MSAMSum: Towards Benchmarking Multi-lingual Dialogue Summarization

Anonymous ACL submission

Abstract

Dialogue summarization helps users capture salient information from various types of dialogues has received much attention recently. However, current works mainly focus on English dialogue summarization, leaving other languages under exploration. Therefore, we present a multi-lingual dialogue summarization dataset, namely MSAMSum, which covers dialogue-summary pairs in six languages. Specifically, we derive MSAMSum from the standard SAMSum (Gliwa et al., 2019) using sophisticated translation techniques and further employ two methods to ensure the integral translation quality and summary factual consistency. Given the proposed MSAMum, we systematically set up five multi-lingual settings for this task, including a novel mix-lingual dialogue summarization setting. To illustrate the utility of our dataset, we benchmark various experiments with pre-trained models under different settings and report results in both supervised and zero-shot manners. We also discuss some future works towards this task to motivate future researches[1].

1 Introduction

Recent years have witnessed increased interest in dialogue summarization (Feng et al., 2021a; Tuggener et al., 2021). It aims to distill the most important information from various types of dialogues, which can alleviate the problem of communication data overload. Towards this research direction, various datasets have been proposed to promote this task.

The AMI (Carletta et al., 2005) and ICSI (Janin et al., 2003) datasets provide the initial opportunity for meeting summarization. With the advent of data-hungry neural models and pre-trained language models, Gliwa et al. (2019) come up with the first high quality large-scale dialogue summarization dataset, namely SAMSum, which resurges this task. Then, various datasets are proposed to meet different needs and scenarios (Chen et al., 2021a; Malykh et al., 2020; Rameshkumar and Bailey, 2020; Zhong et al., 2021; Zhu et al., 2021; Chen et al., 2021b; Zhang et al., 2021; Fabbri et al., 2021). Despite the encouraging progresses achieved, current works overwhelmingly focused on English. Meanwhile, with the help of instantaneous translation systems[2], a dialogue involving multinational participants becomes more and more common and frequent. Therefore, it is valuable to provide them with dialogue summaries in a preferred language.

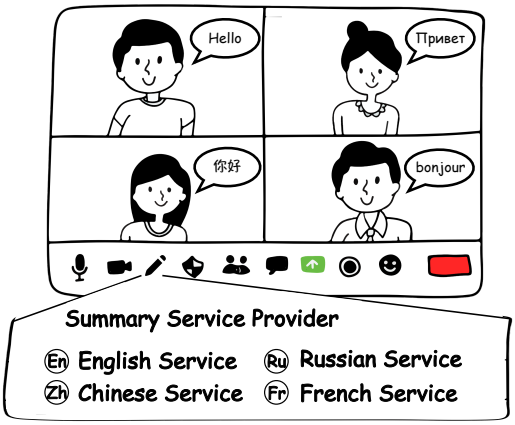


Figure 1: A multi-lingual meeting scenario, in which multinational people participate in one meeting concurrently. It is valuable to provide them with summaries in a preferred language.

To this end, we propose a *multi-lingual dialogue summarization task*. The practical benefits of this task are twofold: it not only provides rapid access to the salient content, but also enables the dissemination of relevant content across participants of other languages. Intuitively, to achieve this goal, we need to answer two key questions, one is *Where do we get data resources for this multi-lingual research?* the other is *How do we perform various multi-lingual settings?*

[1] Our dataset and codes will be available at: xxx.

[2] https://translatebyhumans.com/en/services/interpretation/zoom/

For the first question, we seek for potential available resources that can support our multi-lingual research. Although creating English datasets has proven feasible, the need for dialogues and summary-written experts in different languages makes the collection of multi-lingual datasets highly costing or even intractable. To mitigate this challenge, we devote our efforts to constructing the multi-lingual dataset via sophisticated translation techniques following Zhu et al. (2019). Firstly, we select SAMSum (Gliwa et al., 2019) as our source English dataset because of its large scale and wide domain coverage. Then, we translate it into five other official languages of the United Nations via high-performance translation API, including Chinese, French, Arabic, Russian and Spanish. Furthermore, We employ two methods: *round-trip translation* and *textual entailment* to filter out low-quality translations and ensure the factual consistency at both the dialogue-level and summary-level. Finally, *we obtain our MSAMSum dataset as the data resource for this multi-lingual research*.

For the second question, given the well-constructed MSAMsum dataset, we set up various settings for our multi-lingual dialogue summarization task, including ONE-TO-ONE, MANY-TO-ONE, ONE-TO-MANY and MANY-TO-MANY. The ONE-TO-ONE setting can be further divided into *Mono-lingual* and *Cross-lingual* settings. To further boost the research on multi-lingual dialogue summarization, we creatively propose one new setting, namely MIX-TO-MANY, which takes a mix-lingual dialogue as input and produce summaries in different languages. This setting is in line with the real world scenario that multinational participants can use their mother tongue to communicate with each other by means of instantaneous translation systems (depicted in Figure 1). To sum up, *we set up five settings for the research on the whole scene of multi-lingual dialogue summarization*.

To illustrate the utility of our MSAMSum, we conduct extensive experiments under five multi-lingual settings based on the current multi-lingual pre-trained model mBART-50 (Tang et al., 2020), and evaluate it in both supervised and zero-shot manners. The results reveal the feasibility of multi-lingual dialogue summarization task. The case study also shows that the multi-lingual model is able to produce fluent and factual consistency summaries in different languages. We further conclude several future works to prompt future researches.

