CLIDSUM: A Benchmark Dataset for Cross-Lingual Dialogue Summarization

Jiaan Wang[1,2,*], Fandong Meng[1,†], Ziyao Lu[1], Duo Zheng[4]
Zhixu Li[3,†], Jianfeng Qu[2] and Jie Zhou[1]

[1]Pattern Recognition Center, WeChat AI, Tencent Inc, China

[2]School of Computer Science and Technology, Soochow University, Suzhou, China

[3]Shanghai Key Laboratory of Data Science, School of Computer Science, Fudan University, Shanghai, China

[4]Beijing University of Posts and Telecommunications, Beijing, China

jawang1@stu.suda.edu.cn, fandongmeng@tencent.com, zhixuli@fudan.edu.cn

Abstract

We present CLIDSUM, a benchmark dataset towards building cross-lingual summarization systems on dialogue documents. It consists of 67k+ dialogue documents and 112k+ annotated summaries in different target languages. Based on the proposed CLIDSUM, we introduce two benchmark settings for supervised and semi-supervised scenarios, respectively. We then build various baseline systems in different paradigms (pipeline and end-to-end) and conduct extensive experiments on CLIDSUM to provide deeper analyses. Furthermore, we propose mDIALBART which extends mBART via further pre-training, where the multiple objectives help the pre-trained model capture the structural characteristics as well as key content in dialogues and the transformation from source to the target language. Experimental results show the superiority of mDIALBART, as an end-to-end model, outperforms strong pipeline models on CLIDSUM. Finally, we discuss specific challenges that current approaches faced with this task and give multiple promising directions for future research. We have released the dataset and code at https://github.com/krystalan/ClidSum.

1 Introduction

Cross-Lingual Summarization (XLS) aims at generating a summary in the target language from a given document in a different source language. Existing XLS datasets mainly focus on single-participant documents, such as news reports (Zhu et al., 2019; Nguyen and Daumé III, 2019; Bai et al., 2021), how-to guides (Ladhak et al., 2020) and encyclopedia articles (Perez-Beltrachini and Lapata, 2021). Many efforts have been devoted to improving XLS for single-participant documents and achieved great success (Yao et al., 2015; Shen

Figure 1: An example of cross-lingual dialogue summarization. Blue and orange sentences are the German and Chinese summaries of the given English dialogue, respectively.

et al., 2018; Ouyang et al., 2019; Cao et al., 2020; Zhu et al., 2020; Maurya et al., 2021; Chi et al., 2021; Liang et al., 2022; Wang et al., 2022).

Different from single-participant documents, dialogue is a discourse produced by more than one person (Haviland, 1990). The multi-participant dialogue documents, which record the communications between human and human/machine, have attracted wide research attention due to their key role in daily interpersonal interaction (Zhang et al., 2020c). Meanwhile, the globalization progress has prompted conversations among interlocutors of different native languages and brought many scenarios, e.g., international academic conferences and business meetings. Thus, it is valuable to provide speakers with the summary in their familiar language to help them efficiently grasp the gist of a foreign language dialogue. Nevertheless, dialogue-oriented XLS is still under-explored due to the lack of corresponding datasets.

To this end, we introduce Cross-Lingual Dialogue Summarization (XLDS) task that aims to summarize a dialogue in the source language into a different language. To promote the XLDS research, we construct CLIDSUM (**C**ross-**LI**ngual **D**ialogue **SUM**marization), the first large-scale XLDS benchmark dataset with three features: (1) The proposed CLIDSUM is based on two existing

---

monolingual dialogue summarization datasets, i.e., SAMSum (Gliwa et al., 2019) and MediaSum (Zhu et al., 2021). We choose these two datasets under the consideration of their quality and diversity. (2) To make these datasets suitable for XLDS, we employ professional translators to translate original English summaries of SAMSum and MediaSum to German and Chinese. Eventually, the translated corpora constitute CLIDSUM, which totally contains ~56.4k En⇒De and ~56.4k En⇒Zh XLDS samples[1]. (3) Besides the supervised benchmark setting that has been discussed by most previous XLS work, we argue that it is necessary to utilize large-scale monolingual dialogue summarization pairs for XLDS due to the dearth of cross-lingual samples. Thus, we design a semi-supervised setting, where a large number of monolingual pairs together with a relatively small scale of cross-lingual pairs are used to build XLDS systems.

Based on CLIDSUM, we build and evaluate various baseline systems, including *summarize-then-translate*, *translate-then-summarize* and *end-to-end* paradigms: (1) For *summarize-then-translate* baselines, we use a monolingual dialogue summarizer to generate summaries and further translate them to the target language through a machine translation (MT) model/service. (2) For *translate-then-summarize* baselines, we utilize a MT model/service to translate the dialogue documents from the source language to the target language, and then obtain summaries via a monolingual dialogue summarizer. (3) As for *end-to-end* paradigm, we adopt mBART-50 (Tang et al., 2021), a multi-lingual sequence-to-sequence (Seq2Seq) model which has been pre-trained on a large-scale corpora across 50 languages using denoising objectives. Though such a general model achieves surprising results in many downstream tasks, its performance will naturally degrade when there is a relatively large gap between the pre-training and fune-tuning stages (Lai et al., 2021; Zhong et al., 2022). Therefore, to narrow the gap and fully use large-scale monolingual dialogue summarization pairs provided in semi-supervised setting, we propose mDIALBART that extends mBART-50 through the second stage of pre-training with four objectives: action infilling, utterance permutation, monolingual dialogue summarization and machine translation. Specifically, action infilling and utterance permutation encourage the model to

effectively capture the structural characteristics in dialogues. Monolingual dialogue summarization gives the model the ability to summarize dialogues. Machine translation enables the model to learn the transformation from the source language to the target language. The pre-training corpora are provided in the semi-supervised setting, which means we do not use additional data other than CLIDSUM. The experimental results show that mDIALBART outperforms strong pipeline baselines on CLIDSUM. We also conduct human studies to compare generated summaries from different methods and discuss specific challenges that current approaches faced with XLDS. We hope that our work could promote the development of XLDS.

Our main contributions are concluded as follows:

- We introduce XLDS task and present CLIDSUM, the first large-scale XLDS benchmark dataset collecting about 56.4k En⇒De and 56.4k En⇒Zh samples via crowd-sourcing. Furthermore, we design supervised and semi-supervised benchmark settings on CLIDSUM.
- We elaborately build and evaluate various baselines of different paradigms (i.e., translate-then-summarize, summarize-then-translate and end-to-end) to provide deeper analyses.
- We propose the mDIALBART model that extends mBART-50 via the second pre-training stage. Experimental results show that mDIALBART outperforms existing baselines.
- To provide a deeper understanding of the XLDS task and the CLIDSUM dataset, we analyze the results of different methods and conclude the critical difficulties of XLDS on CLIDSUM.

