465.1 / 47.5	472.4 / 60.0	468.1 / 72.8	456.9 / 52.7	449.1 / 49.8	465.1 / 54.3

Table 3: Statistics of re-splitted WikiLingua. *# Samples* denotes the number of samples in training / validation / test set. *# Avg. Tokens* represents the average tokens in the documents and summaries, respectively. Green , light green and gray indicate the high-resource , low-resource and zero-shot directions, respectively.

$\{s^{src}_{g_i}\}^{|S^{src}_*|}_{i=1}$ ($g_i \in \{1, 2, ..., |D|\}$), which are further translated to a different target language $S^{trg}_* = \{s^{trg}_{g_i}\}^{|S^{trg}_*|}_{i=1}$ via Google Translation[7]. In this manner, the source-language document D paired with source/target-language gap sentences S^{src}_*/S^{trg}_* could constitute a pseudo pre-training sample.

Quality Controlling. Since machine translation results might contain flaws, we further employ *round-trip translation* strategy as suggested by Zhu et al. (2019) and Feng et al. (2022). For each gap sentence $s^{src}_{g_i}$ in D, the translated counterpart $s^{trg}_{g_i}$ is translated back to the source language, which we denote as $s^{src'}_{g_i}$. If the ROUGE-1 score between $s^{src}_{g_i}$ and $s^{src'}_{g_i}$ is less than the pre-defined threshold λ, the corresponding pseudo sample will be discarded.

Input Format. To help the model trade off between (1) generating new sentences instead of translating part of input sentences, and (2) learning the translation pattern[8] (Zhu et al., 2020), half of source-language gap sentences in D are randomly masked with a special token `<mask-sent>`.[9]

5 Experiments

5.1 Benchmark Datasets

In order to evaluate M2MS models, two requirements should be met in datasets, *i.e.*, (1) involving multiple languages and summarization directions, and (2) having abundant samples in each direction. Thus, we choose WikiLingua (Ladhak et al., 2020) and CrossSum (Hasan et al., 2021a).

The original WikiLingua dataset, which involves 18 languages, is designed for CLS task. The 18 languages constitute 306 (18×17) cross-lingual directions, each of which contains about 18k CLS samples in average. For each document, WikiLingua also contains its summary in the original language. Therefore, the dataset could be used to evaluate M2MS models. However, the original splitting is for CLS. Thus, we re-split WikiLingua with the special consideration for M2MS: for each document in the test (or validation) set of one direction, the document and its parallel documents[10] are not allowed to appear in the training and validation (or test) sets of other directions. This rule reduces the likelihood that learning shortcuts. We also intentionally create several zero-shot directions.

We focus on six languages in this work: English (En), Chinese (Zh), French (Fr), Hindi (Hi), Turkish (Tr) and Thai (Th). After re-splitting, the statistics are shown in Table 3. There are **9 high-resource directions** each of which contains more than 10k training samples. The other **8 directions** with less than 10k training samples are considered as **low-resource directions**. The remaining 19 zero-shot directions have no training sample. According to *whether both the source and target languages appear in the whole training set*, we further divide them into **11 non-trivial and 8 conventional zero-shot directions**. Note that Tr never appears in the training set of any direction, thus, in other words, the non-trivial zero-shot directions involve Tr while the conventional counterparts do not. We call Tr an *unseen language*. Though there is no training data in a conventional zero-shot direction, both its source and target languages might

[7] `https://cloud.google.com/translate`

[8] In CLS, Zhu et al. (2019) find some words in summaries are directly translated from the source words.

[9] We also attempt to mask all gap sentences or do not mask any gap sentences, the results underperform that of masking half of the gap sentences.

[10] For each document, WikiLingua usually contains its parallel documents in other languages.

Non-Trivial Zero-Shot Directions									
Direction	Tr⇒Others					Avg.	Any⇒Tr		
	Tr⇒En	Tr⇒Fr	Tr⇒Hi	Tr⇒Zh	Tr⇒Th		En⇒Tr	Fr⇒Tr	Tr⇒Tr
mT5 (580M)	9.5 / 61.6	10.0 / 63.8	7.8 / 59.1	12.6 / 59.6	14.0 / 59.6	10.8 / 60.7	2.2 / 48.9	2.1 / 48.8	2.0 / 48.2
mBART (610M)	10.6 / 62.1	10.8 / 64.4	9.9 / 60.1	13.5 / 60.0	15.2 / 59.9	12.0 / 61.3	2.1 / 49.0	2.1 / 49.2	2.1 / 48.4
PISCES (610M)	20.2 / 68.2	19.6 / 68.9	15.7 / 64.9	21.2 / 66.7	22.9 / 64.9	19.9 / 66.7	3.1 / 53.8	2.8 / 53.4	3.7 / 52.9

Conventional Zero-Shot Directions									
Direction	Fr⇒Hi	Hi⇒Fr	Hi⇒Zh	Zh⇒Hi	Hi⇒Th	Th⇒Hi	Zh⇒Th	Th⇒Zh	Avg.
mT5 (580M)	18.8 / 66.9	23.0 / 71.2	23.1 / 68.3	17.5 / 66.1	25.8 / 65.9	17.8 / 66.4	27.2 / 66.9	24.6 / 69.2	22.2 / 67.6
mBART (610M)	19.6 / 68.1	23.6 / 72.1	24.0 / 69.1	18.1 / 66.9	26.7 / 67.4	18.8 / 67.4	27.8/ 67.6	25.0 / 69.4	23.0 / 68.5
PISCES (610M)	21.4 / 69.1	26.1 / 72.9	26.1 / 70.4	20.3 / 68.5	29.1 / 68.5	21.4 / 69.0	29.9 / 68.9	27.0 / 71.0	25.2 / 69.8

Low-Resource Directions									
Direction	Hi⇒Hi	Th⇒Th	En⇒Hi	Hi⇒En	En⇒Th	Th⇒En	Fr⇒Th	Th⇒Fr	Avg.
mT5 (580M)	24.7 / 70.7	32.5 / 69.6	20.8 / 68.5	27.1 / 72.3	29.9 / 68.3	27.8 / 73.1	28.1 / 67.3	25.3 / 72.2	27.0 / 70.2
mBART (610M)	25.3 / 71.1	33.1 / 70.1	21.9 / 69.3	27.8 / 72.8	30.7 / 69.1	28.6 / 73.4	29.0 / 68.5	26.0 / 72.9	27.8 / 70.9
PISCES (610M)	26.5 / 71.8	34.2 / 70.7	23.7 / 70.3	29.5 / 73.6	31.9 / 70.1	30.0 / 74.0	30.0 / 69.2	27.4 / 73.8	29.2 / 71.7