2 Related Work

2.1 Multi-lingual Summarization

Multi-lingual summarization is a valuable research direction, which can benefit users from various countries (Cao et al., 2020). Especially, cross-lingual summarization, which receives a document in a source language and produces a summary in a another language, has attracted lots of research attentions (Wan et al., 2010). For a long time, pipeline systems combining both machine translation and summarization tools are used to solve this problem (Ouyang et al., 2019). However, pipeline systems do have their own drawbacks, like error propagation and system latency. Therefore, researchers turn to end-to-end neural methods. Zhu et al. (2019) first propose two cross-lingual summarization datasets using machine translation techniques. Afterwards, various models (Zhu et al., 2020b; Xu et al., 2020; Wang et al., 2021) and datasets (Ladhak et al., 2020; Hasan et al., 2021; Varab and Schluter, 2021) are proposed for this task. These works have achieved great progresses and have proved the feasibility of end-to-end multi-lingual summarization. In this paper, for the first time, we study the dialogue summarization task under various multi-lingual settings.

2.2 Dialogue Summarization

The earlier publicly available meeting datasets AMI (Carletta et al., 2005) and ICSI (Janin et al., 2003) have prompted dialogue summarization for a long time. Recently, the introduction of SAMSum (Gliwa et al., 2019) dataset has resurged this direction. Researchers propose various methods to tackle this problem by incorporating auxiliary information, modeling the interaction and dealing with long input sequences (Chen and Yang, 2020; Feng et al., 2021b; Zhu et al., 2020a). Additionally, various valuable datasets are carried out to meet different needs, which further accelerate the development of dialogue summarization (Zhong et al., 2021; Zhu et al., 2021; Zhang et al., 2021). What is more, Mehnaz et al. (2021) study dialogue summarization under the Hindi-English code-switched setting and get the best performance based on multi-lingual pre-trained language models. Nonetheless, the current datasets and models are mainly tailored for English, which leave other languages under exploration. To mitigate this challenge, we propose the MSAMSum to study the multi-lingual dialogue summarization task.

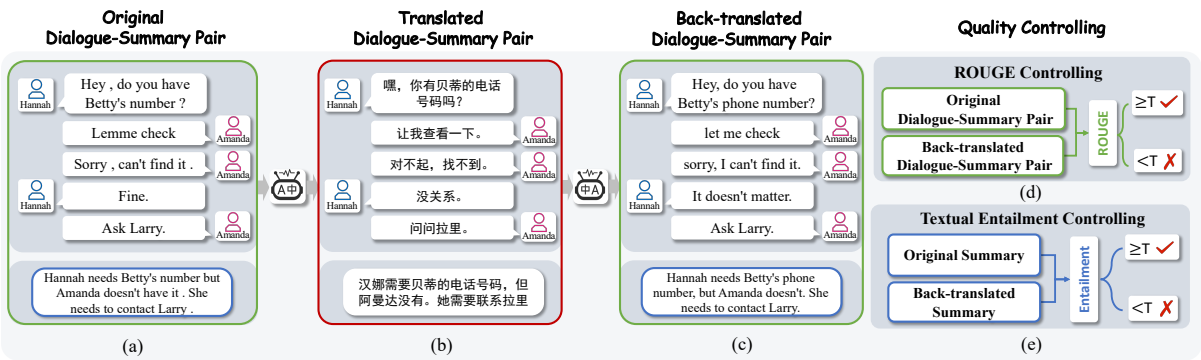


Figure 2: Illustration of our data construction process. (a) Given the original English data in the SAMSum (Gliwa et al., 2019), we translate it into another language (e.g., Chinese). Furthermore, we employ two quality controlling methods: *round-trip translation* and *textual entailment*. (c) For the first method, we back-translate the Chinese data into English and (d) calculate the ROUGE score between the original one and the back-translated one. (e) For the second one, we calculate the entailment score between back-translated summary and the original summary. If both scores exceed the pre-defined threshold, the translated dialogue-summary pair is retained.

3 The MSAMSum Dataset

In this section, we introduce our MSAMSum dataset, including (1) Why we choose SAMSum dataset? (2) How we translate the original SAMSum dataset? (3) How we control the translation quality? and (4) Statistics for the newly created MSAMSum dataset. The whole dataset construction process is shown in Figure 2.

3.1 Dataset Selection

Current dialogue summarization datasets are mainly tailored for English (Gliwa et al., 2019; Chen et al., 2021a,b; Zhang et al., 2021), resulting in existing works not centring on other languages. In order to support our multi-lingual research, we follow Zhu et al. (2019), which uses state-of-the-art machine translation techniques to construct datasets in different languages.

Before launching the translation of the current dataset, we first need to choose a suitable dataset. After carefully comparing several datasets, we finally choose SAMSum (Gliwa et al., 2019) as our source English dataset according to the following two reasons: (1) it is a human-labeled large-scale dataset; (2) it covers a wide range of domains.

3.2 Machine Translation

For each dialogue-summary pair in the selected English SAMSum dataset (shown in Figure 2(a)), we translate the utterances and the summary to the target language (shown in Figure 2(b)) via high-performance machine translation service[3]. To make our work more representative and generalized, we choose five other official languages of the United Nations as our translation target languages[4]. Note that for each dialogue, we perform the translation at the utterance-level since machine translation can achieve good results with utterances of moderate length. After this process, we can get dialogue-summary pairs in Chinese (Zh), French (Fr), Arabic (Ar), Russian (Ru), Spanish(ES) and also original English (En).