2 Related Work

Cross-Lingual Summarization. We divide existing XLS datasets into *synthetic datasets* and *multilingual website datasets* according to the construction methods. (1) *Synthetic datasets* are constructed through translating the summaries of existing text summarization datasets to a target language, such as En2ZhSum, Zh2EnSum (Zhu et al., 2019) and En2DeSum (Bai et al., 2021). These datasets are further equipped with the round-trip translation strategy (Zhu et al., 2019; Zhang et al., 2021; Lai et al., 2022) to filter out low-quality samples. (2) *Multi-lingual website datasets* are collected from websites which provide multi-lingual versions for their articles. For instance, WikiLingual (Ladhak et al., 2020) collects articles from the WikiHow

---

[1]En: English; De: German; Zh: Chinese.

website, where many English articles are translated to non-English versions by human writers. Each article also links to parallel articles in other languages, if available. Thus, it is handy to collect different language versions for one article. Next, the summary of each language-specific article is extracted through a heuristic strategy. In this way, the article in one language and its summary in a different language could constitute an XLS sample. In the similar way, Global Voices (Nguyen and Daumé III, 2019) and XWikis (Perez-Beltrachini and Lapata, 2021) collect multi-lingual articles from Global Voices and Wikipedia websites, respectively. Early XLS methods (Wan et al., 2010; Zhang et al., 2016; Ayana et al., 2018; Ouyang et al., 2019) are based on pipeline paradigms due to the scarcity of parallel corpus. Recently, Zhu et al. (2019) propose the first large-scale XLS dataset and further explore the multi-task learning on end-to-end (e2e) XLS systems which achieve great improvement over pipeline methods. Subsequently, many efforts have contributed to the e2e XLS systems. Among them, Zhu et al. (2020) exploit the translation patterns in XLS. Cao et al. (2020) propose a framework that jointly learns to align and summarize for XLS. Xu et al. (2020) explore pre-training strategy on XLS. Liang et al. (2022) adopt conditional variational auto-encoder (CVAE) (Sohn et al., 2015) to deal with XLS. Different from existing XLS work, we shift research attention from single-participant documents to multi-participant dialogues.

Dialogue Summarization. Dialogue summarization aims to summarize the dialogue document into a shorter passage. AMI (Carletta et al., 2005) and ICSI (Janin et al., 2003) are two early meeting corpora that both contain hundreds of samples. Recently, the SAMSum corpus (Gliwa et al., 2019) is introduced, which contains over 16k chat conversations with human-labeled summaries. MediaSum (Zhu et al., 2021) collects about 464k interview transcripts and uses the overview descriptions as summaries. GupShup (Mehnaz et al., 2021) develops the first code-switched dialogue summarization dataset whose dialogues are in Hindi-English. Based on these datasets, a lot of work (Chen and Yang, 2020; Wu et al., 2021; Xiachong et al., 2021; Chen and Yang, 2021; Feng et al., 2021) models the conversation characteristics and achieves great performance. All these efforts are given to monolingual or code-switched scenarios while we focus

on cross-lingual scenario in this paper.

3   CLIDSUM

In this section, we first discuss how we choose existing monolingual dialogue summarization datasets for the construction of CLIDSUM (§ 3.1). Then, we introduce the details of the annotation process that translates original summaries to target languages (§ 3.2). We next give statistic analysis of CLIDSUM (§ 3.3). Finally, we formulate XLDS task (§ 3.4) and give the details of benchmark settings (§ 3.5).

3.1   Data Selection

As discussed in Section 2, there are two types of XLS datasets: *synthetic datasets* and *multi-lingual website datasets*. To our knowledge, there is no public website that provides multi-lingual dialogue-summary pairs. Therefore, we decide to construct XLDS dataset through translating original summaries of existing dialogue summarization datasets. After carefully comparing existing datasets, we finally choose SAMSum (Gliwa et al., 2019) and MediaSum (Zhu et al., 2021) datasets due to the following reasons: (1) these datasets are of high quality, both of which consist of real-world or human-labeled monolingual dialogue-summary pairs; (2) these datasets collect dialogues in a wide range of domains in daily life.

3.2   Data Annotation

Since the size of MediaSum (~464k) is much larger than any other dialogue summarization datasets (typically less than 30k), it is costly to translate all summaries from MediaSum to target languages via crowd-sourcing. Hence, we randomly select 20k samples from its training set, which together with all samples of validation set and testing set form MediaSum40k (totally 40k samples) subset. The remaining monolingual data is denoted as MediaSum424k. We then translate original English summaries from SAMSum and MediaSum40k to German and Chinese through data annotation.

There are 55 professional translators, 3 data reviewers and 1 data expert participating in the annotation process. All En⇒Zh translators have passed the TEM-8 while all En⇒De translators have passed both the TEM-8 and the PGH[2]. The

---

[2] *Test for English Majors-Band 8* (TEM-8) and *Prüfung für das Germanistik Hauptstudium* (PGH) are two language qualifications in China, which measure the overall English and German proficiency of undergraduates majoring in English and German, respectively.

	Dataset	Domain	Trans.	Src Lang.	Tgt Lang.	Documents	Doc. Length	Src Summ. Length	Tgt Summ. Length
	En2ZhSum	News	Auto	En	Zh	370,687	755.0	55.2	96.0
	Zh2EnSum	News	Auto	Zh	En	1,699,713	103.7	17.9	13.7
	En2DeSum	News	Auto	En	De	437,797	31.0	8.5	7.5
CLIDSUM	XSAMSum	Chat	Manual	En	De/Zh	16,369	83.9	20.3	19.9/33.0
	XMediaSum40k	Interview	Manual	En	De/Zh	40,000	1555.4	14.4	14.8/30.0
	MediaSum424k	Interview	-	En	-	423,596	1553.6	14.4	-

Table 1: Statistics of CLIDSUM and previous synthetic XLS datasets. *Trans.* indicates the translation method (automatic or manual) to construct dataset. *Src Lang.* and *Tgt Lang.* denote the source and target languages (En: English, Zh: Chinese or De: German) for each dataset. *Documents* represents the size of each dataset. *Doc. Length*, *Src Summ Length* and *Tgt Summ Length* show the average length of documents, source summaries and target summaries (word-level for English and German while character-level for Chinese) for each dataset, repetively.

Dataset	Split	Dial.	Turns	Speakers
(X)SAMSum	Train	14,732	11.2	2.4
	Dev.	818	10.8	2.4
	Test	819	11.2	2.4
(X)MediaSum40k	Train	20,000	30.4	9.3
	Dev.	10,000	29.3	9.2
	Test	10,000	30.6	9.3

Table 2: Statistics of the dialogue characteristics in (X)SAMSum and (X)MediaSum40k. *Dial.* indicates the number of dialogue documents for each subset of these datasets. *Turns* and *Speakers* represent the average number of utterances and speakers in dialogues.

data reviewers and the expert are proficient in English, German and Chinese, and have rich experiences in checking translation quality. During the annotation process, 20% of the summaries translated by each translator are checked by a reviewer. If the accuracy[3] is lower than 95%, the translator needs to modify all his/her translated summaries under the guidance of the reviewer. To guarantee the overall quality, after the above process for one translation direction of SAMSum or MediaSum40k, 2% of the translated summaries are randomly sampled and checked by the data expert. If the accuracy is lower than 95%, the corresponding translators and reviewers need to revise their translated summaries again. This quality control loop is executed 7 [4], 1, 1 and 2 times for SAMSum(En⇒Zh), SAMSum(En⇒De), MediaSum40k(En⇒Zh) and MediaSum40k(En⇒De), respectively.

