Revisiting Cross-Lingual Summarization: A Corpus-based Study and A New Benchmark with Improved Annotation

Yulong Chen[1,2] **Huajian Zhang**[2] **Yijie Zhou**[1] **Xuefeng Bai**[2] **Yueguan Wang**[2]
Ming Zhong[3] **Jianhao Yan**[2] **Yafu Li**[2] **Judy Li**[4] **Michael Zhu**[4] **Yue Zhang**[2,5] *
[1] Zhejiang University [2] Westlake University [3] UIUC
[4] Sichuan Lan-bridge Information Technology Co., Ltd.
[5] Westlake Institute for Advanced Study
yulongchen1010@gmail.com *yue.zhang@wias.org.cn*

Abstract

Most existing cross-lingual summarization (CLS) work constructs CLS corpora by simply and directly translating pre-annotated summaries from one language to another, which can contain errors from both summarization and translation processes. To address this issue, we propose ConvSumX, a cross-lingual conversation summarization benchmark, through a new annotation schema that explicitly considers source input context. ConvSumX consists of 2 sub-tasks under different real-world scenarios, with each covering 3 language directions. We conduct thorough analysis on ConvSumX and 3 widely-used manually annotated CLS corpora and empirically find that ConvSumX is more faithful towards input text. Additionally, based on the same intuition, we propose a 2-Step method, which takes both conversation and summary as input to simulate human annotation process. Experimental results show that 2-Step method surpasses strong baselines on ConvSumX under both automatic and human evaluation. Analysis shows that both source input text and summary are crucial for modeling cross-lingual summaries.

1 Introduction

With the advance in deep learning and pre-trained language models (PLMs) (Devlin et al., 2019; Lewis et al., 2020; Raffel et al., 2020), much recent progress has been made in text summarization (Liu and Lapata, 2019; Zhong et al., 2022a; Chen et al., 2022a). However, most work focuses on English (En) data (Zhong et al., 2021; Gliwa et al., 2019; Chen et al., 2021), which does not consider cross-lingual sources for summarization (Wang et al., 2022b). To address this limitation, cross-lingual summarization (CLS) aims to generate summaries in a target language given texts from a source language (Zhu et al., 2019), which has shown values to both academic and industrial communities (Bai et al., 2021; Perez-Beltrachini and Lapata, 2021).

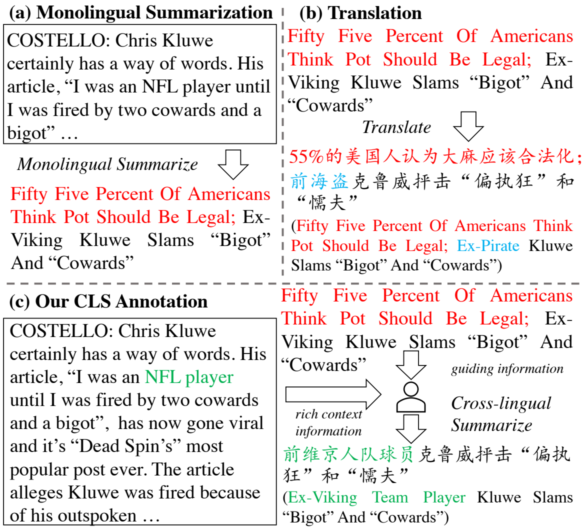
(a) Monolingual Summarization
COSTELLO: Chris Kluwe certainly has a way of words. His article, "I was an NFL player until I was fired by two cowards and a bigot" …
Monolingual Summarize
Fifty Five Percent Of Americans Think Pot Should Be Legal; Ex-Viking Kluwe Slams "Bigot" And "Cowards"
(b) Translation
Fifty Five Percent Of Americans Think Pot Should Be Legal; Ex-Viking Kluwe Slams "Bigot" And "Cowards"
Translate
55%的美国人认为大麻应该合法化；前海盗克鲁威抨击“偏执狂”和“懦夫”
(Fifty Five Percent Of Americans Think Pot Should Be Legal; Ex-Pirate Kluwe Slams "Bigot" And "Cowards")
(c) Our CLS Annotation
COSTELLO: Chris Kluwe certainly has a way of words. His article, "I was an NFL player until I was fired by two cowards and a bigot", has now gone viral and it's "Dead Spin's" most popular post ever. The article alleges Kluwe was fired because of his outspoken …
Fifty Five Percent Of Americans Think Pot Should Be Legal; Ex-Viking Kluwe Slams "Bigot" And "Cowards"
guiding information
rich context information
Cross-lingual Summarize
前维京人队球员克鲁威抨击“偏执狂”和“懦夫”
(Ex-Viking Team Player Kluwe Slams "Bigot" And "Cowards")

