The Cross-lingual Conversation Summarization Challenge

Yulong Chen[*♠♡], **Ming Zhong**[*♣], **Xuefeng Bai**[*♠♡], **Naihao Deng**[‡],
Jing Li[♢], **Xianchao Zhu**[♢], **Yue Zhang**[♡]
[♠] Zhejiang University
[♡] School of Engineering, Westlake University
[♣] University of Illinois at Urbana-Champaign
[‡] University of Michigan, Ann Arbor
[♢] Sichuan Lan-bridge Information Technology Co., Ltd.
yulongchen1010@gmail.com *yue.zhang@wias.org.cn*

Abstract

We propose the shared task of cross-lingual conversation summarization, *ConvSumX Challenge*, opening new avenues for researchers to investigate solutions that integrate conversation summarization and machine translation. This task can be particularly useful due to the emergence of online meetings and conferences. We use a new benchmark, covering 2 real-world scenarios and 3 language directions, including a low-resource language, for evaluation. We hope that *ConvSumX* can motivate research to go beyond English and break the barrier for non-English speakers to benefit from recent advances of conversation summarization.

1 Task Overview

The cross-lingual conversation summarization (*ConvSumX*) task [1] asks models to output a salient, concise and coherent summary in target languages (*e.g.*, Chinese), given a conversation in a source language (*e.g.*, English). In particular, *ConvSumX* contains 2 tracks: daily dialogue summarization and query-based meeting minute. Each covers 3 language directions: `English-to-Chinese` (`En2Zh`), `English-to-French` (`En2Fr`) and `English-to-Ukrainian` (`En2Uk`). Figure 1 gives examples in *ConvSumX*, where we show summaries in 4 languages (including English). Both automatic and manual evaluations are used to measure the model performance, while the evaluation is highly inclined to human evaluation (Section 3.5).

2 Motivation

Thanks to the availability of large-scale corpora (Gliwa et al., 2019; Chen et al., 2021a; Zhong et al., 2021b), research on conversation summarization has made great progress (Zhong et al., 2021a; Ni et al., 2021; Ghazvininejad et al., 2021; Lin et al., 2022). However, existing corpora in this area focus on English while ignoring other languages (Feng et al., 2021a). Such English-dominated corpora lead to a barrier for non-English speakers to benefit from conversation summarization research, which becomes more severe in the era of epidemic, where international meetings are held online and participants communicate in English.

ConvSumX integrates conversation summarization and machine translation, involving the language shift from one to another and stylistic shift from long spoken conversations to concise written monologues. Ideally, using the *first translate, then summarize* and *vice versa* pipelines can solve the task. However, besides the difficulties in monolingual conversation summarization (Chen et al., 2021b; Feng et al., 2021b), pipeline methods suffer from problems caused by machine translation systems. For *translation-first* systems, Zhu et al. (2019) find that machine translation introduces errors for summarizers on news text. In addition, existing machine translation systems show poor performance on conversation text (Wang et al., 2021). For *summarization-first* systems, translating summaries without conversation context can lead to inconsistent translation, in particular for polysemous words. Take $\texttt{C}^{\texttt{En2S}^{\texttt{Zh}}}$ [2] for example. The summary "*Bob is going to the bank.*", where "*bank*" can be translated into "岸边" (river bank) or "银行" (financial bank), requires models to determine the proper translation by considering conversation context. Such issues can be also found in end-to-end systems developed for cross-lingual news summarization and directly using those methods can lead to error propagation (Zhu et al., 2019; Xu et al., 2020; Liang et al., 2022). Thus, more sophisticated designs that take care of conversation natures or data selection strategies that can make better use of silver data are in need.

[*] Equal Contribution.
[1] https://cylnlp.github.io/convsumx-challenge/
[2] The setting means the input conversation text is in English, and the output summary is in Chinese.