In the end, we collect 56,369 Chinese and 56,369 German summaries. We denote the summary-translated SAMSum and MediaSum40k as XSAMSum and XMediaSum40k, respectively, which also inherit the data splitting of original datasets. The

XSAMSum, XMediaSum40k and MediaSum424k together constitute CLIDSUM benchmark dataset.

3.3 Data Analysis

We compare CLIDSUM with previous *synthetic datasets* in Table 1. CLIDSUM is the first large-scale XLS dataset towards dialogues, and it is also the first manually translated XLS dataset. While one may argue that the scale of previous synthetic datasets is much larger than XSAMSum and XMediaSum40k, it is worth noting that 1) the scale of all dialogue summarization datasets is less than 30k except MediaSum, and 2) the quality of XSAMSum and XMediaSum40k is much higher than those automatically constructed ones. In addition, as shown in Table 2, the dialogues from (X)MediaSum40k contain more turns (~30 vs. ~11) and more speakers (~9.3 vs. ~2.4) than (X)SAMSum, leading to more challenges for XLDS research.

3.4 Task Overview

Given a dialogue $\mathcal{D} = \{u_1, u_2, ..., u_{|\mathcal{D}|}\}$ in the source language, where u_i denotes the i-th utterance in $\mathcal{D}$, the XLDS task aims to generate corresponding summary $Y_t = (y_1, y_2, ..., y_{|Y_t|})$ in a different target language, where y_j denotes the j-th token in Y_t.

3.5 Benchmark Settings

We design two benchmark settings for supervised and semi-supervised scenarios, respectively. In the *Supervised Setting*, an XLDS system is established on the training set of XSAMSum or XMediaSum40k and evaluated on the corresponding test set. In the *Semi-Supervised Setting*, the training set of XMediaSum40k and the whole MediaSum424k are used to build XLDS systems which are evaluated on the test set of XMediaSum40k.

---

[3] A translated summary will be regarded as incorrect once it has an error judged by humans.

[4] The accuracy achieves 90% three times, 93% once, 94% twice and finally reaches 98%.

4 Baselines

4.1 Pipeline Method

The main idea of the pipeline method is decomposing XLDS task into dialogue summarization and machine translation sub-tasks. It can be further divided into *summarize-then-translate* and *translate-then-summarize* paradigms.

Summarize-then-translate. In this paradigm, a monolingual dialogue summarizer is used to generate summary in the same language with the given dialogue, and then obtain target language summary through machine translation. The following models are adopted as the dialogue summarizers:

- PEGASUS (Zhang et al., 2020a) is a pre-trained abstractive summarization model.
- T5 (Raffel et al., 2020) is a jointly pre-trained Seq2Seq model for many downstream NLP tasks.
- BART (Lewis et al., 2020) is another pre-trained Seq2Seq model using denoising objectives during pre-training stage.
- mBART-50 (Tang et al., 2021) is a multi-lingual version of BART, which could be also employed on monolingual tasks, though it does not take advantage of its multi-lingual ability.
- MV-BART (Chen and Yang, 2020) is a BART-based multi-view dialogue summarization model which utilizes conversation information from different views, e.g., topic view and stage view.
- BART($\mathcal{D}_{\text{ALL}}$) (Feng et al., 2021) encodes additional dialogue characteristics to BART. The characteristics are extracted by a DialoGPT (Zhang et al., 2020c) annotator.

To translate the generated summaries from the source to the target language, previous XLS work (Ladhak et al., 2020; Perez-Beltrachini and Lapata, 2021) only adopts sophisticated translation services, e.g., Google Translation and AWS Translation. Though great performances, the black-box APIs provided by translation services are constantly updating, which leads to poor reproducibility. Thus, we decide to adopt both translation services and models. Specifically, we consider three translation methods: (1) Google Translation[5]; (2) OPUS-MT (Tiedemann and Thottingal, 2020) releases lots of transformer-based MT models which are trained on OPUS corpus[6] with different translation directions; (3) Trans-WMT20: we train two transformer-based Seq2Seq models on WMT20 parallel corpus[7]

---

[5] https://cloud.google.com/translate
[6] https://opus.nlpl.eu/
[7] https://www.statmt.org/wmt20/

from scratch, including En⇒De and En⇒Zh.

Translate-then-summarize. This paradigm first translates English dialogues from CLIDSUM to German and Chinese through the same translation methods as the *summarize-then-translate*, and then generate summaries via mBART-50.

4.2 End-to-End Method

The end-to-end method needs simultaneously learn both dialogue summarization and machine translation. We finetune mBART-50 with input dialogues from the source language, and summaries from the target language. Note that mBART-50 is used in both pipeline and end-to-end paradigms, where the languages of input and output sequences are the same when used in the pipeline paradigm while different in the end-to-end paradigm. To indicate the input and output languages, mBART-50 appends language identifiers (e.g., `En`, `De` and `Zh`) at both the encoder and the decoder sides.

5 mDIALBART

As suggested by previous work (Zhu et al., 2020; Ladhak et al., 2020), building an end-to-end model is preferable to a pipeline one due to: 1) the pipeline models suffer from error propagation; 2) the translation systems in pipeline paradigm require either a large parallel corpus to train MT models or the monetary cost of paid MT services; 3) the pipeline models have a recurring latency during inference (Ladhak et al., 2020). Thus, it is valuable and urgent to explore end-to-end models on XLDS. To this end, we propose mDIALBART which extends mBART-50 via a second stage of pre-training, where the objectives help the pre-trained model better adapt to the XLDS task (cf., Figure 2).

5.1 Pre-Training Tasks and Corpora

Action Infilling (AcI). Action triples (i.e., *"who-doing-what"*) help the pre-trained model explicitly be aware of the actions within utterances for generating more factual summaries (Chen and Yang, 2021). Following Chen and Yang (2021), we extract action triples through `OpenIE` systems (Angeli et al., 2015). Then we randomly sample text spans from action triples and replace each span with a `[MASK]` token. About 15% of tokens in an utterance will be masked. The action infilling task requires the pre-trained model to reconstruct original dialogues based on the masked dialogues.
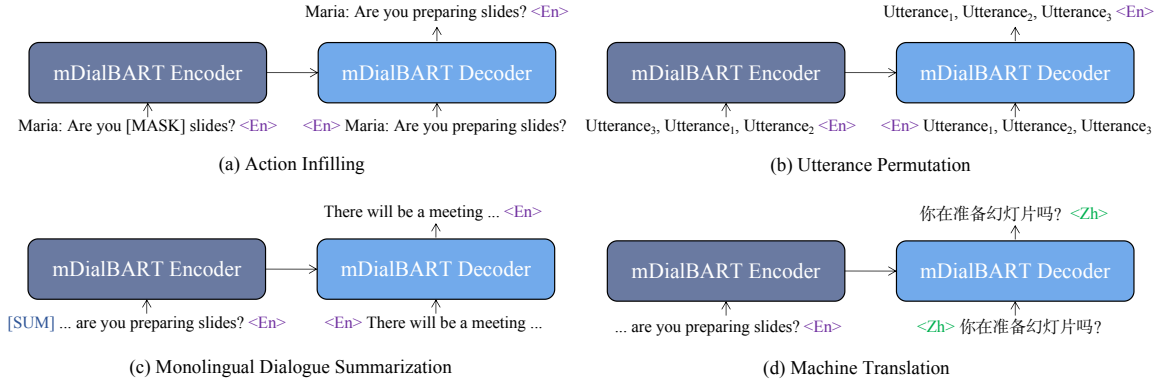
---

translation-task.html

Figure 2: The second stage of pre-training in mDIALBART. <En> and <Zh> are two language identifiers indicating the input and output languages. [SUM] is a special token to indicate the summarization task.
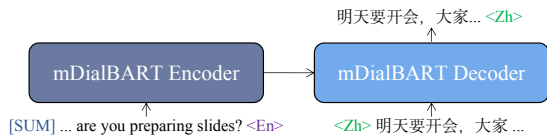

Figure 3: Finetune mDIALBART on XLDS.