3.3 Quality Controlling

To ensure the data quality, we further leverage two quality controlling methods. First, we employ *round-trip translation* strategy at both dialogue and summary level to filter out low-quality translations. Second, at the summary level, we use *textual entailment* strategy to verify factual consistency.

3.3.1 Round-trip Translation

Round-trip translation is the process of translating a text into another language (forward translation), then translating the result back into the original language (back translation), using MT service. Given the translated dialogue-summary pair in target language (shown in Figure 2(b)), we back-translate it into the original English version (shown in Figure 2(c)). Afterward, we follow Zhu et al. (2019) and calculate the ROUGE-1 score (Lin, 2004) between the original dialogue-summary pair and the back-translated dialogue-summary pair (shown in Figure 2(d)). In detail, we first calculate the ROUGE-1 score for the corresponding utterances and the sum-

[3]https://cloud.google.com/translate

[4]https://www.un.org/en/our-work/official-languages

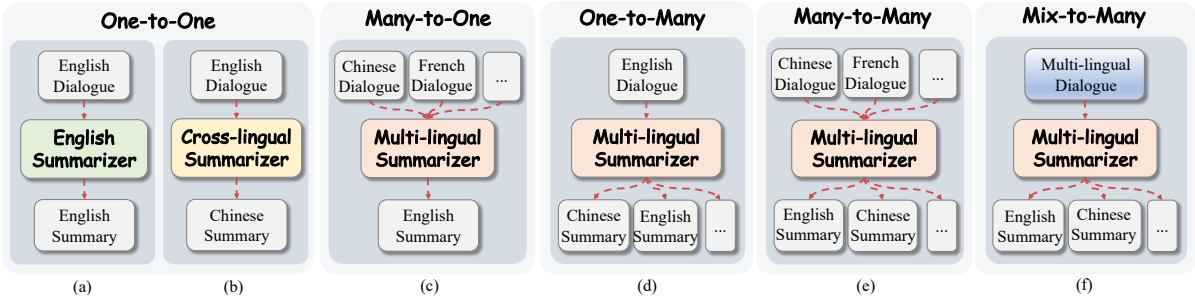


Figure 3: Illustration of different multi-lingual settings. We set up five settings in total, according to the number of input and output languages the model can handle. Concretely, the ONE-TO-ONE is the basic setting, the MANY-TO-ONE model encodes N languages and decodes to English, while the ONE-TO-MANY model encodes English and decodes into N languages, the MANY-TO-MANY model encodes and decodes N languages. Besides, we originally explore one new MIX-TO-MANY setting, where the model takes a mix-lingual dialogue (utterances in a dialogue belongs to different languages) as input and outputs summaries in different languages.

mary respectively, and then get the final ROUGE-1 score by averaging all scores. If the final ROUGE-1 score exceeds the pre-defined threshold, the translated dialogue-summary pair (shown in Figure 2(b)) is retained. Otherwise, the pair will be filtered[5].

3.3.2 Textual Entailment

Since the summary serves as the core part of dialogue summarization, it not only needs coarse-grained surface-level high quality but also fine-grained factual consistency (Huang et al., 2021). To this end, we adopt the textual entailment method to access whether the translated summary is consistent with the original summary. Specifically, we obtain the entailment score for the translated English summary and the original English summary via state-of-the-art entailment model[6], as shown in Figure 2(e). If the entailment score exceeds the pre-defined threshold, the translated dialogue-summary pair is retained. Otherwise, the pair will be filtered.

3.4 Datasets Alignment and Statistics

Following the above steps, we can get translated and pure datasets in different languages. Note that these datasets are of different sizes, which is caused by the quality controlling process. To unify our experiments, we get the intersection of these datasets in six languages, resulting in the final MSAMSum dataset (statistics in Table 1)[7].

[5]We show detailed round-trip translation ROUGE scores in the supplementary file.

[6]https://github.com/pytorch/fairseq/blob/main/examples/roberta/README.md

[7]We show the statistics for different parts before alignment in the supplementary file.

[8]https://forum.wordreference.com/threads/english-to-arabic-length-change.1495268/

		Train	Valid	Test
	#	5307	302	320
	Avg.Turns	11.01	10.48	11.15
En	Avg.Tokens	115.72	115.19	118.21
	Avg.Sum	22.18	22.33	22.06
Zh	Avg.Chars	242.08	237.39	246.95
	Avg.Sum	34.65	35.36	35.08
Fr	Avg.Tokens	99.33	99.01	102.5
	Avg.Sum	19.30	19.47	19.16
Ar	Avg.Tokens	57.17	55.85	56.63
	Avg.Sum	18.81	18.71	18.80
Ru	Avg.Tokens	89.00	88.53	91.11
	Avg.Sum	15.99	16.07	. 6.11
Es	Avg.Tokens	89.83	89.35	92.08
	Avg.Sum	18.67	18.60	18.68

Table 1: Statistics for MSAMSum dataset. "#" means the number of dialogue-summary pairs, "Avg.Turns", "Avg.Tokens", "Avg.Chars" and "Avg.Sum" mean the average number of turns of dialogues, tokens of dialogues, characters of dialogues and tokens of summaries respectively. Note that sentences in Arabic tend to be shorter than those in other languages[8].