High-Resource Directions										
Direction	En⇒En	Fr⇒Fr	Zh⇒Zh	En⇒Fr	Fr⇒En	En⇒Zh	Zh⇒En	Fr⇒Zh	Zh⇒Fr	Avg.
mT5 (580M)	30.9 / 74.0	29.8 / 74.3	31.0 / 72.5	26.8 / 73.3	27.5 / 72.7	27.5 / 70.9	28.0 / 73.0	26.4 / 70.3	25.5 / 72.5	28.2 / 72.6
mBART (610M)	31.7 / 74.6	29.7 / 74.7	31.5 / 72.7	27.8 / 73.9	28.3 / 73.4	28.2 / 71.5	28.8 / 73.4	27.1 / 70.9	26.5 / 73.3	28.8 / 73.2
PISCES (610M)	32.4 / 75.0	30.3 / 75.0	32.1 / 73.0	28.5 / 74.3	29.0 / 73.8	28.8 / 71.9	29.7 / 73.9	27.4 / 71.3	27.6 / 73.7	29.5 / 73.5

Table 4: Experimental results on WikiLingua. Avg. indicates the average score for each cluster of directions. PISCES is significantly better than mBART with t-test $p < 0.01$ in all directions.

have training data with a pivot language, making it less challenging than the non-trivial ones. Taking the conventional zero-shot direction Hi⇒Zh as an example, the training data in Hi⇒En and En⇒Zh could bridge the gap between Hi and Zh. For statistics of the CrossSum dataset used in our experiments, please refer to Appendix C.1.

5.2 Experimental Setup

Baselines. We use mBART-50 (Tang et al., 2021) and mT5 (Xue et al., 2021) as baselines, which have achieved state-of-the-art performances on many CLS/MLS datasets (Perez-Beltrachini and Lapata, 2021; Hasan et al., 2021a; Feng et al., 2022).

Metrics. We adopt ROUGE-1/2/L (Lin, 2004) and BERTSCORE (Zhang et al., 2020b) in our experiments. The ROUGE scores measure the lexical overlap between the generated summaries and corresponding references, while the BERTSCORE measures the semantic similarity. These metrics are calculated by *multi-lingual rouge*[11] and *bert-score*[12] toolkits, respectively. The BERTSCORE is based on *bert-base-multilingual-cased* model. The statistical significance test (Koehn, 2004) is also employed for a fair comparison.

Implementation Details. The implementation details of the pre-training objectives, pre-training corpora and fine-tuning hyper-parameters are given in Appendix B.

5.3 Quantitative Results

Table 4 shows the results on WikiLingua in terms of average ROUGE score (RS) and BERTSCORE (BS). Full results on ROUGE-1/2/L are given in Appendix D. The experimental results on CrossSum also verify the superiority of PISCES, which are provided in Appendix C.2.

PISCES vs. Baselines. Our PISCES outperforms mBART-50 and mT5 in all directions, indicating its superiority. Specifically, PISCES achieves an average increase of 7.9 RS and 5.4 BS over mBART-50 in non-trivial zero-shot directions when the target language is not Tr. Compared with mBART-50, the average improvement in conventional zero-shot directions is 2.2 RS / 1.3 BS, while the counterpart in low-resource directions is 1.4 RS / 0.8 BS. As for high-resource directions, PISCES outperforms mBART-50 by 0.7 RS and 0.3 BS in average. It is not difficult to find that the fewer resources in a direction, the greater the improvement brought by our PISCES. This finding also indicates the potentiality of our model when faced with the real-world scenario, since there are thousands of languages in the world and most directions are low-resource or zero-shot. Through the cross-lingual and task-specific pre-training stages, PISCES facilitates the transfer of task knowledge from high-resource directions to the low-resource and zero-shot ones.

Non-Trivial Zero-Shot Direction. As shown in Table 4, we divide the non-trivial zero-shot directions into two categories (*i.e.*, Tr⇒Others and Any⇒Tr) according to whether Tr is the target language. We

[11] https://github.com/csebuetnlp/xl-sum/tree/master/multilingual_rouge_scoring

[12] https://github.com/Tiiiger/bert_score

	Fr⇒Hi	Hi⇒Fr	Hi⇒Zh	Zh⇒Hi
PISCES	**21.4 / 69.1**	**26.1 / 72.9**	**26.1 / 70.4**	**20.3 / 68.5**
w/o TS	20.7 / 68.6	25.2 / 72.8	25.1 / 69.9	19.5 / 67.9
w/o CL	20.6 / 68.8	25.2 / **72.9**	25.3 / 70.0	19.5 / 67.8
w/o TS & CL	19.6 / 68.1	23.6 / 72.1	24.0 / 69.1	18.1 / 66.9

Table 5: Results of ablation studies.

Model	WikiLingua								
	En⇒Zh			Zh⇒En			En⇒En		
	IF	CC	GM	IF	CC	GM	IF	CC	GM
mT5	2.93	3.12	3.06	2.98	3.29	3.03	3.16	3.84	3.52
mBART	3.09	3.38	3.14	3.15	3.53	3.26	3.27	3.96	3.71
PISCES	**3.17**	**3.56**	**3.41**	**3.24**	**3.76**	**3.54**	**3.52**	**4.28**	**4.16**

Table 6: Human evaluation results. "IF", "CC" and "GM" denote informativeness, conciseness and grammaticality, respectively.

discover that the results in Any⇒Tr directions[13] are significantly worse than the Tr⇒Others counterparts. This finding suggests that generating summaries in *unseen languages* is more difficult than understanding documents in *unseen languages*. This is because the encoder could partly understand the *unseen languages* through the shared vocabulary and the similar syntax constituent with other languages. But for the decoder, we only change its language tag to expect it can generate summaries in *unseen languages*. This requires the decoder to *simultaneously* (1) capture the relationships between the unseen language tag and the unseen language tokens and (2) summarize documents. However, the pre-trained model only meets the requirement (1) in the pre-training stage[14], while requirement (2) in the fine-tuning stage, making it hard to simultaneously meet both requirements, and consequently, cannot generate summaries in unseen languages. We reserve this challenge for future work.

Ablations. We conduct ablation studies to investigate the effect of the cross-lingual and task-specific pre-training stages. We run the following ablations:

- **PISCES w/o TS**. To demonstrate the effectiveness of the task-specific pre-training, we also pre-train a variant PISCES model which does not include the task-specific pre-training stage.
- **PISCES w/o CL**. To measure the effectiveness of the cross-lingual pre-training, we remove this stage in the whole pre-training process, resulting in another variant PISCES.
- **PISCES w/o TS & CL** removes both the cross-lingual and task-specific pre-training stages, which is the same as mBART-50.