Utterance Permutation (UP). We shuffle all utterances from dialogues and encourage the model to restore them.

Monolingual Dialogue Summarization (MDS). We adopt MDS as a pre-training task to enable the model to learn the summarization ability.

Machine Translation (MT). Since the XLDS task needs translation capability, we employ machine translation as one of the pre-training tasks.

We pre-train mDIALBART on XMediaSum40k and MediaSum424k corpora which are provided by the semi-supervised setting of CLIDSUM. The dialogues as well as monolingual dialogue-summary pairs used in AcI, UP and MDS come from MediaSum424k. For MT, we use the original English paired with the translated summaries from the training set of XMediaSum40k as pre-training samples.

5.2 Multi-Task Pre-Training

For each iteration of the second pre-training stage, training samples from the four tasks are randomly selected and used to calculate the summation loss and update the parameters. Thus, it is necessary to make the pre-trained model distinguish different tasks, to avoid the confusion brought by the joint training. Specifically, AcI, UP and MT could be regarded as reconstruction tasks. To distinguish MDS from others, a special token [SUM] is appended at the beginning of input sequences of MDS.

5.3 Fine-tuning on XLDS

Figure 3 shows the fine-tuning process on XLDS task. The input sequences of XLDS are prepended with token [SUM] to leverage dialogue summarization ability obtained from the second pre-training stage described in the previous section. Besides, the language identifiers at the encoder and the decoder sides are different to utilize the cross-lingual ability learned from the MT pre-training task.

6 Experiments

We evaluate mDIALBART and various baselines on CLIDSUM. Four automatic metrics are adopted in our experiments: ROUGE-1 (R1) and ROUGE-2 (R2) (Lin, 2004) evaluate unigram and bigram overlap between the generated summaries and correspondings references, respectively. ROUGE-L (R-L) (Lin, 2004) is applied to find the length of the longest common subsequence. BERTScore (B-S) (Zhang et al., 2020b) evaluates the semantic similarity of generated sentences against the references. For evaluation toolkits and the implementation details of all models, please refer to Appendix A.

6.1 Main Results

Table 3 shows the experimental results. We first analyze the performance of pipeline baselines, and then compare them with end-to-end baseline. Second, we compare mDIALBART with all baselines to demonstrate its superiority. Lastly, we introduce a simple data augmentation strategy to further improve the performance of end-to-end methods. For generated cases, please refer to Appendix B.

Pipeline Baselines. Comparing the performance between translate-then-summarize (Trans-Sum) and summarize-then-translate (Sum-Trans), we

#			XSAMSum$_{En\Rightarrow De}$ R1 / R2 / R-L / B-S	XSAMSum$_{En\Rightarrow Zh}$ R1 / R2 / R-L / B-S	XMediaSum40k$_{En\Rightarrow De}$ R1 / R2 / R-L / B-S	XMeidaSum40k$_{En\Rightarrow Zh}$ R1 / R2 / R-L / B-S
			Sum-Trans Methods			
1	PEGASUS	Trans-WMT	37.8 / 13.2 / 30.6 / 70.7	35.5 / 11.6 / 29.2 / 74.7	19.8 / 6.5 / 17.5 / 59.0	21.6 / 5.8 / 18.9 / 64.8
2		OPUS-MT	41.5 / 16.7 / 34.0 / 73.4	33.3 / 11.0 / 27.8 / 72.1	18.3 / 5.6 / 16.1 / 58.1	16.0 / 3.0 / 14.0 / 61.8
3		Google Trans.	43.4 / 18.9 / 36.1 / 74.4	38.5 / 14.3 / 31.9 / 75.8	21.1 / 7.7 / 18.7 / 60.2	22.8 / 6.6 / 20.0 / 65.5
4	T5	Trans-WMT	39.7 / 14.6 / 32.1 / 71.6	36.4 / 12.5 / 30.1 / 75.2	18.6 / 6.3 / 16.0 / 56.5	21.7 / 5.7 / 18.5 / 64.4
5		OPUS-MT	43.2 / 17.9 / 35.4 / 74.2	34.7 / 11.7 / 29.1 / 72.9	17.5 / 5.4 / 15.1 / 57.0	17.0 / 3.3 / 14.6 / 61.5
6		Google Trans.	45.3 / 20.3 / 37.4 / 75.2	39.7 / 15.2 / 33.0 / 76.6	20.3 / 7.6 / 17.6 / 59.1	22.9 / 6.7 / 19.6 / 64.4
7	BART	Trans-WMT	40.2 / 15.4 / 33.0 / 71.7	37.5 / 12.9 / 30.9 / 75.5	20.0 / 6.7 / 17.7 / 59.3	23.4 / 7.0 / 20.6 / 65.6
8		OPUS-MT	44.3 / 19.2 / 36.7 / 74.7	36.7 / 12.7 / 30.5 / 73.6	20.4 / 7.0 / 18.1 / 59.7	17.9 / 3.8 / 15.7 / 62.7
9		Google Trans.	46.5 / 21.7 / 39.0 / 75.8	41.3 / 16.7 / 34.3 / 77.1	23.7 / 9.9 / 21.1 / 61.9	24.7 / 8.0 / 21.9 / 66.3
10	mBART	Trans-WMT	39.1 / 14.0 / 31.3 / 70.9	36.5 / 12.1 / 30.1 / 75.0	17.6 / 5.4 / 15.6 / 58.1	21.1 / 5.7 / 18.5 / 64.4
11		OPUS-MT	41.3 / 16.2 / 33.5 / 73.0	32.1 / 10.3 / 27.0 / 71.6	18.0 / 5.7 / 15.8 / 58.7	15.8 / 2.6 / 13.8 / 61.1
12		Google Trans.	43.3 / 17.6 / 34.9 / 74.0	39.5 / 15.2 / 32.7 / 76.1	20.8 / 7.9 / 18.4 / 60.5	22.6 / 6.7 / 19.8 / 65.1
13	MV-BART	Trans-WMT	42.2 / 15.8 / 33.8 / 72.5	39.1 / 13.9 / 32.1 / 76.1	20.5 / 6.7 / 17.9 / 59.5	23.7 / 7.7 / 20.6 / 65.9
14		OPUS-MT	46.3 / 19.8 / 37.7 / 75.3	38.1 / 13.9 / 31.5 / 74.0	20.9 / 7.4 / 18.7 / 60.2	18.5 / 3.8 / 16.2 / 62.7
15		Google Trans.	**48.3** / <u>22.6</u> / <u>39.8</u> / **76.4**	**42.4** / **17.2** / <u>34.9</u> / <u>77.4</u>	24.0 / 9.9 / 21.1 / 61.9	24.8 / 8.6 / 22.1 / 66.6
16	BART($\mathcal{D}_{ALL}$)	Trans-WMT	42.1 / 16.8 / 34.2 / 72.8	39.2 / 13.9 / 32.2 / 76.3	20.4 / 6.7 / 17.9 / 59.6	23.6 / 7.6 / 20.8 / 66.0
17		OPUS-MT	45.3 / 19.2 / 37.0 / 74.7	35.9 / 13.2 / 30.2 / 73.8	20.7 / 7.3 / 18.7 / 60.0	18.3 / 3.7 / 16.0 / 62.5
18		Google Trans.	<u>48.0</u> / **23.1** / **40.3** / <u>76.3</u>	<u>42.1</u> / **17.2** / **35.1** / **77.6**	23.9 / 9.9 / 21.2 / 62.0	24.8 / 8.6 / 22.2 / 66.8
			Trans-Sum Methods			
19	Trans-WMT	mBART	41.2 / 15.5 / 32.9 / 73.1	38.3 / 13.9 / 31.2 / 76.1	19.3 / 6.7 / 15.9 / 59.7	23.7 / 7.7 / 20.8 / 65.8
20	OPUS-MT	mBART	42.0 / 16.8 / 34.2 / 73.7	36.5 / 13.0 / 29.8 / 75.3	20.0 / 7.5 / 17.6 / 60.0	23.1 / 7.1 / 20.3 / 65.6
21	Google Trans.	mBART	43.5 / 17.8 / 35.1 / 74.1	40.0 / 14.9 / 32.6 / 76.6	20.9 / 8.2 / 18.5 / 60.4	24.1 / 8.2 / 21.4 / 65.9
			End-to-end Methods			
22	mBART		43.4 / 17.6 / 35.0 / 74.0	39.6 / 15.5 / 32.9 / 76.6	20.6 / 7.7 / 18.2 / 60.4	23.8 / 7.8 / 21.0 / 66.0
23	+DA†		- / - / - / -	- / - / - / -	23.1 / 9.8 / 20.6 / 61.8	25.9 / 9.2 / 23.2 / 66.9
24	mDIALBART (our)†		- / - / - / -	- / - / - / -	<u>24.9</u> / <u>10.3</u> / <u>22.2</u> / <u>62.8</u>	<u>27.4</u> / <u>10.5</u> / <u>24.5</u> / <u>68.0</u>
25	+DA†		- / - / - / -	- / - / - / -	**26.7** / **12.1** / **24.0** / **63.8**	**28.7** / **11.1** / **25.7** / **68.4**