4 Multi-lingual Settings

In this section, we introduce various multi-lingual dialogue summarization settings, including a newly proposed MIX-TO-MANY setting. All settings are depicted in Figure 3.

4.1 ONE-TO-ONE

The ONE-TO-ONE setting can be viewed as a specific type of multi-lingual setting, where the model can merely handle the input of one language and the output of one language. According to whether the

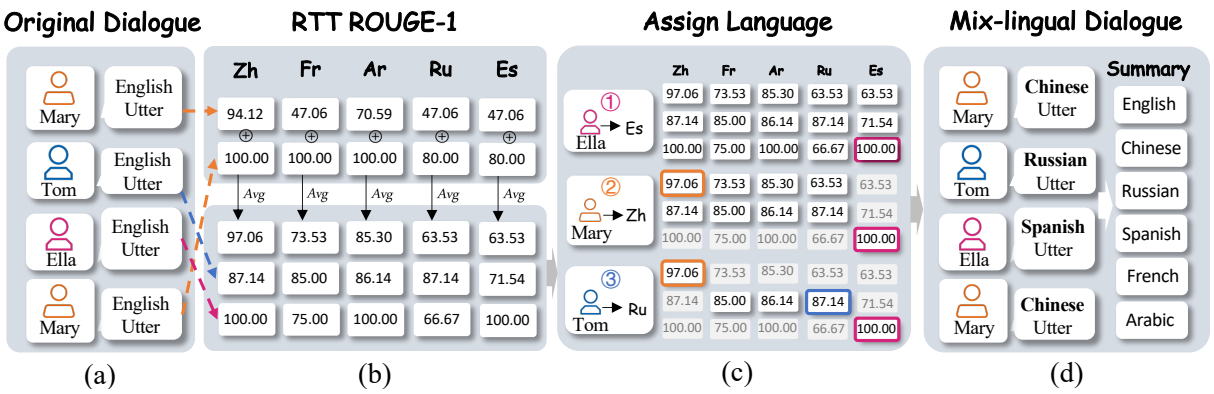


Figure 4: Illustration of the mix-lingual dialogue construction process. Given one English dialogue, we first group utterances for the same participant and get the averaged round-trip translation ROUGE-1 score for each language. Then, we adopt a greedy search strategy to assign each participant a language. Finally, we can get the mix-lingual dialogue associated with summaries in different languages.

input and output belong to the same language, this setting can be further divided into *Mono-lingual* setting (shown in Figure 3(a)) and *Cross-lingual* setting (shown in Figure 3(b)).

Experimental Setting: For *mono-lingual* experiments, we train six models based on {En→En}, {Zh→Zh}, {Fr→Fr}, {Ar→Ar}, {Ru→Ru} and {Es→Es} mono-lingual pairs respectively. For *cross-lingual* experiments, we train two models based on {En→Zh} and {Zh→En} cross-lingual pairs respectively. All eight models are tested in supervised manner.

4.2 MANY-TO-ONE and ONE-TO-MANY

MANY-TO-ONE models are able to process dialogues in various languages and output the summary in one language, as shown in Figure 3(c). On the contrary, ONE-TO-MANY models have the ability to produce summaries in various languages given a fixed language input, as shown in Figure 3(d). Both settings require models with multi-lingual capabilities.

Experimental Setting: For MANY-TO-ONE experiments, we train one model based on all {En→En, Zh→En, Fr→En, Ar→En, Ru→En, Es→En} pairs. For ONE-TO-MANY experiments, we train one model based on all {En→En, En→Zh, En→Fr, En→Ar, En→Ru, En→Es} pairs. These two models are tested in supervised manner.

4.3 MANY-TO-MANY

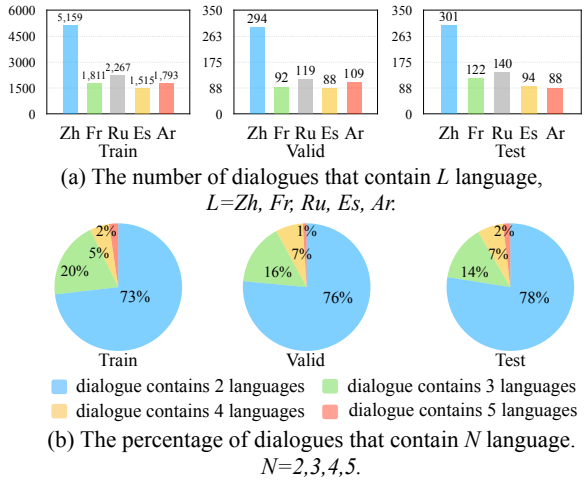
As shown in Figure 3(e), MANY-TO-MANY models can take dialogues in various languages as inputs and produce summaries in various languages. Thanks to the pre-trained multi-lingual language models (Liu et al., 2020; Tang et al., 2020), based on which, MANY-TO-MANY models can perform zero-shot summarization even though the input-output language pair is not seen during the training process.

Experimental Setting: For MANY-TO-MANY experiments, we train one model based on all {En→En, Zh→Zh, Fr→Fr, Ar→Ar, Ru→Ru, Es→Es} pairs and test it in both supervised and zero-shot manners.

4.4 MIX-TO-MANY

Nowadays, dialogue participants from different countries can use their mother tongue to communicate with each other based on instantaneous translation systems. To investigate the possibility of generating summaries directly from mix-lingual dialogues (utterances in different languages), we come up with an innovative new setting: MIX-TO-MANY, as shown in Figure 3(f).