Figure 1: An En-Zh summary from Wang et al. (2022a) (best viewed in color). We compare "*pipeline*: (a)→(b)" annotation protocol and our annotation (c) protocol. Pipeline annotation results in errors from both summarization (red: unmentioned content/hallucination) and translation (cyan: incorrect translation) processes. To address this issue, we explicitly annotate target-language summaries with faithfulness rectification (green) based on input context, with the guidance of mono-lingual summaries.

Most existing work (Zhu et al., 2019; Bai et al., 2021; Feng et al., 2022) constructs CLS corpora by translating summaries from existing mono-lingual summarization datasets into other languages, which is de facto a "*pipeline*" annotation protocol (first *summarize*, then *translate*) as shown in Figure 1. However, such an annotation method can suffer from two major problems: First, summaries from mono-lingual summarization corpora (summarization process) can contain errors (Liu et al., 2022), which are likely to be preserved in translated summaries. For example, the English summary in Figure 1-(a) contains unmentioned con-

[*] Yue Zhang is the corresponding author.

tent/hallucination (red text), which leads to the same discrepancy as in the translated summary (Figure 1-(b), red text). Second, the translation process can further introduce errors, in particular for polysemous words. For example, in Figure 1-(b), the term, "*Ex-Viking*" (which refers to previous members of the Minnesota Vikings team), is mistakenly translated into "前海盗" (which means "*ex-pirate/buccaneer*"). To determine proper translation, it require more information beyond the scope of short summaries.

To qualitatively understand the above problems, we conduct human evaluation and error analysis on existing popular CLS corpora. Empirical results show that existing corpora suffer from the two aforementioned problems, containing significantly many hallucinations and factual errors.[1] In particular, we find that overall $20 \sim 67\%$ of summaries in CLS datasets contain errors, where $7 \sim 46\%$ and $13 \sim 47\%$ of summaries suffer from summarization and translation processes, respectively. This suggests that the pipeline protocol, which is widely used in CLS research, can result in low-quality data and negatively impact the validity of modeling research. In addition, fine-grained error analysis shows that $55.6 \sim 89.1\%$ of translation errors can be resolved with the help of input context.

Motivated by the above findings and to address this issue, we propose the protocol that cross-lingual summaries should be sourced from original input text where mono-lingual summaries can serve as a quick review for salient information. With this concept, we annotate cross-lingual summaries (S^{tgt}) by relying on source text (D^{src}) and source-language summaries (S^{src}) as shown in Figure 1-(c). Such an annotation protocol brings three advantages: First, compared with translation only given S^{src}, rich context information from D^{src} helps annotators to disambiguate word senses and comprehend S^{src} accurately, e.g., "前维京人队" (which means "*ex-viking team player*") in Figure 1-(c); Second, D^{src} is more reliable and can provide ground-truth information to correct potential errors in S^{src}, e.g., red text in Figure 1-(a); Third, compared with writing S^{tgt} only given D^{src}, S^{src} can serve as supplement guidance to help annotators be aware of what should be involved in the summaries, ensuring that salient information in S^{src} and S^{tgt} is aligned.