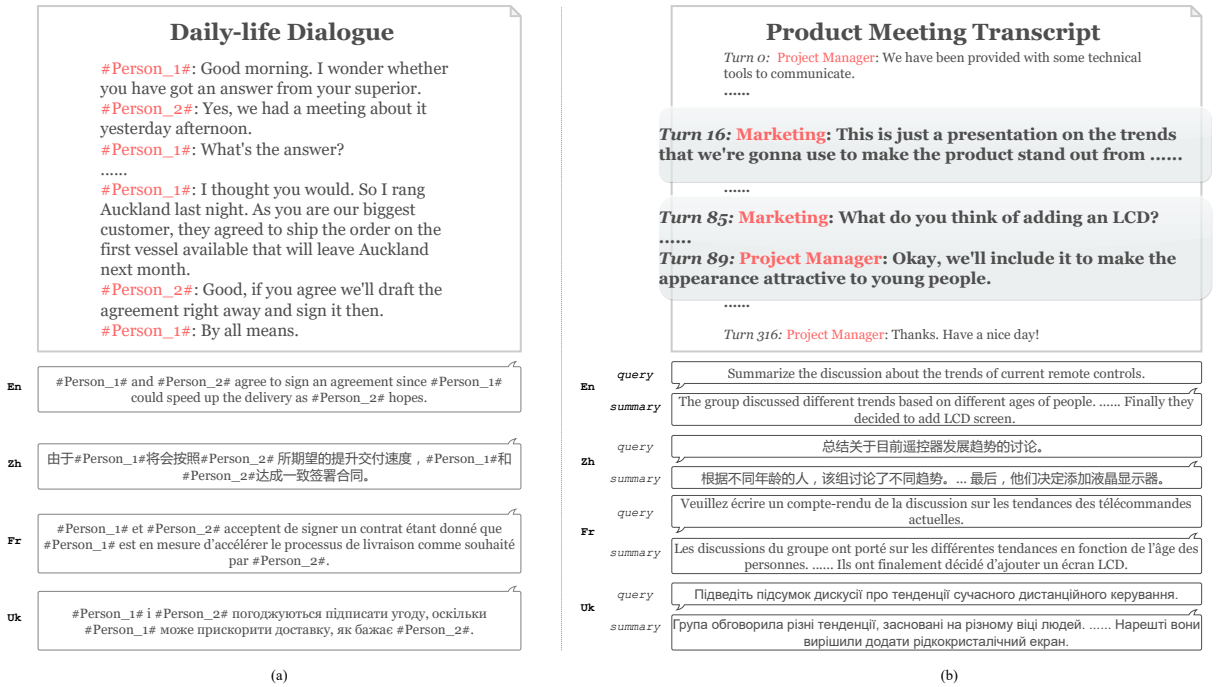

Figure 1: Examples of *ConvSumX*: (a) daily-life dialogue summarization and; (b) query-based meeting minute. From top to bottom, the languages are English (`En`), Chinese (`Zh`), French (`Fr`) and Ukrainian (`Uk`). The English conversations and summaries (and queries) are from DIALOGSUM and QMSum, respectively.

To this end, *ConvSumX Challenge* encourages researchers to investigate different solutions to cross-lingual conversation summarization. First, from the perspective of downstream applications, *ConvSumX* is useful for both business and personal uses. Second, from the perspective of research, *ConvSumX Challenge* looks for a general method that can deal with cross-lingual conversation summarization. Although only 3 typical target languages are presented in this shared task, we hope that *ConvSumX* can motivate research to a broader range of target languages. Third, from the perspective of social good, *ConvSumX* aims to break the barrier of accessing information for non-English speakers and to make them benefit from the advance of conventional English-dominated conversation summarization technologies.

We hope that *ConvSumX Challenge* can gain interest from both communities of text summarization and machine translation and also push the progress on related fields for other languages, including more low-resource languages.

3 Task Description

Formally, the task of *ConvSumX Challenge* asks participants to provide a system that can output a summary in a target language given the input conversation text in a source language.

3.1 Setting

The *ConvSumX Challenge* focuses on the low-resource/few-shot setting and cross-lingual/domain transfer technologies. The low-resource/few-shot here is stated from the perspective of *lacking large gold training data*. The term *gold data* refers to cross-lingual ⟨`conversation-summary`⟩ pairs that are annotated by translators who are expert at both source and target languages.

The reasons are: (1) gold data are limited as the annotation is very costly, in particular when conversations involve domain expert knowledge (*e.g.,* academic meeting). In contrast, machine translation and monolingual summarization data are abundant and useful (Perez-Beltrachini and Lapata, 2021); (2) we seek for a general solution that can be applied to not only the target languages in this paper, but also other languages. However, for practical consideration, we provide large silver data (Section 3.3.2). We also encourage participants to make use of other external resources to solve the task.

The above setting is widely adopted by existing cross-lingual summarization datasets in other domains, such as the first large-scale cross-lingual summarization corpus, NCLS dataset (Zhu et al., 2019) and its succeeding works (Xu et al., 2020; Bai et al., 2021; Liang et al., 2022).