Table 3: Experimental results on CLIDSUM. The **bold** and <u>underline</u> denote the best and the second scores, respectively. † represents that the evaluated models utilize the monolingual dialogue-summary pairs provided by the semi-supervised setting. Sum-Trans: Summarize-then-translate, Trans-Sum: Translate-then-summarize.

Methods	En⇒De		En⇒Zh	
	BLEU ↑	TER ↓	BLEU ↑	TER ↓
Trans-WMT	25.7	74.4	<u>35.5</u>	<u>55.7</u>
OPUS-MT	<u>27.3</u>	<u>59.9</u>	24.1	70.2
Google Trans.	**40.4**	**47.9**	**47.8**	**44.4**

Table 4: Translation results on WMT20 (En⇒De and En⇒Zh *newstest2020*). ↑ indicates higher is better. ↓ indicates lower is better.

find that Trans-Sum outperforms Sum-Trans accompanied with the same summarizer (row 19-21 vs. row 10-12) due to the limited amount of monolingual dialogue-summary pairs in the source language. Besides, the English-centric dialogue summarization work (Chen and Yang, 2020; Feng et al., 2021) reduces the summarization error and helps Sum-Trans overtake Trans-Sum (row 15/18 vs. row 21). Moreover, as shown in Table 4, we test the performance of the adopted MT methods on *WMT20-newstest2020*. Google Trans performs best in both En⇒De and En⇒Zh directions, and OPUS-MT outperforms Trans-WMT in En⇒De translation, but is worse in En⇒Zh translation. With the same dialogue summarizer, the XLDS performance of pipeline methods is consistent with the performance of the corresponding MT method.

End-to-End vs. Pipeline Baselines. Comparing the performance between end-to-end and pipeline baselines, mBART-50 achieves competitive results with the strong Trans-Sum baselines (row 22 vs. row 21), but performs worse than the strong Sum-Trans baselines (row 22 vs. row 15/18). This is because: (1) the strong pipeline baselines adopt the sophisticated translation service which leverages a large amount of parallel data; (2) the end-to-end models need both the abilities to translate and summarize, which requires a large amount of cross-lingual training data. Nevertheless, existing datasets can not fully meet this requirement.

mDIALBART vs. All Baselines. To further explore the end-to-end paradigm on XLDS, we propose mDIALBART that extends mBART-50 through the second pre-training stage. Experimental results show that mDIALBART outperforms all baselines in all automatic metrics (row 24). Note that the data used in the second pre-training stage is provided in the semi-supervised setting of CLIDSUM. That is, mDIALBART is a semi-supervised XLDS system and thus we only evaluate mDIALBART on the test set of XMediaSum40k.

Data Augmentation. Moreover, we construct a large number of pseudo-XLDS samples by translating original English summaries from Media-Sum424k to German and Chinese via Google

Model	En⇒De R1 / R2 / R-L / B-S	En⇒Zh R1 / R2 / R-L / B-S
mBART	20.6 / 7.7 / 18.2 / 60.4	23.8 / 7.8 / 21.0 / 66.0
w/ AcI	21.8 / 8.4 / 19.3 / 61.1	24.9 / 8.4 / 22.0 / 66.4
w/ UP	21.4 / 8.0 / 18.8 / 60.8	24.2 / 8.1 / 21.4 / 66.1
w/ MDS	23.3 / 9.3 / 20.7 / 62.1	26.0 / 9.0 / 23.0 / 67.1
w/ MT	21.1 / 7.9 / 18.6 / 60.6	24.3 / 8.4 / 21.5 / 66.2
mDialBART	24.9 / 10.3 / 22.2 / 62.8	27.4 / 10.5 / 24.5 / 68.0

Table 5: Ablation study on XMediaSum40k.

Model	Gram.	Info.	Conci.
BART($\mathcal{D}_{ALL}$)+Google	0.240	0.117	0.283
Google+mBART	-0.254	-0.126	-0.343
mBART	-0.334	-0.277	-0.300
mDIALBART	**0.348**	**0.286**	**0.360**

Table 6: Human studies on XMediaSum40k (En⇒Zh).

Translation. The pseudo-XLDS samples together with the real XLDS samples from XMediaSum40k jointly train end-to-end models via a simple curriculum learning (Bengio et al., 2009) strategy, i.e., the training data is converted from pseudo data to real data in each epoch. With the help of the pseudo-XLDS samples, mBART-50 and mDIALBART significantly improve their performance on XMediaSum40k (row 23/25 vs. row 22/24). We also find that our mDIALBART outperforms mBART+DA (row 24 vs. row 23), indicating the effectiveness of the second stage pre-training. Through carefully designed pre-training objectives, the pre-trained model is more effective than simply using the data augmentation strategy.