To this end, we first simulate the real scenario and construct mix-lingual dialogue-summary pairs, the whole construction process is shown in Figure 4. Given each English dialogue in MSAMSum (shown in Figure 4(a)), we first group utterances by participants, which results in several groups for different participants (shown in Figure 4(b)). Then, for each group, we calculate the average round-trip translation ROUGE-1 score for each language (shown in Figure 4(c)). Afterward, we adopt a greedy search strategy to assign each participant a language (shown in Figure 4(d)). The goal of our strategy is twofold: choose as many languages as possible and as high-quality translations as possi-



(a) The number of dialogues that contain *L* language, *L=Zh, Fr, Ru, Es, Ar.*

(b) The percentage of dialogues that contain *N* language. *N=2,3,4,5.*

Figure 5: Statistics for mix-lingual dialogues. (a) We show the language distribution by calculating the number of dialogues containing one specific language; (b) We provide the distribution of the number of languages included in the dialogue.

ble. Finally, we can get the mix-lingual dialogue, in which utterances are in different languages. The number of mix-lingual dialogues is in line with MSAMSum. The statistics for mix-lingual dialogues are shown in Figure 5. Finally, we pair the mix-lingual dialogue with summaries in different languages (shown in Figure 4(e)).

Experimental Setting: For MIX-TO-MANY experiments, we train one model based on all {Mix→En, Mix→Zh, Mix→Fr, Mix→Ar, Mix→Ru, Mix→Es} pairs and test it in supervised manner.

5 Experiments

In this section, we first introduce our model mBART-50. After, we describe the evaluation metrics. Finally, we show the implementation details.

5.1 Backbone Model

We employ mBART-50 (Tang et al., 2020) as our multi-lingual summarizer, which is a Transformer-based model and pre-trained on a huge volume of multi-lingual data. It is derived from mBART (Liu et al., 2020) and extends the language processing capabilities from 25 languages to 50 languages in total. The architecture of mBART-50 is based on the BART (Lewis et al., 2020), which adopts position-wise feed-forward network, multi-head attention (Vaswani et al., 2017), residual connection (He et al., 2016) and layer normalization (Ba et al., 2016) modules to map the source dialogue into distributed representations and further generate the target summary.

To handle various input and output languages, mBART-50 needs to receive inputs with language identifiers (e.g., En, Zh) at both the encoder and the decoder side. According to the practical experience, we set both the source language identifier and target language identifier at the start of the source and target sequences respectively.

5.2 Evaluation Metrics

The most widely used metrics for summarization are ROUGE scores (Lin, 2004). However, the original ROUGE is specifically designed for English. To make this metric suitable for our experiments, we employ the multi-lingual ROUGE (Hasan et al., 2021) as our evaluation metrics, which takes segmentation and popular stemming algorithms for various languages into consideration[9].

5.3 Implementation Details

For MSAMSum construction, we set round-trip translation ROUGE-1 threshold to 80.00 and the textual entailment threshold to 0.9. For experiments, we use the standard mBART-50 implementation provided by Huggingface/transformers[10]. For fine-tuning process, the learning rate is set to 5e-06, the dropout rate is 0.1, the warmup is set to 2000 and the batch size is 4. In the test process, beam size is 5, the minimum decoded length is 10 and the maximum length is 150. All our experiments are conducted based on the Tesla-V100-32GB GPU.

6 Results

In this section, we describe experimental results and show our analyses for different settings.

6.1 ONE-TO-ONE Results

Table 2 shows the results for ONE-TO-ONE setting, including both the *mono-lingual* and the *cross-lingual* experiments. According to the 52.98 ROUGE-1 score achieved by fine-tuning BART-large on full English SAMSum dataset (Chen and Yang, 2020), we can see that our experiments achieve impressive results. For *mono-lingual* experiments, Ar→Ar results perform worse than others to some extent, we attribute this to the fact that the Arabic language processing capability of the

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

[10] https://huggingface.co/facebook/mbart-large-50-many-to-many-mmt

ONE-TO-ONE			
Src→Tgt	R-1	R-2	R-L
Mono-lingual			
En→En	49.16	24.18	40.15
Es→Es	43.95	20.01	35.87
Zh→Zh	40.11	16.93	33.48
Fr→Fr	41.77	19.20	34.47
Ru→Ru	37.95	15.74	31.76
Ar→Ar	28.66	6.61	23.07
Cross-lingual			
Zh→En	45.75	20.18	36.90
En→Zh	42.62	17.43	34.88

Table 2: Test set results on the different language pairs of MSAMSum dataset by fine-tuning mBART-50 under the ONE-TO-ONE setting, where "R" is short for "ROUGE".

MANY-TO-ONE			
Src→Tgt	R-1	R-2	R-L
En→En	48.18	22.43	38.63
Zh→En	45.01	17.76	35.49
Fr→En	44.22	18.49	35.30
Ar→En	31.09	08.00	24.18
Ru→En	44.20	17.53	35.06
Es→En	44.50	17.97	35.56

Table 3: Test set results on the different language pairs of MSAMSum dataset by fine-tuning mBART-50 under the MANY-TO-ONE setting.

pre-trained mBART-50 is relatively weak, which is in line with the size of original pre-training corpus (Lewis et al., 2020). For *cross-lingual* experiments, surprisingly, we find that En→Zh get better results compared with Zh→Zh, which may due to the model's strong English comprehension ability.