How to Download Photos from Your iPhone to a Computer	
Turkish Document	iPhone'un şarj kablosunun bir ucunu iPhone'un şarj girişine tak, ardından USB ucunu bilgisayarının USB girişlerinden birine tak. Kilidini açmak için parolanı (veya TouchID'ni ya da FaceID'ni) gir ve iPhone'undaki Home düğmesine bas. Devam etmeden önce, istenirse "Bu bilgisayara güvenilsin mi?" kısmında Güven seçeneğine dokun. Mac'in Dock'unda çok renkli bir çarkıfeleğe benzeyen Fotoğraflar uygulaması simgesine tıkla. Fotoğraflar uygulaması iPhone'unu bağladığında otomatik olarak açılabilir. iPhone'un simgesi, uygulamanın penceresinin sol üst köşesinde görünmelidir. Fotoğrafların alınıp içeri aktarılacağı yer olarak pencerenin sol tarafında iPhone'unun adına tıkla. Bunu penceredeki resimlere tıklayarak yap. Bilgisayarında olmayan tüm fotoğrafları içeri aktarmak istiyorsan bu adımı atla. Bu, pencerenin sağ üst köşesindedir. Seçtiğin fotoğraf sayısı bu butonda görünecektir (örneğin, 5 Seçileni İçeri Aktar). iPhone'undaki Mac bilgisayarında olmayan tüm fotoğrafları aktarmak istiyorsan Tüm Yeni Ögeleri İçeri Aktar seçeneğine tıkla. Bu, pencerenin sol tarafındadır. Az önce aktardığın fotoğraflar bu sayfada listelenir.
mBART	Examine the iphone's keyboard. Click the "screen" button to view the photos. Click the "screen" button to view the list of available photos.
PISCES	**Connect your iphone to** computer. **Unlock your iphone.** Click **the "photos" app**. **Select the photos you** wish **to download.** Click the "choose photos" option. **Select the photos you** wish **to download.** Click the "download" button.
Ground Truth	**Connect your iphone to** your mac. **Unlock your iphone.** Open **the photos app.** Select your iphone. **Select the photos you**'d like **to download.** Click import selected. Click imports.

Table 7: An example of Tr⇒En summarization.

As shown in Table 5, we conduct ablation studies in several conventional zero-shot directions (results in more directions are provided in Appendix E). In each case, the RS and BS are lower than vanilla PISCES. In addition, both PISCES w/o TS and PISCES w/o CL outperform PISCES w/o TS & CL. Thus, the effectiveness of both stages is proved.

5.4 Qualitative Results

Human Evaluation. Following Zhu et al. (2020); Liang et al. (2022b), we conduct the human evaluation on 50 random samples extracted from WikiLingua (En⇒Zh, Zh⇒En and En⇒En, respectively). Three graduate students are invited to assess the generated summaries from three aspects: informativeness (IF), conciseness (CC) and grammaticality (GM). The scoring adopts a 5-point scale from 1 (worst) to 5 (best). Table 6 shows the average results. The IF, CC and GM scores of PISCES are significantly better than those of mT5 or mBART-50, demonstrating the effectiveness of our model.

Case Study. Table 7 shows an example Turkish document, the generated summary and the ground truth summary. Though the summary generated by PISCES contains a repeated sentence, it has good overlaps with the ground truth. But for mBART-50, the generated summary is not relevant to the core idea of the document. This observation indicates that, through the cross-lingual and task-specific pre-training, PISCES could better transfer the task knowledge from high-resource directions to zero-shot ones, and even has the ability to generate summaries for the documents whose language does not occur in the fine-tuning stage.

[13] Results on Hi/Zh/Th⇒Tr are given in Appendix D

[14] Though PISCES has been pre-trained with pseudo M2MS samples, there is still a large gap between the pseudo samples and downstream samples, *e.g.*, text style and domain.

Error Analysis. To further study how future research could advance M2MS, we take a closer look at the generation errors of PISCES and analyze them in Appendix F.

6 Conclusion

In this paper, we unify MLS and CLS to M2MS. Through carefully-designed preliminary studies, we discuss that unifying MLS and CLS to M2MS is valuable. In addition, we propose PISCES, the first pre-trained M2MS model, which contains three pre-training stages to enable the model learn the multilingual language modeling, cross-lingual ability and summarization ability. Extensive experiments show its superiority compared with the state-of-the-art baselines (mBART-50 and mT5). The case study further demonstrates that our model could even generate summaries for the documents whose language does not occur in the fine-tuning stage.

Ethical Considerations

In this section, we consider potential ethical issues of our model. In this paper, we propose PISCES which utilizes mBART-50 (Tang et al., 2021) as the meta pre-trained model and further suffers from the cross-lingual pre-training and task-specific pre-training stages. The pre-training samples are constructed from OPUS (Tiedemann and Thottingal, 2020) and mC4 (Xue et al., 2021) corpora. To construct the pseudo M2MS samples in the task-specific pre-training stage, Google Translation is also adopted to translate gap sentences. Therefore, PISCES might involve the same biases and toxic behaviors exhibited by language models, pre-training corpora and Google Translation.

Limitations

While we show that PISCES outperforms mBART-50 on WikiLingua (Ladhak et al., 2020), there are some limitations worth considering in future work: (1) PISCES still struggles to generate summaries in unseen languages (Section 5.3); (2) In this work, we focus on six languages in total, and future work could extend our method to more languages.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No.62072323, 62102276), Shanghai Science and Technology Innovation Action Plan (No. 22511104700), the Natural Science Foundation of Jiangsu Province (Grant No. BK20210705), the Natural Science Foundation of Educational Commission of Jiangsu Province, China (Grant No. 21KJD520005) and the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

Yu Bai, Heyan Huang, Kai Fan, Yang Gao, Zewen Chi, and Boxing Chen. 2021. Bridging the gap: Cross-lingual summarization with compression rate. *ArXiv preprint*, abs/2110.07936.

Yue Cao, Hui Liu, and Xiaojun Wan. 2020a. Jointly learning to align and summarize for neural cross-lingual summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6220–6231, Online. Association for Computational Linguistics.

Yue Cao, Xiaojun Wan, Jinge Yao, and Dian Yu. 2020b. Multisumm: Towards a unified model for multi-lingual abstractive summarization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):11–18.

Yulong Chen, Ming Zhong, Xuefeng Bai, Naihao Deng, Jing Li, Xianchao Zhu, and Yue Zhang. 2022. The cross-lingual conversation summarization challenge. *arXiv preprint arXiv:2205.00379*.

Zewen Chi, Li Dong, Shuming Ma, Shaohan Huang, Saksham Singhal, Xian-Ling Mao, Heyan Huang, Xia Song, and Furu Wei. 2021. mT6: Multilingual pretrained text-to-text transformer with translation pairs. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1671–1683, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zewen Chi, Li Dong, Furu Wei, Wenhui Wang, Xian-Ling Mao, and Heyan Huang. 2020. Cross-lingual natural language generation via pre-training. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7570–7577. AAAI Press.

Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2022. MSAMSum: Towards benchmarking multi-lingual dialogue summarization. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 1–12, Dublin, Ireland. Association for Computational Linguistics.