Using CLS protocol, we build ConvSumX, a new benchmark to facilitate future CLS research. ConvSumX focuses on conversational text in a few-shot setting. Compared with monologue (e.g., news), conversational text is less explored yet is also practically useful in real-world scenarios (Chen et al., 2022b). ConvSumX contains two sub-tasks, namely DialogSumX and QMSumX, based on two English conversation summarization datasets DialogSum (Chen et al., 2021) and QMSum (Zhong et al., 2021), respectively. Each covers three language-directions, taking En as the source, and Mandarin (Zh), French (Fr) and Ukrainian (Ukr) as target languages. We empirically compare different annotations using the pipeline protocol and our CLS protocol with human evaluation. Analysis shows that by considering input context, our protocol can significantly reduce annotation errors, suggesting ConvSumX is a high-quality benchmark in terms of cross-lingual faithfulness.

Based on the same intuition that D^{src} and S^{src} can serve as a critical complement to each other, we propose a 2-Step framework for CLS, which fine-tunes a multi-lingual PLM using concatenated S^{src} and D^{src} as input, and S^{tgt} as output. Experimental results show that our conceptual framework yields surprisingly better performance over strong baselines on ConvSumX. Analysis and human evaluation show that our method can effectively generate more faithful cross-lingual summaries in a low-resource setting, and verify that source input text and summaries are supplementary to each other in modeling cross-lingual summaries.

To summarize, our contributions are the following:

1. We systematically review the pipeline annotation protocol and show that such a protocol can result in low-quality data (§ 2);
2. We propose the concept that CLS should be sourced from both source input text and source-language summaries and under our protocol, we present ConvSumX benchmark (§ 3), where QMSumX is the first query-focused CLS dataset.
3. Under the same concept, we propose a simple yet effective 2-Step framework for CLS (§ 4), which demonstrates the necessity of both source input text and mono-lingual summary for CLS modeling.

We release ConvSumX at `https://github.com/cylnlp/ConvSumX`.

[1] The term *error* later in this paper refers to errors that are hallucinations or can cause factual misunderstandings, except when otherwise specified.

2 Analyzing Existing CLS Corpora

We conduct a corpus-based study on existing popular human-annotated CLS corpora, namely NCLS, XSAMSum and XMediaSum, covering both monologue and dialogue texts.

NCLS (Zhu et al., 2019) is the first large cross-lingual news summarization corpus, which is constructed by automatically translating existing mono-lingual summarization datasets and using a round-trip strategy with human post-editing on test sets.

XSAMSum and **XMediaSum** are both from ClidSum (Wang et al., 2022a), where they manually translate summaries from two English dialogue summarization datasets, namely SAMSum (Gliwa et al., 2019) and MediaSum (Zhu et al., 2021), into Mandarin and German.

2.1 Error Analysis on *Pipeline* Annotation

Since all 3 corpora have the task of summarizing English (En) documents into Mandarin (Zh) summaries, we perform human evaluation on this language direction. For each corpus, we randomly extract 100 instances from its training and testing sets, respectively, resulting in a total of 600 instances to evaluate. Each instance consists of English document (D^{en}) and summary (S^{en}), and Mandarin summary (S^{zh}).

We invite two expert translators, who are native in Mandarin and professional in English, as our judges and ask them to first evaluate whether the S^{zh} contains errors or not, by evaluating the S^{zh} against D^{en} (IAA[2]: 0.67, substantial agreement). If S^{zh} is found errors, the judges are asked to identify where such errors come from (IAA: 0.80, substantial agreement). Specifically, if this error is also found in S^{en}, we regard that it is caused by the mono-lingual summarization process; if this error is only found in S^{zh} but not in S^{en}, we regard that it is caused by the translation process. In this process, we only focus on factual errors, and minor syntax errors are ignored.

Corpora		Overall	Summ.	Trans.
NCLS	Train	67	46	47
	Test	60	36	40
XMediaSum	Train	27	11	19
	Test	27	10	18
XSAMSum	Train	35	13	23
	Test	20	7	13

Table 1: Error analysis on 3 CLS corpora. We randomly sample 100 instances from the training and test sets, respectively, and count the number of factual errors. Summ. indicates that the mono-lingual summarization process contains errors, which are preserved in cross-lingual summaries; Trans. denotes that the errors in cross-lingual summaries are caused by the translation process. One sample can have multiple errors.