6.2 MANY-TO-ONE and ONE-TO-MANY Results

Table 3 and table 4 show results for MANY-TO-ONE and ONE-TO-MANY settings respectively. For both settings, we find that the results of the multi-lingual model varied less between pairs compared with ONE-TO-ONE models. For the MANY-TO-ONE model, the results of En→En and Zh→En are slightly worse than results of corresponding single ONE-TO-ONE models. This is because the MANY-TO-ONE model needs to handle multiple languages, which may cause the parameters interference problem (Lin et al., 2021), and is therefore inferior to a single expert model. In contrast, the

ONE-TO-MANY			
Src→Tgt	R-1	R-2	R-L
En→En	49.84	24.73	40.67
En→Es	47.27	21.82	37.87
En→Zh	43.86	18.25	35.56
En→Fr	44.33	19.58	35.20
En→Ru	41.26	15.76	33.00
En→Ar	39.71	14.96	32.82

Table 4: Test set results on the different language pairs of MSAMSum dataset by fine-tuning mBART-50 under the ONE-TO-MANY setting.

MANY-TO-MANY						
Src→Tgt	En	Zh	Fr	Ar	Ru	Es
En	**36.79**	*30.83*	*30.76*	*20.93*	*28.35*	*34.51*
Zh	*18.46*	**35.56**	*30.65*	*25.93*	*30.03*	*33.01*
Fr	*22.90*	*31.77*	**36.25**	*26.25*	*29.94*	*34.01*
Ar	*14.64*	*20.69*	*20.72*	**23.47**	*19.74*	*22.94*
Ru	*22.57*	*32.02*	*30.08*	*25.27*	**33.28**	*32.58*
Es	*27.74*	*32.09*	*31.97*	*25.75*	*30.11*	**37.21**

Table 5: Test set R-L results on the different language pairs of MSAMSum dataset by fine-tuning mBART-50 under the MANY-TO-MANY setting. Results in **bold** are achieved by supervised summarization. Results in *italics* are achieved by zero-shot summarization.

ONE-TO-MANY model improves the performance of both En→En and En→Zh results, which shows the ONE-TO-MANY training setting enhances the model's English understanding ability. Additionally, both Ar→En and En→Ar get relatively lower results, which coincide with the findings in ONE-TO-ONE experiments.

6.3 MANY-TO-MANY Results

Table 5 shows ROUGE-L results for the MANY-TO-MANY setting[11]. We test each language pair in the cartesian product of six languages, which results in two types of manners: supervised and zero-shot summarization. For the supervised manner (results in **bold**), almost all results show the best performance. For the zero-shot manner (results in *italics*), we find that despite the model is fine-tuned based on mono-lingual dialogue-summary pairs, it still has the strong ability to perform summarization across different languages. In line with previous experiments, we find the MANY-TO-MANY model that balances across various languages inevitably loses some performances compared with the ONE-TO-ONE model. Nonetheless, the MANY-

[11] We show all ROUGE-1, ROUGE-2 and ROUGE-L scores in the supplementary file.

MIX-TO-MANY			
Src→Tgt	R-1	R-2	R-L
Mix→En	44.68	17.78	35.17
Mix→Es	43.51	18.08	34.75
Mix→Zh	40.76	15.76	33.14
Mix→Fr	41.50	17.04	32.76
Mix→Ru	38.26	13.38	30.75
Mix→Ar	36.06	12.09	29.60

Table 6: Test set results on the different language pairs of MSAMSum dataset by fine-tuning mBART-50 under the MIX-TO-MANY setting.

TO-MANY model, which greatly reduces the deployment cost while preserving the performance, is an important research direction in the future.

6.4 MIX-TO-MANY Results

Table 6 shows the results for the MIX-TO-MANY setting. As the first step towards this direction, we find that current multi-lingual pre-trained models can obtain encouraging results. The Mix→Es, Mix→Zh, Mix→Fr and Mix→Ru models achieve comparable results with respect to the corresponding ONE-TO-ONE model. These results verify that despite the multi-lingual model only deals with one language at a time in the pre-training progress, after fine-tuning, it can handle mix-lingual inputs concurrently. Surprisingly, the Mix→Ar results even surpass the performance of singe Ar→Ar model. We think this is due to the mix-lingual dialogue essentially acts as an utterance-level code-switching data, which helps the representation space of the low-resource language align with other languages. This also inspire us that it would be better to generate the low-resource language summary directly from the mix-lingual dialogue.

6.5 Case Study

Figure 6 shows summaries in different languages generated by the ONE-TO-MANY model for an example English dialogue. We can see that all the generated summaries achieve good ROUGE performance, with English being the highest. We find that the multi-lingual model can generate fluent summaries while preserving the important information of the dialogue. Besides, the model also has the ability to accurately express participants information (e.g., Elliot, Jordan) and keep entities' factual consistency (e.g., 8 pm) across different languages.