George Giannakopoulos, Jeff Kubina, John Conroy, Josef Steinberger, Benoit Favre, Mijail Kabadjov, Udo Kruschwitz, and Massimo Poesio. 2015. MultiLing 2015: Multilingual summarization of single and multi-documents, on-line fora, and call-center conversations. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 270–274, Prague, Czech Republic. Association for Computational Linguistics.

Tahmid Hasan, Abhik Bhattacharjee, Wasi Uddin Ahmad, Yuan-Fang Li, Yong-Bin Kang, and Rifat Shahriyar. 2021a. Crosssum: Beyond english-centric cross-lingual abstractive text summarization for 1500+ language pairs. *ArXiv*, abs/2112.08804.

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

Yichong Huang, Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2021. The factual inconsistency problem in abstractive text summarization: A survey. *arXiv preprint arXiv:2104.14839*.

Timo Johner, Abhik Jana, and Chris Biemann. 2021. Error analysis of using BART for multi-document summarization: A study for English and German language. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 391–397, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Anton Leuski, Chin-Yew Lin, Liang Zhou, Ulrich Germann, Franz Josef Och, and Eduard Hovy. 2003a. Cross-lingual c*st*rd: English access to hindi information. *ACM Transactions on Asian Language Information Processing*, 2(3):245–269.

Anton Leuski, Chin-Yew Lin, Liang Zhou, Ulrich Germann, Franz Josef Och, and Eduard H. Hovy. 2003b. Cross-lingual c*st*rd: English access to hindi information. *ACM Trans. Asian Lang. Inf. Process.*, 2:245–269.

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, Online. Association for Computational Linguistics.

Qian Li, Shu Guo, Yang Luo, Cheng Ji, Lihong Wang, Jiawei Sheng, and Jianxin Li. 2023. Attribute-consistent knowledge graph representation learning for multi-modal entity alignment. *Proceedings of the ACM Web Conference 2023*.

Yunlong Liang, Fandong Meng, Jinan Xu, Jiaan Wang, Yufeng Chen, and Jie Zhou. 2022a. Summary-oriented vision modeling for multimodal abstractive summarization. *arXiv preprint arXiv:2212.07672*.

Yunlong Liang, Fandong Meng, Chulun Zhou, Jinan Xu, Yufeng Chen, Jinsong Su, and Jie Zhou. 2022b. A variational hierarchical model for neural cross-lingual summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2088–2099, Dublin, Ireland. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Marina Litvak, Natalia Vanetik, Mark Last, and Elena Churkin. 2016. MUSEEC: A multilingual text summarization tool. In *Proceedings of ACL-2016 System Demonstrations*, pages 73–78, Berlin, Germany. Association for Computational Linguistics.

Thong Nguyen and Luu Anh Tuan. 2022. Improving neural cross-lingual summarization via employing optimal transport distance for knowledge distillation. *Proc. of AAAI*.

Constantin Orăsan and Oana Andreea Chiorean. 2008. Evaluation of a cross-lingual Romanian-English multi-document summariser. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Laura Perez-Beltrachini and Mirella Lapata. 2021. Models and datasets for cross-lingual summarisation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9408–9423, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Carol Pfaff. 1979. Constraints on language mixing: Intrasentential code-switching and borrowing in spanish/english. *Language*, 55:291.

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. *Journal of Machine Learning Research*, 21(140):1–67.

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

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online. Association for Computational Linguistics.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Natalia Vanetik and Marina Litvak. 2015. Multilingual summarization with polytope model. In *SIGDIAL Conference*.

Daniel Varab and Natalie Schluter. 2021. MassiveSumm: a very large-scale, very multilingual, news summarisation dataset. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10150–10161, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Xiaojun Wan. 2011. Using bilingual information for cross-language document summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1546–1555, Portland, Oregon, USA. Association for Computational Linguistics.

Xiaojun Wan, Huiying Li, and Jianguo Xiao. 2010. Cross-language document summarization based on machine translation quality prediction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 917–926, Uppsala, Sweden. Association for Computational Linguistics.

Bin Wang, Chen Zhang, Yan Zhang, Yiming Chen, and Haizhou Li. 2022a. Analyzing and evaluating faithfulness in dialogue summarization. *arXiv preprint arXiv:2210.11777*.

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, Online. Association for Computational Linguistics.

Jiaan Wang, Yunlong Liang, Fandong Meng, Beiqi Zou, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2023. Zero-shot cross-lingual summarization via large language models. *arXiv preprint*.

Jiaan Wang, Fandong Meng, Ziyao Lu, Duo Zheng, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2022b. ClidSum: A benchmark dataset for cross-lingual dialogue summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7716–7729, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jiaan Wang, Fandong Meng, Tingyi Zhang, Yunlong Liang, Jiarong Xu, Zhixu Li, and Jie Zhou. 2022c. Understanding translationese in cross-lingual summarization. *arXiv preprint arXiv:2212.07220*.

Jiaan Wang, Fandong Meng, Duo Zheng, Yunlong Liang, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2022d. A Survey on Cross-Lingual Summarization. *Transactions of the Association for Computational Linguistics*, 10:1304–1323.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. 2022. PRIMERA: Pyramid-based masked sentence pre-training for multi-document summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5245–5263, Dublin, Ireland. Association for Computational Linguistics.

Ruochen Xu, Chenguang Zhu, Yu Shi, Michael Zeng, and Xuedong Huang. 2020. Mixed-lingual pre-training for cross-lingual summarization. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference* *on Natural Language Processing*, pages 536–541, Suzhou, China. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Jin-ge Yao, Xiaojun Wan, and Jianguo Xiao. 2015. Phrase-based compressive cross-language summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 118–127, Lisbon, Portugal. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020a. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.

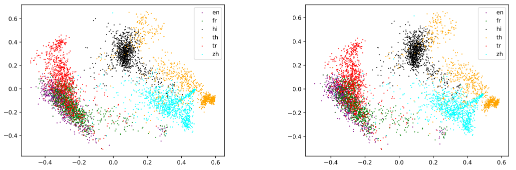
Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Shaohui Zheng, Zhixu Li, Jiaan Wang, Jianfeng Qu, An Liu, Lei Zhao, and Zhigang Chen. 2023. Long-document cross-lingual summarization. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, WSDM '23, page 1084–1092, New York, NY, USA. Association for Computing Machinery.

Junnan Zhu, Qian Wang, Yining Wang, Yu Zhou, Jiajun Zhang, Shaonan Wang, and Chengqing Zong. 2019. NCLS: Neural cross-lingual summarization. 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)*, pages 3054–3064, Hong Kong, China. Association for Computational Linguistics.

Junnan Zhu, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2020. Attend, translate and summarize: An efficient method for neural cross-lingual summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1309–1321, Online. Association for Computational Linguistics.