Table 1 shows the evaluation result. Overall, we see that all CLS corpora show high error frequencies ($20 \sim 67\%$), indicating existing CLS can be less accurate. In particular, all mono-lingual summarization annotation contains errors ($7 \sim 46\%$), which are preserved in the CLS corpora. Moreover, the cross-lingual annotation process can invite more errors ($13 \sim 47\%$). This verifies our assumption that the pipeline annotation protocol, which ignores valuable input context, can lead to poor data quality.

In particular, NCLS contains the most errors, which can be because in addition to the different quality of their original mono-lingual summaries, S^{zh} in NLCS are automatically translated by MT systems. Although human post-editing is conducted on the test set, factual errors are still frequent in the test set compared with the training set. This can be because their post-editing focuses on poor fluency and translationese, while correcting factual errors or hallucinations requires information from the source text, which is not presented to human editors. In addition, the averaged number of words in NCLS is much larger than in XMediaSum and XSAMSum,[3] making translation more difficult.

The major contradiction between frequent errors according to our analysis and the high data quality reported by (Zhu et al., 2019) and (Wang et al., 2022a) can be explained by different reference sources, where our results show that these datasets have limitations in the choice of source for reference. For example, when only given S^{en} ("*Fifty Five Percent... Ex-Viking...*") as reference, an S^{zh} ("55%的美国人...前海盗") can be considered as a correct translation (Figure 1-b). However, when evaluated against D^{en}, S^{zh} is considered to have hallucination errors ("55%的美国人(*fifty five percent...*)") and impropoer translation ("前海盗(*ex-pirate*)", which should have been translated into to "前维京人队员(*ex-viking team member*)").

[2] We measure Inter-Annotator Agreement (IAA) by calculating their Pair-wise Cohen kappa score on 60 quiz instances.

[3] Avg. token. length in English summaries: NCLS (55.2) XMediaSum (14.4), XSAMSum(20.3).

Corpora		Translation Errors					
		W.S.	Ter.	C.	S.R.	Oth.	All
NCLS	Train	25	6	2	4	12	49
	Test	23	5	5	8	5	46
XMS	Train	8	3	1	3	8	23
	Test	5	3	0	3	8	19
XSS	Train	9	5	4	4	5	27
	Test	4	1	1	2	5	13

Table 2: Fine-grained categorization of translation errors. Here we report the error count of each type. W.S, Ter., C., S.R., and Oth. stand for Word Sense, Terminology, Coreference, Sentence Relation, and Others. Note that one summary can have multiple errors.

2.2 In-depth Analysis on Translation Errors

To further understand why directly translating English summaries can invite so many errors, we perform an error analysis on summaries containing translation errors and categorize them. In particular, the two judges first identify whether the translation error can be resolved by considering the input context, or not, assuming that the errors can be caused by lacking input context (e.g., polyseme translation), and other translation errors (e.g., inconsistent translation). We categorize the former error types based on their linguistic typologies (avg. IAA: 0.62, substantial agreement):

Word Sense (W.S.): the translation of a word/phrase is incorrect under source input context.

Terminology (Ter.): the translation of a word/phrase can be semantically correct but is improper in source input domains.

Coreference (C.): the translation of coreference expressions refer to incorrect objectives.

Sentence Relation (S.R.): The relation between two sentences/clauses is induced incorrectly or the translation of a sentence is incorrect because of misunderstanding the interrelation/structure of a sentence.

Others (Oth.): simple errors such as typos or less accurate translation.