6.2 Ablation Study

As shown in Table 5, we conduct ablation studies on XMediaSum40k to evaluate the contribution of each pre-training task. All four tasks contribute to the second pre-training stage. The most important one is MDS, which is most relevant to XLDS in all pre-training tasks. Both MDS and XLDS need the pre-trained model to understand the main content of dialogues and further summarize them. MT and UP bring less benefits than others due to the fewer pre-training samples and the lower required capability, respectively. Specifically, the number of MT pre-training samples is 20k, which is significantly fewer than other tasks. UP only requires the pre-trained model to restore the order of utterances in noisy dialogues rather than predicting new words. Such NLU-style task would make it easier for models to learn shortcuts rather than the semantic understanding and reasoning (Du et al., 2021). For ablation studies of combining two or three tasks, please refer to Appendix C.

6.3 Human Study

We conduct human studies to further evaluate the performances of the strong baselines under each paradigm and our pre-trained model, i.e., BART($\mathcal{D}_{ALL}$) + Google Trans, Google Trans + mBART, mBART and mDIALBART. We randomly select 50 samples from the test set of XMedia-Sum40k. Seven crowd workers with high levels of fluency in English and Chinese are asked to assess the generated Chinese summaries from three aspects: *grammaticality* (Gram.), *informativeness* (Info.) and *conciseness* (Conci.). Following the Best-Worst Scaling method (Kiritchenko and Mohammad, 2017), crowd workers are asked to select the best and the worst generated summaries on each criteria. The result scores are calculated based on the percentage of times each model is selected as best minus the times it is selected as worst. Thus, the final scores should range from -1 (worst) to 1 (best). Table 6 shows the results of human studies. Our mDIALBART outperforms strong baselines in all metrics, indicating its superiority. The Fleiss' Kappa scores (Fleiss, 1971) of Gram., Info. and Conci. are 0.37, 0.26 and 0.43, respectively, indicating a good inter-agreement among our evaluators.

6.4 XLDS Difficulties

To further study the specific challenges of end-to-end XLDS and give multiple promising directions for future research, we take a closer look at CLID-SUM and model generation errors. We conclude the following difficulties worthy of research attention:

Multiple Topics. The dialogues in MediaSum40k are interview transcripts from NPR and CNN (Zhu et al., 2021), and each transcript usually records multiple topics from different events. Previous dialogue summarization work (Chen and Yang, 2020; Feng et al., 2021) has proved the effectiveness of topic information and proposed topic-aware summarizers. Thus, it is also necessary to explore topic-aware end-to-end XLDS models.

Low Resource. End-to-end XLDS models learn both dialogue summarization and machine translation simultaneously, which require a large amount of training data. Intuitively, as a more difficult task, XLDS needs more training samples than MT when building an end-to-end model. Following previous MT work (Lin et al., 2020), a parallel corpus is considered as an extremely low resource when the number of its samples is less than 100k. Therefore, it is hard to learn XLDS well when only utilizing

supervised data provided by CLIDSUM. We believe the low-resource XLDS is also worth studying.

Domain Adaption. There are various dialogue domains in daily life (e.g., chat, interview and debates), and it is impractical to construct XLDS datasets for each domain. Therefore, it is valuable to utilize XLDS data of one domain to improve the model performance of others. Moreover, we find that mDIALBART performs slightly worse than mBART baseline when fine-tuning on XSAMSum, indicating its limited domain adaption capability.

7 Conclusion

In this paper, we introduce XLDS task and present CLIDSUM, the first large-scale XLDS benchmark dataset. We also propose mDIALBART, a XLDS-oriented pre-trained model that extends mBART-50 via the second pre-training stage. The carefully designed pre-training tasks help mDIALBART better adapt to the XLDS task. Experiments on CLIDSUM demonstrate that our mDIALBART outperforms strong baseline models.

Ethical Considerations

We discuss the main ethical considerations of CLIDSUM benchmark dataset as follows: (1) Licenses. CLIDSUM is derived from the SAMSum (Gliwa et al., 2019) and the MediaSum (Zhu et al., 2021), both of which are well-constructed and published datasets. We will follow the CC BY-NC-ND 4.0[8] license of SAMSum to make XSAMSum public. Following the requirements of MediaSum, we restrict our usage to research purpose only, and will release the translated MediaSum under CC-BY-NC-SA 4.0 license[8]. (2) Compensation. During the translation annotation, the salary for annotating each summary is determined by the average time of annotation and local labor compensation standard. (3) Data characteristics. We refer readers to the content and (Gliwa et al., 2019; Zhu et al., 2021) for more detailed characteristics. (4) Potential problems. While principled measures are taken to ensure the quality of the dataset, there might still be potential problems with the dataset quality, which may lead to incorrect translations in applications. We also consider potential ethical issues of our mDIALBART pre-trained model: mDIALBART inherits mBART-50 (Tang et al., 2021) and is further trained on the XMediaSum40k and MediaSum424k

---

[8]https://en.wikipedia.org/wiki/Creative_Commons_license

corpora. Therefore, mDIALBART could reflect the same biases and toxic behaviors exhibited by language models, such as biases about race and gender (Sheng et al., 2020).

Limitations

While we show that mDIALBART performs best on XMediaSum40k, there are limitations that provide avenues for future work. (1) As we discussed in Section 6.4, the domain adaption capability of mDIALBART is limited. We think future work can focus on designing general or unified XLDS pre-trained models which could be applied in multiple dialogue domains. (2) Not all the parameters of mDIALBART are completely useful, such as the token embeddings of irrelevant languages, which reduces the inference speed.

Acknowledgements

References

Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D. Manning. 2015. Leveraging linguistic structure for open domain information extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 344–354, Beijing, China. Association for Computational Linguistics.

Ayana, Shiqi Shen, Yun Chen, Cheng Yang, Zhiyuan Liu, and Maosong Sun. 2018. Zero-shot cross-lingual neural headline generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26:2319–2327.

Yu Bai, Yang Gao, and Heyan Huang. 2021. Cross-lingual abstractive summarization with limited parallel resources. 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 6910–6924, Online. Association for Computational Linguistics.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009*, volume 382 of *ACM International Conference Proceeding Series*, pages 41–48. ACM.

Yue Cao, Hui Liu, and Xiaojun Wan. 2020. 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.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Maël Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska Masson, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre D. Wellner. 2005. The ami meeting corpus: A pre-announcement. In *MLMI*.

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)*, pages 4106–4118, Online. Association for Computational Linguistics.

Jiaao Chen and Diyi Yang. 2021. Structure-aware abstractive conversation summarization via discourse and action graphs. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1380–1391, Online. Association for Computational Linguistics.

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.

Mengnan Du, Varun Manjunatha, Rajiv Jain, Ruchi Deshpande, Franck Dernoncourt, Jiuxiang Gu, Tong Sun, and Xia Hu. 2021. Towards interpreting and mitigating shortcut learning behavior of NLU models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 915–929, Online. Association for Computational Linguistics.

Xiachong Feng, Xiaocheng Feng, Libo Qin, Bing Qin, and Ting Liu. 2021. Language model as an annotator: Exploring DialoGPT for dialogue 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)*, pages 1479–1491, Online. Association for Computational Linguistics.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

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*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

John B. Haviland. 1990. Linguistic anthropology: Linguistics: The cambridge survey, volume 4. language: The socio-cultural context. frederick j. newmeyer, ed. *American Anthropologist*, 92:1036–1037.