A Word Embeddings of the Unseen Language and Other Languages



(a) mBART (`M2MS`) (b) mBART (`U-CLS`)

Figure 3: Visualization of word embeddings from mBART (`M2MS`) and mBART (`U-CLS`). Tr is the unseen language.

To verify the word embeddings of the unseen language drift away from those of other languages after adding the monolingual training data, based on MUSE dictionary, we choose top frequent 1000 English words and the words with the same meaning in other five languages (*i.e.*, Fr, Hi, Zh, Th and Tr). Then, we calculate the embeddings of these words based on mBART (`M2MS`) and mBART (`U-CLS`), respectively. For the word that consists of multiple tokens, the word embedding is the average of embeddings of those tokens. As shown in Figure 3, we utilize Principal Component Analysis (PCA) to visualize the word embeddings from mBART (`M2MS`) and mBART (`U-CLS`). In the PCA space, we further calculate the central point of each language by averaging the word embeddings in the language. Then, we find the average distance between the central point of Tr and other languages is 0.426 / 0.407 for mBART (`M2MS`) / mBART (`U-CLS`). This distance in vanilla mBART-50 (Tang et al., 2021) is 0.398. Therefore, the monolingual training data used in mBART (`M2MS`) makes the word embeddings of the unseen language drift away from those of other languages.

B Implementation Details

Pre-Training Details. We use mBART-50 (Tang et al., 2021) as the meta pre-trained model, and futher pre-train it via cross-lingual and task-specific pre-training stages. The implementation of mBART-50 is based on the Transformers (Wolf et al., 2020) library with default settings (12 encoder layers, 12 decoder layers and 1024 hidden states). In cross-lingual pre-training, we dynamically mask 0-15% tokens in the source-language sentences, and construct 20.6M samples from OPUS parallel corpora (Tiedemann and Thottingal, 2020). In task-specific pre-training, we construct 3.1M training samples from mC4 corpus (Xue et al., 2021). We set the total length of gap sentences to k% of the document length, and k is dynamically

Direction	MultiUN	CCMatrix	CCAligned	MultiCCAligned	XLEnt	Europarl	QED	TED	WMT	**Sum**
En⇔Fr	-	-	-	-	-	349291	152623	77188	4648	583750
En⇔Hi	-	2959722	-	-	405366	-	1211	9039	568	3375906
En⇔Th	-	-	1947729	-	246976	-	52140	30765	-	2277610
En⇔Tr	-	-	2496997	-	761750	-	94212	72674	3819	3429452
En⇔Zh	-	-	-	-	1258289	-	-	3158	3658	1265105
Fr⇔Hi	-	-	-	619040	97082	-	660	8816	-	725598
Fr⇔Th	-	-	-	737469	67292	-	34418	30024	-	869203
Fr⇔Tr	-	-	-	1321431	183282	-	61412	69931	-	1636056
Fr⇔Zh	1494829	-	-	-	211039	-	2041	3088	-	1710997
Hi⇔Th	-	-	-	436284	65870	-	484	4526	-	507164
Hi⇔Tr	-	1099853	-	-	111573	-	544	8384	-	1220354
Hi⇔Zh	-	445148	-	-	97732	-	15	650	-	543545
Th⇔Tr	-	-	-	617566	86156	-	40026	29602	-	773350
Th⇔Zh	-	-	-	-	54637	-	2390	2169	-	59196
Tr⇔Zh	-	1435286	-	-	169774	-	1885	3125	-	1610070
Total	1494829	5940009	4444726	3731790	3816818	349291	444061	353139	12693	20587356

Table 8: Statistics of the constructed cross-lingual pre-training samples. Each entry shows the number of samples for each language pair in the corresponding corpus.

En⇔Fr	En⇔Hi	En⇔Th	En⇔Tr	En⇔Zh	Fr⇔Hi	Fr⇔Th	Fr⇔Tr	Fr⇔Zh	Hi⇔Th	Hi⇔Tr
190916	190916	190916	190916	88636	188351	190916	190916	190916	158518	190578
Hi⇔Zh	Th⇔Tr	Th⇔Zh	Tr⇔Zh	En⇒En	Fr⇒Fr	Hi⇒Hi	Th⇒Th	Tr⇒Tr	Zh⇒Zh	**Total**
172039	190916	24160	190916	95458	95458	95458	95458	95458	95458	3113274

Table 9: Statistics of the constructed task-specific pre-training samples.

selected from $[5, 10, 15]$. The pre-defined λ in the round-trip translation is 0.7. All experimental results listed in this paper are the average of 3 runs.

Table 8 and Table 9 show the statistics of the constructed samples in the cross-lingual pre-training and task-specific pre-training stages, respectively. The cross-lingual pre-training and task-specific pre-training stages are conducted on 8 NVIDIA Tesla V100 GPUs with 32GB memory. In the cross-lingual pre-training stage, we pre-train the model for 150K steps, with early stopping, 32 batch size, 3e-5 learning rate following Xiao et al. (2022) and 10K warmup steps. In the task-specific pre-training stage, we pre-train the model for 100K steps, with early stopping, 4 batch size, 3e-5 learning rate and 10K warmup steps.

Fine-Tuning and Testing Details. In the fine-tuning stage, we fine-tune the PISCES model on 8 NVIDIA Tesla V100 GPUs (32G) with 4 batch size, 10 epochs, 2K warmup steps, 3e-5 learning rate, and set the maximum number of tokens for input sequences to 1024. To balance the high-resource and low-resource language data, following Xue et al. (2021), we sample the training examples according to the probability $p(D) \propto |D|^{\alpha}$, where $p(D)$ is the probability of sampling training examples from a give direction during fine-tuning and $|D|$ is the number of original examples in the direction. We set the hyperparameter α to 0.5. To fine-tune mT5 baseline on M2MS, the language tags (*e.g.*, `<En>` and `<Zh>`) are appended at the inputs of both encoder and decoder sides. In the test process, we set the beam size and the maximum decoded length to 5 and 128, respectively.

C Experiments on CrossSum

C.1 Data Statistics.

Table 10 lists the data statistics of the CrossSum dataset (Hasan et al., 2021a) used in our experiments. The data splitting mainly inherits from the original CrossSum except for zero-shot directions and monolingual directions: (1) If the number of samples in a direction (*e.g.*, Fr⇒Hi) is less than 1k, we will regard the direction as a zero-shot direction and evenly split its samples into validation and test sets. (2) Considering the number of samples in cross-lingual directions is hundred-level or thousand-level, we truncate the number of samples in each monolingual direction (*e.g.*, En⇒En) to 10k to make a balance. The corresponding splitting follows 8:1:1. If the number of samples in a monolingual direction (*e.g.*, Th⇒Th) is less than 10k, its splitting follows the original CrossSum.

C.2 Experimental Results.