Track	Data Source	Domain	Query	# Conv.	# Summ.	Train/Dev/Test
Track 1	DIALOGSUM	Daily-life Dialogue	✗	131.0	13.8	12,460/500/500
Track 2	QMSum	Product Meeting	✓	6,007.7	70.5	690/145/151
		Academic Meeting	✓	13,317.3	52.7	259/54/56

Table 1: Statistics of *ConvSumX*. # Conv. and # Summ. are averaged token lengths of conversations and summaries.

3.2 Tracks

The *ConvSumX Challenge* consists of 2 tracks, focusing on different scenarios, respectively.

- **Track 1** focuses on cross-lingual summarization for real-life dialogues. This track is in line with the INLG 2021 *DialogSum Challenge* (Chen et al., 2021b) while we extend *DialogSum* into a cross-lingual setting. *ConvSumX* can be particularly useful in scenarios such as travelling abroad where summarizers can serve as personal assistants.
- **Track 2** focuses on cross-lingual meeting minutes. Compared with daily conversations, meetings are much longer and contain richer topic switches and more professional knowledge. Generating cross-lingual meeting minutes can help non-English speakers to quickly access information of their interest, especially in cases where conferences are mostly held in English. In particular, Track 2 asks a system to generate a summary in the target language, given an input meeting text in the source language and a query in the target language.

3.3 Data

3.3.1 Data Selection

The data of *ConvSumX* are derived from two public English datasets, namely DIALOGSUM (Chen et al., 2021a) and QMSum (Zhong et al., 2021b). Table 1 shows the statistics.

DIALOGSUM is a large-scale dialogue summarization dataset, consisting of face-to-face spoken dialogues that focus on real-life scenarios. In particular, DIALOGSUM provides multi-references for each test dialogue. We ask annotators to first choose the best reference summary and then annotate it into the target languages.

QMSum is a query-based meeting minute dataset, covering 3 domains, namely academic, product and committee. We choose academic meeting and product discussion meeting for annotation as they are more in line with our motivation.

3.3.2 Annotation

Each summary in the dev and test sets of DIALOGSUM and QMSum is annotated into 3 target languages by expert translators [3]. Note that the annotation is not the simple translation of summaries, instead, each annotation needs to take care of original English conversations to ensure that the annotated summary is consistent with the input (Section 2).

In addition to manually annotated dev and test sets, following Zhu et al. (2019), we construct silver training data using machine translation. In particular, we translate summaries in target languages using multiple engines, including Google translate [4], NiuTrans [5] and LanMT [6]. Besides, to provide resources for pipeline methods, we translate the conversation texts using the same methods. Note that we do not filter these silver data. Instead, we leave this issue as an open question for the participants.

3.4 Protocol

We propose the following schedule:

- **Phase 1** (from Jul, 2022): The shared task is announced at the INLG 2022 conference, and the data are available on the shared task website; participants can register to the task.
- **Phase 2** (from Dec, 2022): The leaderboard is open; participants can submit their systems to the organizers and the online leaderboard keeps updating the best performance on each track using automatic evaluation metrics.
- **Phase 3** (from Mar, 2023): The submission is closed; organizers conduct manual evaluation.
- **Phase 4** (Jun, 2023): The *ConvSumX Challenge* shared task is fully completed. Organizers submit participant reports and challenge reports to INLG 2023 and present at the conference. The hidden test set is made public.

[3] More information can be found in Appendix A.
[4] https://translate.google.com
[5] https://niutrans.com
[6] https://www.dtranx.com

In particular, in Phase 1 participants can train and validate summarization systems on their hardwares using data provided by the organizers. Participants are encouraged to use external resources to train their systems. Such resources include, but are not limited to: monologue summarization data, machine translation data, and other public or additional cross-lingual summarization data that are manually/automatically created by the participants. However, for fairness and reproducibility, participants should specify what and how external resources are used in their system reports. In Phase 3, after the submission deadline, the organizers will start to evaluate summaries generated by final submitted models with the help from linguistic experts. For fairness, the test set will not be publicly available during the shared task.

Please note that the above schedule can be modified accordingly when the schedule of INLG 2023 is released. The leaderboard and the detailed schedule will be announced on the shared task website at https://cylnlp.github.io/convsumx-challenge/.

3.5 Evaluation

The evaluation of the *ConvSumX Challenge* considers both automatic and manual evaluation metrics.

3.5.1 Automatic Evaluation

Following previous cross-lingual summarization work (Zhu et al., 2019), we use ROUGE scores (Lin, 2004) for automatic evaluation. ROUGE scores evaluate the model performance by considering the overlap of n-grams in the system-generated summary against the reference summary. Although recent works claim that ROUGE fails to measure important information regarding factual consistency (Zhang et al., 2020; Fabbri et al., 2021), we choose ROUGE because: (1) it directly reflects model's ability of obtaining salient information and; (2) it can be easily applied to multiple languages including low-resource languages.