Adam L. Janin, Don Baron, Jane Edwards, Daniel P. W. Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, and Chuck Wooters. 2003. The icsi meeting corpus. *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, 1:I–I.

Svetlana Kiritchenko and Saif Mohammad. 2017. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 465–470, Vancouver, Canada. 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.

Siyu Lai, Hui Huang, Dong Jing, Yufeng Chen, Jinan Xu, and Jian Liu. 2021. Saliency-based multi-view mixed language training for zero-shot cross-lingual classification. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 599–610, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Siyu Lai, Zhen Yang, Fandong Meng, Xue Zhang, Yufeng Chen, Jinan Xu, and Jie Zhou. 2022. Generating authentic adversarial examples beyond meaning-preserving with doubly round-trip translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4256–4266, Seattle, United States. Association for Computational Linguistics.

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.

Yunlong Liang, Fandong Meng, Yufeng Chen, Jinan Xu, and Jie Zhou. 2021. Modeling bilingual conversational characteristics for neural chat 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)*, pages 5711–5724, Online. Association for Computational Linguistics.

Yunlong Liang, Fandong Meng, Chulun Zhou, Jinan Xu, Yufeng Chen, Jinsong Su, and Jie Zhou. 2022. 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.

Zehui Lin, Xiao Pan, Mingxuan Wang, Xipeng Qiu, Jiangtao Feng, Hao Zhou, and Lei Li. 2020. Pre-training multilingual neural machine translation by leveraging alignment information. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2649–2663, Online. Association for Computational Linguistics.

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

Laiba Mehnaz, Debanjan Mahata, Rakesh Gosangi, Uma Sushmitha Gunturi, Riya Jain, Gauri Gupta, Amardeep Kumar, Isabelle G. Lee, Anish Acharya, and Rajiv Ratn Shah. 2021. GupShup: Summarizing open-domain code-switched conversations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6177–6192, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Khanh Nguyen and Hal Daumé III. 2019. Global Voices: Crossing borders in automatic news summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 90–97, Hong Kong, China. Association for Computational Linguistics.

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)*, pages 2025–2031, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Shi-qi Shen, Yun Chen, Cheng Yang, Zhi-yuan Liu, and Mao-song Sun. 2018. Zero-shot cross-lingual neural headline generation. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 26(12):2319–2327.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2020. Towards Controllable Biases in Language Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3239–3254, Online. Association for Computational Linguistics.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 3483–3491.

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.

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, 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.

Jiaan Wang, Fandong Meng, Duo Zheng, Yunlong Liang, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2022. A survey on cross-lingual summarization. *arXiv preprint arXiv:2203.12515*.

Longyue Wang, Zhaopeng Tu, Xing Wang, Li Ding, Liang Ding, and Shuming Shi. 2020. Tencent AI lab machine translation systems for WMT20 chat translation task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 483–491, Online. Association for Computational Linguistics.

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.

Chien-Sheng Wu, Linqing Liu, Wenhao Liu, Pontus Stenetorp, and Caiming Xiong. 2021. Controllable abstractive dialogue summarization with sketch supervision. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5108–5122, Online. Association for Computational Linguistics.

Feng Xiachong, Feng Xiaocheng, and Qin Bing. 2021. Incorporating commonsense knowledge into abstractive dialogue summarization via heterogeneous graph networks. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 964–975, Huhhot, China. Chinese Information Processing Society of China.

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.

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.

Jiajun Zhang, Yu Zhou, and Chengqing Zong. 2016. Abstractive cross-language summarization via translation model enhanced predicate argument structure fusing. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24:1842–1853.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020a. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, 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.

Xinze Zhang, Junzhe Zhang, Zhenhua Chen, and Kun He. 2021. Crafting adversarial examples for neural 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)*, pages 1967–1977, Online. Association for Computational Linguistics.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020c. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2022. Dialoglm: Pre-trained model for long dialogue understanding and summarization. *Proc. of AAAI*.

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*, pages 5927–5934, 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), 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   Implementation Details

A.1   Automatic Evaluation

To calculate ROUGE scores, we employ the *multi-lingual ROUGE* toolkit[9] that considers segmentation and stemming algorithms for various languages. To calculate BERTScore, we use the *bert-score* toolkit[10].

A.2   Translation Models

For OPUS-MT models, we directly utilize the pre-trained *OPUS-MT-en-de*[11] and *OPUS-MT-en-zh*[12] models. For Trans-WMT translation model, we follow Liang et al. (2021) to train transformer-base models (Vaswani et al., 2017) (512 hidden size, 2048 filter size, 8 multi-head attention, 6 encoder layers and 6 decoder layers) on WMT20 parallel corpus[13]. Specifically, for En⇒De translation, we combine six corpora including Euporal, ParaCrawl, CommonCrawl, TildeRapid, News-Commentary, and WikiMatrix. And for En⇒Zh translation, we combine News Commentary v15, Wiki Titles v2, UN Parallel Corpus V1.0, CCMT Corpus, and WikiMatrix. To pre-process the raw data, we employ a series of open-source/in-house scripts, including full-/half-width conversion, unicode conversation, punctuation normalization, and tokenization (Wang et al., 2020). After filtering steps, we generate subwords via joint BPE (Sennrich et al., 2016) with 32K merge operations. Finally, we obtain 45,541,367 sentence pairs for

En⇒De and 22,244,006 sentence pairs for En⇒Zh, respectively.

A.3   Pre-Trained Language Models

The implementation of pre-trained models used in our baselines is provided by the Huggingface Transformers (Wolf et al., 2020), i.e., *PEGASUS-large*[14], *T5-large*[15], *BART-large*[16] and *mBART-50*[17]. In the fine-tuning process of pipeline baselines, we set the batch size of PEGASUS, T5, BART and mBART to 16 for XSAMSum and 24 for XMedia-Sum40k. The corresponding learning rates are set to 4e-5 and 2e-5, respectively. All these models are fine-tuned with 20 epochs. The hyperparameters of MV-BART (Chen and Yang, 2020) and BART($\mathcal{D}_{\text{ALL}}$) (Feng et al., 2021) are the same as original paper. In the fine-tuning process of end-to-end baseline, mBART-50 is fine-tuned with 4 batch size, 5e-6 learning rate and 20 epochs.

A.4   mDIALBART

Our mDIALBART first inherits from the *mBART-50*[17] (1024 hidden size, 16 multi-head attention, 12 encoder layers and 12 decoder layers) and then suffers from the second stage of pre-training, which are conducted utilizing 8 NVIDIA Tesla V100 GPUs with 32GB memory. mDIALBART is trained using *pytorch-lightning*[18] framework with 5e-6 learning rate and 8 batch size. We set the warmup steps and total steps to 5,000 and 1,000,000 steps, respectively. The pre-trained mDIALBART is next fine-tuned on the XLDS task using a single GPU with 4 batch size and 5e-6 learning rate. The maximum number of tokens for input sequences is 1024. In the test process, the beam size is 5, and the maximum decoded length is 150.

B   Case Study

We give the generated summaries of strong baselines and mDIALBART in Figure 4. The dialogue in the left example only discusses one event, i.e., Hillary Clinton declares candidacy for the U.S. Senate seat from New York. All these models can generate a suitable summary that matches this event.