Table 11 shows the experimental results on CrossSum. Our PISCES outperforms mBART-50 by 2.3 ROUGE-1, 2.0 ROUGE-2, 2.0 ROUGE-L and 1.3 BERTSCORE in the average of all directions, which

Trg Src		En	Fr	Hi	Zh	Th	Tr
En	# Samples	8000 / 1000 / 1000	1513 / 188 / 188	3784 / 463 / 481	3981 / 497 / 497	816 / 102 / 102	4542 / 568 / 566
	# Avg. Tokens	638.7 / 30.6	1013.0 / 43.2	899.9 / 41.0	914.6 / 35.5	1058.3 / 51.3	880.8 / 37.6
Fr	# Samples	1513 / 188 / 188	8000 / 1000 / 1000	- / 308 / 308	- / 174 / 174	- / 92 / 93	- / 414 / 415
	# Avg. Tokens	1124.3 / 33.7	710.9 / 40.8	1048.5 / 40.7	1358.3 / 37.6	1501.7 / 47.9	1058.3 / 38.9
Hi	# Samples	3784 / 463 / 481	- / 308 / 308	8000 / 1000 / 1000	1107 / 135 / 137	- / 189 / 189	2956 / 369 / 369
	# Avg. Tokens	862.0 / 31.6	1106.5 / 39.1	775.4 / 40.2	804.3 / 33.6	1186.3 / 49.4	712.8 / 34.3
Zh	# Samples	3981 / 497 / 497	- / 174 / 174	1107 / 135 / 137	8000 / 1000 / 1000	- / 134 / 135	1209 / 151 / 151
	# Avg. Tokens	725.0 / 32.7	1082.0 / 41.9	690.7 / 41.9	768.0 / 40.4	1059.7 / 52.7	642.4 / 36.6
Th	# Samples	816 / 102 / 102	- / 92 / 93	- / 189 / 189	- / 134 / 135	6616 / 826 / 826	- / 238 / 239
	# Avg. Tokens	957.1 / 34.5	1095.2 / 40.6	985.3 / 42.0	1036.3 / 38.7	1055.5 / 62.1	912.2 / 39.7
Tr	# Samples	4542 / 568 / 566	- / 414 / 415	2956 / 369 / 369	1209 / 151 / 151	- / 238 / 239	8000 / 1000 / 1000
	# Avg. Tokens	619.4 / 31.8	775.2 / 41.2	579.3 / 39.0	591.5 / 34.4	762.7 / 53.2	704.9 / 40.2

Table 10: Statistics of CrossSum used in our experiments. *# Samples* denotes the number of samples in training / validation / test set. *# Avg. Tokens* represents the average tokens in the documents and summaries, respectively. gray indicates the zero-shot directions.