English Dialogue	
Elliot : I can't talk rn , I'm rly busy. Elliot : Can I call u back in about 2 hours? Jordan : Not really , I'm going to a funeral. Jordan : I'll call you tonight , ok? Elliot : Sure Elliot : Whose funeral is it? Jordan : My colleague's , Brad. Jordan : I told you about him , he had a liver cancer. Elliot : I'm so sorry man , I hope u're ok. Elliot : I'll call u at 8 pm.	
Generated Summaries (*One-to-many*)	
English	Elliot can't talk because he's busy. Jordan is going to a funeral for his colleague, Brad, who had a liver cancer. Elliot will call him at 8 pm. [71.19-42.11-50.85]
Chinese	乔丹要去参加他的同事布拉德的葬礼。他得了肝癌。埃利奥特将在晚上8点给乔丹打电话。[66.67-40.00-35.09]
Russian	Джордан собирается на похороны своего коллегы Брэда, у него рак печени.Элиот позвонит Джордана в 20: 00. [58.38-30.00-38.10]
French	Elliot ne peut pas parler parce qu'il est occupé.Jordan va au funeral de son collègue, Brad, qui a un cancer du foie.Il appellera Elliot à 20 h. [68.97-42.86-55.17]
Arabic	جوردن هو الذهاب إلى جنازة زميلها براد لديه سرطان الكبد.إيليت سوف ندعو له في الساعة الثامنة مساء. [57.78-27.91-31.11]
Spanish	Elliot no puede hablar porque está ocupado.Jordan va a un funeral de su colega, Brad, que tuvo un cáncer de hepática.Eliot llamará a Jordan a las 8 p.m. [60.71-29.63-39.29]

Figure 6: Example English dialogue in the MSAMSum dataset and summaries in different languages generated by the ONE-TO-MANY model. The scores in square brackets are R-1, R-2 and R-L respectively.

7 Conclusion and Future Work

In this paper, we innovatively explore the multi-lingual dialogue summarization task. To this end, we carefully create MSAMSum as our testbed, which covers dialogue-summary pairs in six languages, including English, Chinese, Russian, French, Arabic and Spanish. Furthermore, we systematically set up five multi-lingual settings to benchmark extensive experiments. Our results indicate that various models can achieve impressive performance based on pre-trained models. Besides, the newly proposed MIX-TO-MANY setting also shows its effectiveness in low-resource scenarios.

In the future, we think several concerns need to be addressed for this task. Firstly, multi-lingual models tend to underperform mono-lingual models; Secondly, low-resource languages tend to perform poorly; Thirdly, the difficulty of aligning fine-grained information in different languages. Future works should pay particular attention to these concerns to facilitate this multi-lingual dialogue summarization research direction.

References

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *In arXiv*.

Yue Cao, Xiaojun Wan, Jin-ge Yao, and Dian Yu. 2020. Multisumm: Towards a unified model for multi-lingual abstractive summarization. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020*. AAAI Press.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, et al. 2005. The ami meeting corpus: A pre-announcement. In *International workshop on machine learning for multimodal interaction*. Springer.

Jiaao Chen and Diyi Yang. 2020. Multi-view sequence-to-sequence models with conversational structure for abstractive dialogue summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online. Association for Computational Linguistics.

Mingda Chen, Zewei Chu, Sam Wiseman, and Kevin Gimpel. 2021a. Summscreen: A dataset for abstractive screenplay summarization. *arXiv preprint arXiv:2104.07091*.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021b. Dialsumm: A real-life scenario dialogue summarization dataset. *arXiv preprint arXiv:2105.06762*.

Alexander Fabbri, Faiaz Rahman, Imad Rizvi, Borui Wang, Haoran Li, Yashar Mehdad, and Dragomir Radev. 2021. ConvoSumm: Conversation summarization benchmark and improved abstractive summarization with argument mining. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6866–6880, Online. Association for Computational Linguistics.

Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2021a. A survey on dialogue summarization: Recent advances and new frontiers. *ArXiv*, abs/2107.03175.

Xiachong Feng, Xiaocheng Feng, Bing Qin, and Xinwei Geng. 2021b. Dialogue discourse-aware graph model and data augmentation for meeting summarization. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, Hong Kong, China. 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*, Online.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society.

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*.

Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, et al. 2003. The icsi meeting corpus. In *ICASSP*. IEEE.

Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. 2020. Wikilingua: A new benchmark dataset for multilingual abstractive summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*.

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*, Online. Association for Computational Linguistics.

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

Zehui Lin, Liwei Wu, Mingxuan Wang, and Lei Li. 2021. Learning language specific sub-network for multilingual machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online. Association for Computational Linguistics.

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.

Valentin Malykh, Konstantin Chernis, Ekaterina Artemova, and Irina Piontkovskaya. 2020. SumTitles: a summarization dataset with low extractiveness. In *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain (Online).

Laiba Mehnaz, Debanjan Mahata, Rakesh Gosangi, Uma Sushmitha Gunturi, Riya Jain, Gauri Gupta, Amardeep Kumar, Isabelle Lee, Anish Acharya, and Rajiv Ratn Shah. 2021. Gupshup: An annotated corpus for abstractive summarization of open-domain code-switched conversations. *arXiv preprint arXiv:2104.08578*.

Jessica Ouyang, Boya Song, and Kathy McKeown. 2019. A robust abstractive system for cross-lingual summarization. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota. Association for Computational Linguistics.

Revanth Rameshkumar and Peter Bailey. 2020. Storytelling with dialogue: A Critical Role Dungeons and Dragons Dataset. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

Don Tuggener, Margot Mieskes, Jan Deriu, and Mark Cieliebak. 2021. Are we summarizing the right way? a survey of dialogue summarization data sets. In *Proceedings of the Third Workshop on New Frontiers in Summarization*, Online and in Dominican Republic. Association for Computational Linguistics.