Table 2 presents the error types and their error counts. First, we see that errors (W.S, Tem., C. and S.R. together: $8 \sim 41$) caused by lacking input context are more than other translation errors (Oth.: $5 \sim 12$). This further suggests the necessity of considering input text when annotating CLS corpora. In addition, word sense sees overall most errors ($26.32 \sim 51.02\%$, avg. 41.81%), which is in line with the intuition that lacking context can mostly lead to word sense ambiguity. Moreover, all categories see error instances, suggesting that such problematic summaries can confuse humans at multiple levels of language understanding.

Appendix A shows detailed information about our judges and Appendix B shows cases of different translation error types and their analysis.

3 ConvSumX

To address the aforementioned issues in pipeline annotation, we propose ConvSumX with a new annotation protocol, focusing on *few-shot* CLS. ConvSumX contains two cross-lingual summarization scenarios, namely daily dialogue summarization, and query-based summarization, covering 3 language directions: En2Zh, En2Fr and En2Ukr.

3.1 Data Source

We choose DIALOGSUM (Chen et al., 2021) and QMSum (Zhong et al., 2021) for ConvSumX by considering their potential to build real-world applications, and annotating their test and dev sets.

DIALOGSUM DIALOGSUM (Chen et al., 2021) is a real-life scenario dialogue summarization dataset, including various types of task-oriented dialogues.

QMSum QMSum (Zhong et al., 2021) is a query-based meeting summarization dataset, covering the academic, product and committee domains. We select data from academic and product for annotation.

3.2 Annotation

As discussed in § 2, the final quality of CLS corpora can be influenced by both summarization process and translation process, most of which can be resolved with the information from input documents. Therefore, instead of merely focusing on summaries in source languages, we ask annotators to write summaries in target languages (S^{tgt}) directly by considering both input documents (D^{src}) and pre-annotated summaries (S^{src}). We refer to our protocol as CLS protocol.

We take English as the source language and choose Mandarin, French and Ukrainian as target languages because they are from different language families, and have different morphological variations and syntactic structures, with the potential to benefit other languages in their families. We invite expert translators, who are native in target

Corpora		Summ.	Query
DIALOGSUM	Dev	34/500	–
	Test	21/500	–
QMSum	Dev	33/199	7/199
	Test	11/209	0/209

Table 3: Error analysis on QMSum and DIALOGSUM. we show the number of error summaries/data size.

languages and professional in English, as our annotators (Appendix A). We ask annotators to first comprehend D^{src}, and then write S^{tgt} with the help of S^{src}. In addition to the standard annotation criteria of DIALOGSUM and QMSum, we ask our annotators specifically pay attention to the following aspects featuring the CLS:

- Cross-lingual Consistency: Although being in different languages, the core semantic information of S^{tgt} should be consistent with D^{src}, in particular for polysemous words or phrases.
- Language Style and Terminology: Annotators should write S^{tgt} in the same language style of S^{src}, and use proper terminologies in some certain domains, such as academic meetings.
- Translationese: The annotated summaries should be natural in the target languages.

For QMSum, annotators are additionally asked to write a query in target languages (Q^{tgt}) with the help of the query in source language (Q^{src}), where Q^{tgt} and S^{tgt} form a QA pair.

Before annotation, we ask each annotator to label training samples (10% of each dataset) until all annotated instances meet our requirements. After annotation, each instance is reviewed by an editor, who is also an expert translator. Editors are asked to first read the annotated summary to identify whether it is natural and readable in target languages, and then evaluate it against source input document to identify whether there are any factual errors. If any errors are found, we ask the corresponding annotator to re-annotate the whole batch and repeat this checking and re-annotation process until all summaries are correct. As mono-lingual summarization process can also contain large errors (§ 2.1), we additionally require annotators to modify English summaries/queries if any errors are found. Table 3 presents the percentage of summaries that contain errors in the original datasets.

Finally, we split the original dev sets into our new training and dev sets and keep the test set unchanged (DialogSumX: 400/100/500 and QMSumX: 157/40/209).