3.5.2 Manual Evaluation

As neural summarizers mostly contain factual errors that cannot be easily detected by automatic metrics (Zhu et al., 2019; Fabbri et al., 2021) and translated words can be various (Freitag et al., 2021), automatic evaluation such as ROUGE can be less accurate. Thus, our evaluation highly relies on manual evaluation. Given that the *ConvSumX* integrates conversation summarization and machine translation, we adopt multiple human evaluation metrics from both tasks to better measure model performance.

In particular, standard summarization metrics include: Fluency, Consistency, Relevance and Coherence (Kryscinski et al., 2019); standard machine translation metrics include: Omission, Untranslation, Mistranslation, Addition and Terminology (Mariana, 2014). However, except for Fluency, summarization metrics evaluate generated summaries from the perspective of input documents in the same language while machine translation metrics evaluate translation from the perspective of source sentences (the English summary in our case). There can be an evaluation inconsistency between these two tasks. In addition, there is an overlap between these two groups of metrics. For example, a mistranslated summary can be regarded as containing consistency errors.

To unify the aforementioned evaluation metrics and obtain fine-grained evaluations, we propose to evaluate system-generated summaries from the following aspects against source conversation texts.

Fluency and Use of Language evaluates the quality of generated sentences, including the grammar and word order. Moreover, it evaluates whether the language in generated summaries is natural and conventional, e.g., the syntactic structure is not normal or the summary contains untranslated words.

Relevance evaluates the importance of information in the generated summary.

Factual and Translation Consistency evaluates the factual alignment of the generated summaries (target languages) against the source conversation (source languages), including information that is not presented in the conversation, wrong causal relation, etc. Moreover, for pipeline methods, if the final summary contains mis-translated words, we consider it inconsistency.

Terminology evaluates the use of language. For example, the generated word can be a right translation but is improper in certain domains (*e.g.,* academic meeting).

Overall score measures the overall quality for each summary.

For each metric above, we randomly extract 10% generated summaries and ask annotators who are native in the target languages to give scores from 1 to 5. The higher, the better.

4 Related Work

Zhu et al. (2019) propose the first large scale cross-lingual news summarization dataset, facilitating the study in this filed using neural network models. Bai et al. (2021) construct an `English-to-German` news summarization dataset using the automatic method of Zhu et al. (2019). Perez-Beltrachini and Lapata (2021) construct a cross-lingual dataset based on Wikipedia, focusing on European languages. In particular, Perez-Beltrachini and Lapata (2021) use the document and the lead paragraph in other languages aligned by Wikipedia inter-language links to construct cross-lingual ⟨`document-summary`⟩ pairs. Similarly, Ladhak et al. (2020) construct the WikiLingua dataset based on multi-lingual WikiHow.

Very recently, Wang et al. (2022) and Feng et al. (2022) construct cross-lingual dialogue summarization datasets. In particular, Wang et al. (2022) manually translate summaries from SAMSum (Gliwa et al., 2019), an online written dialogue summarization dataset, and 40k data in MediaSum (Zhu et al., 2021) into German and Chinese. Feng et al. (2022) construct MSAMSum by automatically translating SAMSum into Chinese, French and Russian. Compared with them, our work focuses on spoken conversation in multiple scenrios, and covers low-resource language (`Ukrainian`). In addition, we also focus on query-based meeting scenarios, which can be more useful in real-world applications.

Similarly to *ConvSumX Challenge*, Ghosal et al. (2021) propose a shared task, AutoMin, at Interspeech 2021. AutoMin focuses on monolingual meeting minutes in English and Czech. In contrast, we focus on the cross-lingual setting and consider more scenarios, domains and languages.

5 Conclusion

We propose the *ConvSumX Challenge* to address the task of cross-lingual conversation summarization, with the hope that *ConvSumX* can encourage researchers to investigate various methods for conversation summarization beyond English, in particular for low and mid-resource languages, and the frontier of cross-lingual conversation summarization can be pushed further.

Copyright and License of Datasets

The *ConvSumX Challenge* uses cross-lingual ⟨`conversation-summary`⟩ pairs that are annotated on the top of two English conversation summarization datasets, namely DIALOGSUM and QMSum, to evaluate models. Both DIALOGSUM and QMSum are free for academic use with the MIT license, which contains no limitation to use, modification or distribution. We will also make our annotated data available for the academia.