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*, Online and Punta Cana, Dominican Republic.

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*.

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*, Uppsala, Sweden. Association for Computational Linguistics.

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*, Online. Association for Computational Linguistics.

Ruochen Xu, Chenguang Zhu, Yu Shi, Michael Zeng, and Xuedong Huang. 2020. Mixed-lingual pretraining 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*, Suzhou, China. Association for Computational Linguistics.

Shiyue Zhang, Asli Celikyilmaz, Jianfeng Gao, and Mohit Bansal. 2021. EmailSum: Abstractive email thread summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online. Association for Computational Linguistics.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021. QMSum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online. Association for Computational Linguistics.

Chenguang Zhu, Yang Liu, Jie Mei, and Michael Zeng. 2021. MediaSum: A large-scale media interview dataset for dialogue summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online. Association for Computational Linguistics.

Chenguang Zhu, Ruochen Xu, Michael Zeng, and Xuedong Huang. 2020a. A hierarchical network for abstractive meeting summarization with cross-domain pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, Online. Association for Computational Linguistics.

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)*, Hong Kong, China. Association for Computational Linguistics.

Junnan Zhu, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2020b. 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*, Online. Association for Computational Linguistics.

MANY-TO-MANY						
Src→Tgt	En	Zh	Fr	Ar	Ru	Es
En	**48.00/22.29/36.79**	*37.51/13.82/30.83*	*38.81/14.56/30.76*	*24.48/8.16/20.93*	*34.50/11.49/28.35*	*42.86/17.38/34.51*
Zh	*24.24/8.37/18.46*	**43.75/19.14/35.56**	*39.80/13.96/30.65*	*32.28/10.10/25.93*	*37.82/12.87/30.03*	*41.97/16.08/33.01*
Fr	*29.71/08.69/22.90*	*39.53/13.73/31.77*	**45.26/21.60/36.25**	*31.92/10.34/26.25*	*37.11/12.17/29.94*	*42.59/16.59/34.01*
Ar	*18.75/3.74/14.64*	*25.27/6.36/20.69*	*26.46/6.30/20.72*	**29.15/7.76/23.47**	*24.48/5.04/19.74*	*29.24/6.89/22.94*
Ru	*30.88/9.99/22.57*	*39.80/14.46/32.02*	*38.29/13.84/30.08*	*30.72/9.49/25.27*	**41.50/15.95/33.28**	*41.53/15.18/32.58*
Es	*37.18/12.14/27.74*	*39.79/15.05/32.09*	*41.04/15.91/31.97*	*31.41/10.18/25.75*	*37.34/12.02/30.11*	**46.40/21.53/37.21**

Table 7: Test set ROUGE-1/ROUGE-2/ROUGE-L results on the different language pairs of MSAMSum dataset by fine-tuning mBART-50 under the MANY-TO-MANY setting. Results in **bold** are achieved by supervised summarization. Results in *italics* are achieved by zero-shot summarization.

A Ethical Considerations

As we propose a new multi-lingual dialogue summarization dataset and conduct experiments based on large pre-trained language models, we make several clarifications to address potential concerns:

- **Dataset:** Since our MSAMSum is derived from the SAMSum (Gliwa et al., 2019), which is a well-constructed and human-labelled dataset. Therefore, our dataset inherits the contents of SAMSum and does not contain toxic information.
- **Model:** The experiments described in this paper are based on the mBART-50-large (Tang et al., 2020) and make use of V100 GPUs. Despite we run dozens of experiments, our results could help reduce parameter searches for future works. We also consider to alleviate such resource-hungry challenge by exploring light-weight distilled models.

B Detailed MANY-TO-MANY Results

Table 7 shows detailed ROUGE-1, ROUGE-2 and ROUGE-L results for MANY-TO-MANY experiments in both supervised and zero-shot manners, as a supplement to Table 5.

C Round-trip Translation ROUGE Scores

Table 8 shows the average ROUGE scores between the English data in SAMSum (Gliwa et al., 2019) and the round-trip translated English data. These results indicate the overall translation quality.

D The Changing of Data Size

Table 9 shows how the data size changes. After quality controlling process, we can get different data size for different languages (before alignment). After taking the intersection of different languages, we get our final MSAMSum (after alignment).

		R-1	R-2	R-L
Train	Zh	84.57	60.87	86.77
	Ru	75.97	47.70	78.91
	Es	75.05	46.43	78.19
	Ar	76.09	48.13	79.02
	Fr	75.53	47.02	78.68
Valid	Zh	84.47	60.80	86.69
	Ru	75.57	46.81	78.56
	Es	74.85	46.19	77.99
	Ar	75.97	48.09	78.93
	Fr	75.24	46.74	78.40
Test	Zh	84.11	59.91	86.32
	Ru	75.74	47.18	78.67
	Es	74.68	45.63	77.84
	Ar	75.56	47.24	78.48
	Fr	75.15	46.39	78.33

Table 8: The average ROUGE scores between each original English data in the SAMSum (Gliwa et al., 2019) and corresponding round-trip translated English data for five languages.

	Train	Valid	Test
Original			
SAMSum	14732	818	819
Before alignment			
Zh	11738	658	660
Ru	6089	329	354
Es	6697	369	370
Ar	6341	340	337
Fr	7523	426	417
After alignment			
Final	5307	302	320

Table 9: The size of datasets at different stages.