Trg Src	Model	En	Fr	Hi	Zh	Th	Tr	Avg.
En	mT5 (580M)	30.1 / 8.3 / 22.3 / 66.3	30.7 / 10.4 / 22.8 / 66.2	30.2 / 8.9 / 24.8 / 67.4	26.1 / 6.6 / 22.6 / 65.8	27.4 / 8.6 / 21.8 / 60.1	25.8 / 9.9 / 22.5 / 66.7	28.4 / 8.8 / 22.8 / 65.4
	mBART (610M)	31.2 / 8.7 / 22.8 / 66.9	32.8 / 12.4 / 24.5 / 66.9	32.6 / 9.5 / 25.5 / 68.3	29.6 / 8.2 / 24.3 / 67.0	30.8 / 10.4 / 23.4 / 62.9	26.3 / 10.2 / 22.8 / 67.2	30.6 / 9.9 / 23.9 / 66.5
	Pisces (610M)	32.0 / 9.1 / 23.7 / 67.4	33.6 / 13.4 / 25.6 / 67.6	33.4 / 10.5 / 26.4 / 68.9	30.5 / 8.6 / 24.9 / 67.5	31.3 / 11.7 / 24.2 / 64.1	27.1 / 10.9 / 23.5 / 67.5	31.3 / 10.7 / 24.7 / 67.2
Fr	mT5 (580M)	30.7 / 10.1 / 23.4 / 66.6	30.8 / 11.9 / 23.5 / 66.2	33.0 / 12.1 / 27.3 / 68.1	39.3 / 21.5 / 33.1 / 69.9	35.6 / 15.9 / 29.2 / 64.6	25.3 / 10.7 / 22.3 / 65.8	32.5 / 13.7 / 26.5 / 66.9
	mBART (610M)	31.9 / 10.3 / 23.8 / 67.4	32.0 / 12.9 / 24.3 / 66.6	36.0 / 16.6 / 30.2 / 69.8	41.3 / 23.4 / 36.7 / 70.9	37.1 / 17.4 / 30.8 / 65.3	29.5 / 14.1 / 26.1 / 68.2	34.6 / 15.8 / 28.7 / 68.0
	Pisces (610M)	33.5 / 11.7 / 25.8 / 68.2	32.7 / 13.4 / 25.0 / 67.1	39.7 / 19.5 / 33.9 / 71.5	43.8 / 25.7 / 38.7 / 73.0	42.8 / 25.3 / 35.7 / 69.2	33.9 / 18.9 / 30.4 / 70.1	37.7 / 19.1 / 31.6 / 69.9
Hi	mT5 (580M)	29.7 / 9.3 / 23.2 / 67.2	29.6 / 10.8 / 23.3 / 66.1	32.0 / 11.3 / 25.7 / 67.4	28.6 / 8.3 / 24.0 / 66.3	29.8 / 11.0 / 23.8 / 62.6	22.0 / 7.4 / 19.8 / 65.5	28.6 / 9.7 / 23.3 / 65.9
	mBART (610M)	31.5 / 9.9 / 24.0 / 67.7	32.5 / 13.3 / 25.5 / 67.4	32.9 / 11.8 / 26.0 / 67.7	29.4 / 8.9 / 24.6 / 66.9	33.7 / 15.1 / 27.6 / 65.0	22.6 / 7.8 / 19.7 / 65.7	30.4 / 11.1 / 24.6 / 66.7
	Pisces (610M)	31.8 / 9.9 / 24.1 / 68.0	35.3 / 15.9 / 28.0 / 68.7	33.8 / 12.5 / 26.8 / 68.3	32.5 / 10.8 / 27.3 / 68.9	38.3 / 19.0 / 31.3 / 67.0	23.9 / 8.6 / 21.0 / 66.4	32.6 / 12.8 / 26.4 / 67.9
Zh	mT5 (580M)	30.9 / 9.8 / 23.1 / 66.5	31.1 / 13.1 / 24.9 / 66.5	31.2 / 9.2 / 24.7 / 68.3	32.5 / 11.6 / 26.8 / 67.0	29.2 / 9.3 / 23.8 / 62.5	24.0 / 9.2 / 21.7 / 66.8	29.8 / 10.4 / 24.2 / 66.3
	mBART (610M)	32.4 / 10.6 / 24.4 / 67.4	35.7 / 17.7 / 28.6 / 68.5	33.5 / 10.9 / 27.5 / 69.5	33.1 / 11.6 / 26.9 / 67.2	35.3 / 15.2 / 28.6 / 64.5	25.3 / 9.3 / 22.2 / 67.3	32.6 / 12.5 / 26.4 / 67.4
	Pisces (610M)	33.4 / 11.0 / 25.5 / 68.2	39.1 / 22.5 / 32.5 / 70.7	34.6 / 11.4 / 27.9 / 70.2	34.2 / 12.3 / 27.8 / 67.6	37.7 / 17.4 / 31.1 / 65.6	27.8 / 11.1 / 24.3 / 68.0	34.5 / 14.3 / 28.2 / 68.4
Th	mT5 (580M)	25.7 / 6.6 / 19.1 / 62.6	25.3 / 9.9 / 19.8 / 63.5	30.3 / 10.3 / 24.6 / 66.4	27.2 / 7.6 / 22.8 / 64.1	33.7 / 13.2 / 25.6 / 63.4	22.0 / 9.5 / 19.7 / 63.9	27.4 / 9.5 / 21.9 / 64.0
	mBART (610M)	27.1 / 6.9 / 19.7 / 63.7	27.6 / 11.0 / 21.5 / 64.2	30.8 / 12.5 / 25.1 / 66.6	33.1 / 15.8 / 28.4 / 67.5	35.9 / 14.7 / 26.9 / 65.2	27.2 / 13.0 / 24.2 / 66.4	30.3 / 12.3 / 24.3 / 65.6
	Pisces (610M)	29.5 / 9.1 / 21.7 / 65.6	38.0 / 19.6 / 30.4 / 69.4	35.8 / 15.6 / 29.1 / 69.2	37.1 / 18.5 / 32.3 / 69.2	36.4 / 15.2 / 27.3 / 65.5	30.2 / 15.4 / 26.7 / 68.2	34.5 / 15.6 / 27.9 / 67.8
Tr	mT5 (580M)	29.4 / 10.1 / 22.4 / 67.1	32.9 / 12.2 / 24.6 / 67.2	29.2 / 7.8 / 23.8 / 67.1	30.1 / 9.5 / 25.1 / 67.6	30.7 / 11.3 / 24.9 / 62.8	28.0 / 11.9 / 24.2 / 67.2	30.0 / 10.5 / 24.2 / 66.5
	mBART (610M)	32.0 / 11.1 / 24.9 / 68.0	36.0 / 17.2 / 28.9 / 68.9	32.5 / 9.5 / 26.0 / 68.6	31.5 / 9.8 / 25.5 / 67.9	37.5 / 18.1 / 31.1 / 66.4	28.8 / 12.7 / 24.9 / 67.8	33.1 / 13.1 / 26.9 / 67.9
	Pisces (610M)	33.3 / 11.5 / 25.3 / 68.6	38.3 / 18.9 / 30.8 / 70.2	33.2 / 10.2 / 26.6 / 69.2	32.2 / 10.1 / 26.0 / 68.7	40.9 / 22.0 / 34.3 / 67.7	30.8 / 14.0 / 26.5 / 68.5	34.8 / 14.4 / 28.2 / 68.8
Avg.	mT5 (580M)	29.4 / 9.0 / 22.2 / 66.0	30.1 / 11.4 / 23.2 / 66.0	31.0 / 9.9 / 25.2 / 67.5	30.6 / 10.9 / 25.7 / 66.8	31.1 / 11.5 / 24.8 / 62.7	24.5 / 9.8 / 21.7 / 66.0	29.4 / 10.4 / 23.8 / 65.8
	mBART (610M)	31.0 / 9.6 / 23.3 / 66.8	32.8 / 14.1 / 25.6 / 67.1	33.1 / 11.8 / 26.7 / 68.4	33.0 / 13.0 / 27.7 / 67.9	35.1 / 15.2 / 28.1 / 64.9	26.6 / 11.2 / 23.3 / 67.1	31.9 / 12.5 / 25.8 / 67.0
	Pisces (610M)	32.2 / 10.4 / 24.3 / 67.7	36.2 / 17.3 / 28.7 / 69.0	35.1 / 13.3 / 28.4 / 69.5	35.1 / 14.3 / 29.5 / 69.1	37.9 / 18.4 / 30.7 / 66.5	29.0 / 13.2 / 25.4 / 68.1	34.2 / 14.5 / 27.8 / 68.3

Table 11: Experimental results on CrossSum (ROUGE-1 / ROUGE-2 / ROUGE-L / BERTSCORE). gray indicates the zero-shot directions. "Avg." denotes the average scores w.r.t each row, each column or all directions.

verifies the effectiveness of PISCES. For the average results in all zero-shot directions, mBART-50 achieves 33.8, 15.7, 28.1 and 67.1 in terms of ROUGE-1/2/L and BERTSCORE. The counterparts of PISCES are 37.9, 19.6, 31.8 and 69.3, showing its superiority in the zero-shot directions.

D Full Results on WikiLingua

Table 12 shows the experimental results in terms of ROUGE-1, ROUGE-2 and ROUGE-L, respectively.

E Ablations in Conventional Zero-Shot Directions

Table 13 shows the ablation results in all conventional zero-shot directions.

F Error Analysis

We first randomly select 100 summaries generated by PISCES on WikiLingua (En⇒Zh). After manually examining the generated summaries, we find the following major error types:

- **Missing Information**: part of the information in the ground truth summary is not mentioned in the generated summary. This is the most frequent error type, and accounts for 39% of the generated summaries.
- **Faithfulness**: the generated summary involves information that is inconsistent with (or not presented in) the source document. We find 32% of the summaries have this error.
- **Redundancy**: the generated summary contains additional information beyond the ground truth summary. 17% of the generated summaries contain this error.
- **Foreign Words**: the generated summary involves words in another language. 9% of the generated Chinese summaries involve some (typically one or two) words in another language.