Corpora		Overall	Summ.	Trans.
DialogSumX	T+D	2	0	2
	Test	0	0	0
QMSumX	T+D	2	0	2
	Test	1	0	1
DialogSum-P	T+D	16	9	9
	Test	11	5	7
QMSum-P	T+D	31	19	18
	Test	19	9	13

Table 4: Comparison between CLS and pipeline annotation protocols. We count the number of different errors on 100 instances, respectively. T+D: Training and Dev sets, which are the original dev set.

3.3 Comparison between ConvSumX with *Pipeline* Annotation Data

To qualitatively compare CLS and pipeline annotation protocols in a fair setting (e.g., to remove the influence of different data sources), we additionally annotate instances using the pipeline approach, i.e., directly translating English summaries into Mandarin. We randomly sample 100 instances from dev/test sets of DIALOGSUM and QMSum, referring to them as DialogSum-P and QMSum-P, respectively. Overall, we have 400 instances to annotate and 800 instances to evaluate.

These data are annotated by the same annotators, using the same quality control process as ConvSumX. To avoid priori knowledge from input context for pipeline annotation, this process is conducted *before* ConvSumX annotation. Then, we perform human evaluation on those translated data and corresponding data in ConvSumX using the same method as described in § 2.1 in an anonymous way. For ConvSumX, we take corrected English summaries as *pseudo* translation for evaluation. Table 4 shows the human evaluation results.

Consistent with our findings (§2.1), DialogSum-P and QMSum-P contain errors ($11 \sim 31$) from both the summarization and translation processes. In contrast, ConvSumX contains fewer errors ($0 \sim 2$),[4] indicating the necessity of our CLS annotation protocol.

[4] All errors that we find in § 3.3 are further corrected in ConvSumX. The final ConvSumX, which is used for training and evaluating models in § 5, contains no errors that we can find.

Corpora	Domain	Lan. Direct	Annotation	D^{src}	S^{src}	S^{tgt}	% E.
En2ZhSum	News	En2Zh	$D^{src} \rightarrow S^{src} \rightsquigarrow S^{tgt}$	755.0	55.2	96.0	33.5
Zh2EnSum	News	Zh2En	$D^{src} \rightarrow S^{src} \rightsquigarrow S^{tgt}$	103.7	17.9	13.7	-
En2DeSum	News	De2En	$D^{src} \rightarrow S^{src} \rightsquigarrow S^{tgt}$	31.0	8.5	7.5	-
XSAMSum	Written chit-chat	En2Zh/De	$D^{src} \rightarrow S^{src} \rightarrow S^{tgt}$	83.9	20.3	33.0/19.9	27.5/-
XMediaSum	Interview	En2Zh/De	$D^{src} \rightarrow S^{src} \rightarrow S^{tgt}$	1555.4	14.4	30.0/14.8	27.0/-
DialogSumX	Real-life dialog	En2Zh/Fr/Ukr	$\{D^{src}, S^{src}\} \rightarrow S^{tgt}$	131.9	19.9	53.0/22.0/17.3	1.0/-/-
QMSumX	Q-F meeting	En2Zh/Fr/Ukr	$\{D^{src}, S^{src}\} \rightarrow S^{tgt}$	1916.2	63.5	114.4/72.1/49.9	1.5/-/-

Table 5: Statistics of ConvSumX and other human-crafted CLS datasets. Lan. Direct: language direction. #: averaged length. D^{src},S^{src} and S^{tgt} are text length. We calculate character length for Mandarin and token length for others. Q-f: Query-focused. % E.: averaged sampled error rate. Both Zh2EnSum (Zhu et al., 2019) and En2DeSum (Bai et al., 2021) are constructed using the same method of En2ZhSum (Zhu et al., 2019). "$\rightarrow$": human annotation. "$\rightsquigarrow$": automatic generation with human post-editing.