Acknowledgement

Yue Zhang is the corresponding author. We thank all reviewers for their insightful comments and Yang Liu, Da Yin, Qian Cao and Jianyu Wang for discussion and proofreading. This project receives a support from the Sichuan Lan-bridge Information Technology Co., Ltd. and China's Language Service Center (Sichuan Province). Our sincere appreciation goes to professional translators from Lan-bridge.

References

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.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021a. DialogSum: A real-life scenario dialogue summarization dataset. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5062–5074, Online. Association for Computational Linguistics.

Yulong Chen, Yang Liu, and Yue Zhang. 2021b. DialogSum challenge: Summarizing real-life scenario dialogues. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 308–313, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating Summarization Evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

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

Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2022. Msamsum: Towards benchmarking multi-lingual dialogue summarization. In *Proceedings of the 2st Workshop on Document-grounded Dialogue and*

Conversational Question Answering (DialDoc 2022), Online. Association for Computational Linguistics.

Xiachong Feng, Xiaocheng Feng, Libo Qin, Bing Qin, and Ting Liu. 2021b. 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, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 1479–1491. Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Marjan Ghazvininejad, Vladimir Karpukhin, and Asli Celikyilmaz. 2021. Discourse-aware prompt design for text generation. *arXiv preprint arXiv:2112.05717*.

Tirthankar Ghosal, Muskaan Singh, Ondřej Bojar, and Anja Nedoluzhko. 2021. Overview of the first shared task on automatic minuting (automin) at interspeech 2021. In *Proceedings of the First Shared Task on Automatic Minuting at Interspeech 2021*, page to appear.

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.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. 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 540–551, Hong Kong, China. 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.

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. *arXiv preprint arXiv:2203.03820*.

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.

Haitao Lin, Junnan Zhu, Lu Xiang, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2022. Other roles matter! enhancing role-oriented dialogue summarization via role interactions.

Valerie R Mariana. 2014. *The Multidimensional Quality Metric (MQM) framework: A new framework for translation quality assessment*. Brigham Young University.

Ansong Ni, Zhangir Azerbayev, Mutethia Mutuma, Troy Feng, Yusen Zhang, Tao Yu, Ahmed Hassan Awadallah, and Dragomir Radev. 2021. SummerTime: Text summarization toolkit for non-experts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 329–338, Online and Punta Cana, Dominican Republic. 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.

Jiaan Wang, Fandong Meng, Ziyao Lu, Duo Zheng, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2022. Clidsum: A benchmark dataset for cross-lingual dialogue summarization. *arXiv preprint arXiv:2202.05599*.

Tao Wang, Chengqi Zhao, Mingxuan Wang, Lei Li, and Deyi Xiong. 2021. Autocorrect in the process of translation — multi-task learning improves dialogue machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*, pages 105–112, 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*, pages 536–541, Suzhou, China. Association for Computational Linguistics.

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

Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021a. Dialoglm: Pre-trained model for long dialogue understanding and summarization. *arXiv preprint arXiv:2109.02492*.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021b. 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*, pages 5905–5921, 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*, 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.

A More Information about Annotation

The cross-lingual ⟨`conversation-summary`⟩ pairs used for *ConvsumX Challenge* are constructed by expert translators from the Sichuan Lan-bridge Information Technology which is recognized as a qualified institution for translation service [7] by the ISO [8]. The entire construction process involves 9 annotators, 3 editors and 1 project manager.

For each language direction (*e.g.,* `En2Zh`), we have 3 annotators and 1 editor. All summaries and queries are first annotated by annotators and then reviewed by an editor. If bad summaries are found by editors (*e.g.,* grammar and inconsistency errors or unnatural language), the annotator would re-annotate the batch until they are qualified.

All annotators/editors are native in the target language (*i.e.*, Chinese, French or Ukrainian), and professional in English. Annotators/editors have following competences:

- translation competence and;
- linguistic and textual competence in the source language and the target language and;
- competence in research, information acquisition and processing and;
- culture competence and;
- technical competence and;
- domain competence.

In addition, annotators/editors shall meet at least one of the following criteria:

- a recognized graduate qualification in translation from an institution of higher education or;
- a recognized graduate quallfication in any other field from an institution of higher education plus two years of full-time professional experience in translating.

To monitor the whole annotation process and conduct quality control, we invite a senior translator as the project manager. The manager, who also satisfies the above requirements, has more than 5-year experience in multi-lingual translation projects that cover the language directions as described in this paper.

[7] Requirements for translation services: https://www.iso.org/standard/59149.html

[8] International Organization for Standardization: https://www.iso.org/home.html.