---

[9] https://github.com/csebuetnlp/xl-sum/tree/master/multilingual_rouge_scoring
[10] https://github.com/Tiiiger/bert_score. The Chinese and German BERTScores are calculated through *chinese-bert-wwm-ext* and *bert-base-german-uncased* models, respectively.
[11] https://huggingface.co/Helsinki-NLP/opus-mt-en-de
[12] https://huggingface.co/Helsinki-NLP/opus-mt-en-zh
[13] https://www.statmt.org/wmt20/translation-task.html

[14] https://huggingface.co/google/pegasus-large
[15] https://huggingface.co/t5-large
[16] https://huggingface.co/facebook/bart-large
[17] https://huggingface.co/facebook/mbart-large-50-many-to-many-mmt
[18] https://github.com/PyTorchLightning/pytorch-lightning

	English Dialogue	
	... **Hillary:** I am a new Democrat. I don't believe government is the source of all our problems, or the solution to them ... **King:** Education and health-care reform among the first lady's primary issues. Her agenda ... ... **Randall:** Mr. Sweeney, what holds the key? **Sweeney:** It is the people, and it's not whether they're going to vote Democratic or Republican, they're going to vote for what is proven. ...	... **Suzanne:** Those massive wildfires burning across a thousand miles of Australia have claimed ... **Shane:** On behalf of the firefighters, that are doing it so tough up there, we do apologize and regret that we weren't able to save everything... ... **Malveaux:** And now to Nigeria, this is where pirates have actually kidnapped two Americans from a U.S. oil supply vessel. This is in the Gulf of Guinea... ... **Malveaux:** All right, Vladimir, thank you so much. We certainly hope they're safe. ...

	Generated Summaries	
MV-BART + Google Trans.	Die first lady der demokraten **erklärt** ihre senats **kandidatur**. 民主党第一夫人**宣布参选**参议院。	**Zwei amerikaner**, die von einem us-ölversorgungsschiff **in** Nigeria **entführt** wurden. 海盗在尼日利亚**绑架美国人**。
BART($\mathcal{D}_{ALL}$) + Google Trans.	**Hillary Clinton** kündigt **kandidatur für den New York** senat an. 希拉里克林顿宣布竞选纽约参议院。	**Piraten** entführen **amerikaner** vor Nigeria. 海盗劫持**两名美国人**。
Google Trans. + mBART	**Hillary Clinton erklärt kandidatur für den** senat der USA. 希拉里宣布竞选纽约州**参议员**。	**Piraten nehmen zwei amerikaner** von einem **us-öl-schiff entführt**. 海盗袭击西非地区。
mBART	**Hillary Clinton** kündigt **kandidatur für den US-Senat** an. 希拉里·克林顿宣布竞选参议员。	**Piraten nehmen zwei amerikaner** vom **ölschiff in** den golf von guinea. 海盗袭击美国船只。
mDialBART	**Hillary Clinton** kündigt senats **kandidatur für** New York an. 希拉里·克林顿宣布参选**美国参议员**。	Großflächige **waldbrände in Australien**; **Piraten nehmen zwei amerikaner aus us-öl-schiff entführt**. 澳大利亚大规模野**火**;海盗从美国石油船上**绑架两名美国人**。
Gold Summaries	**Hillary Clinton erklärt kandidatur für den US-Senat von New York**. 希拉里·克林顿宣布在纽约竞选美国参议员.	**Feuerwehrpilot stirbt bei waldbränden in Australien; Piraten nehmen zwei amerikaner aus us-öl-schiff entführt.** 消防员飞行员在澳大利亚森林大火中丧生;非洲海盗绑架了两名美国人**。**

Figure 4: Example English dialogue in XMediaSum40k and the corresponding generated summaries of strong baselines and mDIALBART. The **bold** indicates the overlap words between generated summaries and gold summaries.

#	Model	En⇒De R1 / R2 / R-L / B-S	En⇒Zh R1 / R2 / R-L / B-S
1	mBART	20.6 / 7.7 / 18.2 / 60.4	23.8 / 7.8 / 21.0 / 66.0
2	+AcI	21.8 / 8.4 / 19.3 / 61.1	24.9 / 8.4 / 22.0 / 66.4
3	+UP	21.4 / 8.0 / 18.8 / 60.8	24.2 / 8.1 / 21.4 / 66.1
4	+MDS	23.3 / 9.3 / 20.7 / 62.1	26.0 / 9.0 / 23.0 / 67.1
5	+MT	21.1 / 7.9 / 18.6 / 60.6	24.3 / 8.4 / 21.5 / 66.2
6	+AcI+UP	22.1 / 8.5 / 19.6 / 61.3	25.1 / 8.5 / 22.2 / 66.6
7	+AcI+MDS	24.0 / 9.4 / 21.2 / 62.3	26.2 / 9.2 / 23.5 / 67.6
8	+AcI+MT	22.8 / 9.4 / 20.3 / 61.6	26.0 / 9.8 / 23.1 / 66.8
9	+UP+MDS	23.8 / 9.4 / 21.1 / 62.3	26.6 / 9.3 / 23.6 / 67.6
10	+UP+MT	22.2 / 9.1 / 19.8 / 61.2	25.3 / 9.7 / 22.6 / 66.5
11	+MDS+MT	24.2 / 10.2 / 21.6 / 62.4	27.2 / 10.2 / 24.2 / 67.6
12	+AcI+UP+MDS	23.7 / 9.7 / 21.1 / 62.2	26.6 / 9.4 / 23.5 / 67.5
13	+AcI+UP+MT	22.9 / 9.5 / 20.4 / 61.6	26.0 / 10.0 / 23.1 / 66.7
14	+AcI+MDS+MT	**24.9** / **10.5** / **22.2** / **62.9**	**27.7** / 10.5 / **24.7** / 67.9
15	+UP+MDS+MT	24.8 / **10.5** / 22.1 / 62.7	27.4 / **10.6** / 24.5 / 67.9
16	mDialBART	**24.9** / 10.3 / **22.2** / 62.8	27.4 / 10.5 / 24.5 / **68.0**

Table 7: Ablation study of mDIALBART.

However, the content of the right dialogue across from many events, we find that all baselines capture only one event (i.e., the pirate event) while our mDIALBART is aware of the two different events in the dialogue and exactly summarize it.

C Effect of Each Pre-Training Task

Table 7 shows the complete ablation studies of mDIALBART on XMediaSum40k. We find that combining two tasks in the second pre-training stage outperforms using only one of them, e.g.,

mBART (+AcI+UP) outperforms both mBART (+AcI) and mBART (+UP). mBART (+MDS+MT) performs best in the setting of combining two tasks due to the high relevance[19] with XLDS.

As for combining three or four tasks, we find that MDS and MT are vital to the second pre-training stage. The performance of the pre-trained model with both MDS and MT is significantly better than others (row 14-16 vs. row 12-13). In addition, the performance of mBART (+AcI+MDS+MT), mBART (+UP+MDS+MT) and mDIALBART is similar. We think this is because both AcI and UP help the pre-trained model capture the structural characteristics in dialogues and these two tasks are redundant to some extent.

---

[19]XLDS could be regarded as the combination of MDS and MT.