3.4 Characteristics of ConvSumX

Table 5 presents a comparison between ConvSumX and other CLS corpora, highlighting the unique features of ConvSumX. Firstly, ConvSumX is designed for spoken conversation summarization and encompasses two real-world scenarios. Notably, QMSumX is the first corpus addressing query-based CLS. Secondly, ConvSumX includes multiple languages from diverse families (French: Romance; Mandarin: Chinese; Ukrainian: Slavic; English: Germanic), positioning it as a valuable resource for studying cross-lingual generalization and language transfer. Furthermore, ConvSumX is the pioneering benchmark for CLS research involving the low-resource language, Ukrainian. Last, ConvSumX is the first CLS benchmark that forsakes the pipeline annotation protocol, which is essentially different from all existing human-crafted corpora. The low error frequencies demonstrate its cross-lingual faithfulness.

4 Method

4.1 Setting

Generally, the task of *few-shot CLS* is defined as: given a source input text D^{src}, few-shot CLS is to generate a summary in a target language S^{tgt} by learning a limited number of gold-annotated $\langle D^{src}, S^{tgt}\rangle$ data, with the help of external knowledge, which can be from mono-lingual summarization data, machine translation data and PLMs.

Specifically, for *query-focused CLS*, the system is asked to generate S^{tgt} given D^{src} with a query in the target language Q^{tgt}.

4.2 Models

We evaluate two standard CLS baselines, namely pipeline method and End2End method, and propose a novel 2-Step framework, which differ from each other in the way the cross-lingual summary is generated. Figure 2 summarizes the main difference between their workflows.

Pipeline Method Previous work decomposes CLS into mono-lingual summarization and machine translation (Zhu et al., 2019), by deploying *first-summarize, then-translate* (*S-T*) or *first-translate, then-summarize* (*T-S*) strategies.

We compare with *S-T* as it can benefit from large mono-lingual summarization and monologue translation data, while *T-S* has been proven much worse (Feng et al., 2022) as both dialogue translation and non-English summarization data are very limited. For QMSumX, we additionally translate Q^{tgt} into Q^{src} before mono-lingual summarization and translation, to which we refer as *T-S-T*.

End2End Method Previous work models the CLS task and has shown better performance on previous datasets compared with pipeline methods (Zhu et al., 2019; Xu et al., 2019).

We compare two End2End methods: First, we directly fine-tune a multi-lingual model on $\langle D^{src}, S^{tgt}\rangle$ (DialogSumX) and $\langle\{Q^{tgt}; D^{src}\}, S^{tgt}\rangle$ (QMSumX), marked as E2E; Second, inspired by Bai et al. (2021), where an End2End model first generates mono-lingual summary and then cross-lingual summary in an auto-regressive way and shows good performance in few-shot setting, we fine-tune a multi-lingual model on $\langle D^{src}, \{S^{src}; S^{tgt}\}\rangle$ (DialogSumX) and $\langle\{Q^{tgt}; D^{src}\}, \{S^{src}; S^{tgt}\}\rangle$ (QMSumX), marked as E2M (M means mixed).

2-Step Method Inspired by our data analysis (§ 2) that mono-lingual summary can help guiding salient information for cross-lingual summary,

