The Cross-lingual Conversation Summarization Challenge

Yulong Chen∗, Ming Zhong∗, Xuefeng Bai∗, Naihao Deng†,
Jing Liˌ, Xianchao Zhuˌ, Yue Zhangˌ

∗ Zhejiang University
ˌ School of Engineering, Westlake University
∗ University of Illinois at Urbana-Champaign
† University of Michigan, Ann Arbor
ˌ 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 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 $\text{CEn2Zh}$ for example. The summary “Bob is going to the bank.”, where “bank” can be translated into “岸” (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.

1The setting means the input conversation text is in English, and the output summary is in Chinese.
### 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).
## 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.

<table>
<thead>
<tr>
<th>Track</th>
<th>Data Source</th>
<th>Domain</th>
<th>Query</th>
<th># Conv.</th>
<th># Summ.</th>
<th>Train/Dev/Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Track 1</td>
<td>DIALOG SUM</td>
<td>Daily-life Dialogue</td>
<td>✗</td>
<td>131.0</td>
<td>13.8</td>
<td>12,460/500/500</td>
</tr>
<tr>
<td>Track 2</td>
<td>QMSum</td>
<td>Product Meeting</td>
<td>✓</td>
<td>6,007.7</td>
<td>70.5</td>
<td>690/145/151</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Academic Meeting</td>
<td>✓</td>
<td>13,317.3</td>
<td>52.7</td>
<td>259/54/56</td>
</tr>
</tbody>
</table>

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

### 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. 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, NiuTrans and LanMT. 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.

---

3More information can be found in Appendix A.
4https://translate.google.com
5https://niutrans.com
6https://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, Un-translation, 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


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 scenarios, 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


Conversational Question Answering (DialDoc 2022), Online. 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 by the ISO ⁷. 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;
- 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 qualification 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.

---

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