Redundancy and missing information are two major flaws caused by the limited summarization ability (Johner et al., 2021). Faithfulness error is another error type that has been noticed in the summarization research field recently (Huang et al., 2021). The neural generative summarization models are highly prone to generate factual inconsistency errors (Huang et al., 2021). Some studies (Kryscinski et al., 2020; Wang et al., 2022a) show that over

Trg Src	Model	En	Fr	Hi	Zh	Th	Tr
En	mT5	40.9 / 17.7 / 34.2	36.0 / 15.0 / 29.4	30.1 / 8.9 / 23.5	37.0 / 13.4 / 32.1	38.6 / 17.8 / 33.4	3.3 / 0.2 / 3.0
	mBART	41.9 / 18.2 / 34.9	37.2 / 15.8 / 30.3	31.7 / 9.6 / 24.5	37.9 / 13.9 / 32.7	39.5 / 18.5 / 34.0	3.2 / 0.2 / 3.0
	PISCES	42.8 / 18.8 / 35.5	38.1 / 16.4 / 31.1	33.7 / 10.8 / 26.6	38.8 / 14.2 / 33.3	40.9 / 19.3 / 35.6	4.5 / 0.7 / 4.2
Fr	mT5	37.0 / 14.3 / 31.2	38.6 / 18.3 / 32.6	27.0 / 7.4 / 22.1	35.6 / 12.4 / 31.3	36.4 / 15.9 / 32.1	3.1 / 0.2 / 2.9
	mBART	38.2 / 15.0 / 31.7	39.2 / 17.9 / 32.0	28.7 / 7.9 / 22.3	36.9 / 12.8 / 31.6	37.9 / 16.6 / 32.6	3.1 / 0.2 / 3.0
	PISCES	39.2 / 15.4 / 32.4	40.0 / 18.3 / 32.5	31.3 / 8.8 / 24.2	37.4 / 13.0 / 31.9	39.2 / 17.3 / 33.6	4.1 / 0.6 / 3.8
Hi	mT5	36.9 / 14.2 / 30.3	31.9 / 11.6 / 25.6	34.9 / 11.9 / 27.2	32.1 / 9.6 / 27.5	34.0 / 14.1 / 29.2	3.2 / 0.3 / 3.0
	mBART	37.9 / 14.6 / 30.8	32.8 / 12.2 / 25.9	35.6 / 12.5 / 27.8	33.2 / 10.6 / 28.2	35.4 / 14.6 / 30.1	3.4 / 0.3 / 3.2
	PISCES	39.8 / 16.0 / 32.7	35.7 / 14.1 / 28.4	37.2 / 13.6 / 28.8	35.9 / 11.8 / 30.7	38.1 / 16.6 / 32.6	4.0 / 0.6 / 3.8
Zh	mT5	38.3 / 14.3 / 31.5	34.5 / 13.8 / 28.3	26.0 / 6.2 / 20.3	41.1 / 16.5 / 35.5	36.1 / 14.7 / 30.8	3.3 / 0.3 / 3.2
	mBART	39.2 / 15.1 / 32.0	36.0 / 14.5 / 29.0	27.0 / 6.6 / 20.8	41.7 / 17.0 / 35.9	36.8 / 15.3 / 31.4	3.4 / 0.2 / 3.2
	PISCES	40.3 / 15.8 / 33.0	37.4 / 15.4 / 29.9	29.6 / 8.2 / 23.2	42.5 / 17.5 / 36.3	39.2 / 17.0 / 33.6	4.3 / 0.6 / 4.0
Th	mT5	37.6 / 15.0 / 30.7	34.1 / 13.9 / 27.8	26.2 / 6.8 / 20.5	34.0 / 11.0 / 28.7	41.1 / 20.3 / 36.0	3.3 / 0.3 / 3.2
	mBART	38.5 / 15.4 / 31.9	35.6 / 14.2 / 28.3	27.8 / 7.3 / 21.4	34.6 / 11.3 / 29.0	42.2 / 20.8 / 36.2	3.3 / 0.3 / 3.1
	PISCES	40.2 / 16.6 / 33.2	37.2 / 15.4 / 29.7	31.0 / 9.3 / 23.9	36.9 / 12.7 / 31.3	43.3 / 21.7 / 37.5	4.3 / 0.7 / 4.0
Tr	mT5	14.2 / 2.2 / 12.1	14.9 / 2.9 / 12.2	11.2 / 1.4 / 10.8	19.2 / 2.6 / 15.9	20.0 / 4.2 / 17.9	3.0 / 0.2 / 2.8
	mBART	15.7 / 2.6 / 13.4	16.0 / 3.2 / 13.2	14.9 / 2.3 / 12.6	19.9 / 3.0 / 17.6	21.4 / 4.8 / 19.3	3.1 / 0.2 / 3.0
	PISCES	28.3 / 8.8 / 23.4	27.3 / 9.3 / 22.2	23.2 / 5.5 / 18.5	29.8 / 8.2 / 25.7	30.8 / 11.3 / 26.7	5.3 / 0.8 / 5.0

Table 12: Experimental results on WikiLingua (ROUGE-1 / ROUGE-2 / ROUGE-L). Green , light green and gray indicate the high-resource , low-resource and zero-shot directions, respectively.

	Fr⇒Hi	Hi⇒Fr	Hi⇒Zh	Zh⇒Hi
PISCES	**21.4 / 69.1**	**26.1 / 72.9**	**26.1 / 70.4**	**20.3 / 68.5**
w/o TS	20.7 / 68.6	25.2 / 72.8	25.1 / 69.9	19.5 / 67.9
w/o CL	20.6 / 68.8	25.2 / **72.9**	25.3 / 70.0	19.5 / 67.8
w/o TS & CL	19.6 / 68.1	23.6 / 72.1	24.0 / 69.1	18.1 / 66.9
	Hi⇒Th	Th⇒Hi	Zh⇒Th	Th⇒Zh
PISCES	**29.1 / 68.5**	**21.4 / 69.0**	**29.9 / 68.9**	**27.0 / 71.0**
w/o TS	28.2 / 68.1	20.3 / 68.3	28.7 / 68.3	25.8 / 70.3
w/o CL	28.0 / 68.0	20.3 / 68.4	29.0 / 68.5	26.0 / 70.4
w/o TS & CL	26.7 / 67.4	18.8 / 67.4	27.8 / 67.6	25.0 / 69.4

Table 13: Results of ablation studies.

30% of the summaries generated by neural models contain this error. We confirm that CLS also involves the faithfulness error. Future work could give deeper and more fine-grained analyses of this error type.

The issue of foreign words could also refer to the code-switching phenomenon (Pfaff, 1979). Note that the generated foreign words are not limited in the source language. In several cases, the generated Chinese summaries of the given English documents even involve Thai words. We also find the semantics of these foreign words are typically coherent with their context. This error type might be caused by the cross-lingual pre-training (which bridges the representation gap of parallel words in different languages) in PISCES.

ACL 2023 Responsible NLP Checklist

A For every submission:

□ A1. Did you describe the limitations of your work?
Left blank.

□ A2. Did you discuss any potential risks of your work?
Left blank.

□ A3. Do the abstract and introduction summarize the paper's main claims?
Left blank.

□ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B □ Did you use or create scientific artifacts?

Left blank.

□ B1. Did you cite the creators of artifacts you used?
Left blank.

□ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Left blank.

□ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Left blank.

□ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Left blank.

□ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Left blank.

□ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Left blank.

C □ Did you run computational experiments?

Left blank.

□ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Left blank.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

□ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Left blank.

□ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Left blank.

□ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Left blank.

D □ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Left blank.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Left blank.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Left blank.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Left blank.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Left blank.