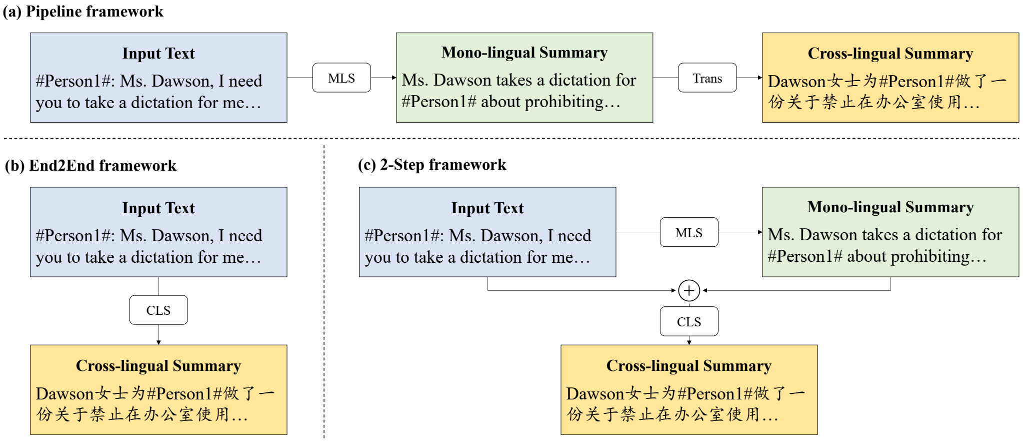
(a) Pipeline framework
Input Text
#Person1#: Ms. Dawson, I need you to take a dictation for me…
MLS
Mono-lingual Summary
Ms. Dawson takes a dictation for #Person1# about prohibiting…
Trans
Cross-lingual Summary
Dawson女士为#Person1#做了一份关于禁止在办公室使用…
(b) End2End framework
Input Text
#Person1#: Ms. Dawson, I need you to take a dictation for me…
CLS
Cross-lingual Summary
Dawson女士为#Person1#做了一份关于禁止在办公室使用…
(c) 2-Step framework
Input Text
#Person1#: Ms. Dawson, I need you to take a dictation for me…
MLS
Mono-lingual Summary
Ms. Dawson takes a dictation for #Person1# about prohibiting…
CLS
Cross-lingual Summary
Dawson女士为#Person1#做了一份关于禁止在办公室使用…

Figure 2: Illustration of pipeline method, end2end method, and our 2-Step method. MLS: mono-lingual summarizer; CLS: cross-lingual summarizer; Tans: translator.

and generating proper translation requires information from source input text, we proposed a 2-Step method. Conceptually, 2-Step is designed to simulate human annotation, where we ask an end2end model to generate S^{tgt} given concatenated S^{src} and D^{src}. Compared with pipeline methods, 2-Step method can explicitly make use of information from source input. Compared with End2End methods, 2-Step can focus on relevant information with the help of mono-lingual summaries.

Similarly, for QMSumX, we obtain the source language summaries by first translating Q^{tgt} into Q^{src} and then using mono-lingual summarizers. During inference, we use model-generated source summaries as S^{src}, which are obtained using the same way of pipeline methods.

Note all individual models are in a seq2seq manner. The terms "*pipeline*", "End2End" and "2-Step" are stated from the perspective between source input text and output cross-lingual summaries.

5 Experiments

Metrics For automatic evaluation, we use ROUGE (Lin, 2004)[5] and BERTSCORE (Zhang et al., 2020)[6]. ROUGE measures the n-gram overlap between generated and reference summaries. BERTSCORE calculates the pariwise cosine similarity between BERT (Devlin et al., 2019) token embeddings of generated and reference summaries. We report the F-1 scores of ROUGE-1 (R1), ROUGE-2 (R2), ROUGE-L (RL) and BERTSCORE (BS).

Implementation Details For mono-lingual generation, we use UNISUMM[7] for model initialization, further pre-training it on original training sets of DIALOGSUM and QMSum, and then prefix-tuning it on our few-shot training data. For cross-lingual generation (MT or CLS), we use `mBART-large-50-many-to-many-mmt`[8] for model initialization and then fine-tune it on our cross-lingual data. All experiments are conducted on NVIDIA A100 GPU. We conduct a hyper-parameter search for learning rate and batch size, from [1.5e-4, 1e-4, 5e-5, 3e-5, 1e-5] and [8, 16, 32, 64], and choose the best checkpoint based on R2 score on our few-shot dev sets.

5.1 Main Results

The main results on DialogSumX (*DX*) and QMSumX (*QX*) are shown in Table 6. In general, we find that our 2-Step system achieves the best results in most languages and the best averaged results on both tasks. In particular, 2-Step system outperforms pipeline method (*S-T*) (avg. improvement: 0.19 R2 and 0.24 BS scores on *DX*; 0.61 R2 and 1.39 BS scores on *QX*). It also outperforms End2End models by a large margin (avg. improvement: 4.73 $\sim$ 5.78 R2 and 2.36 $\sim$ 2.79 BS scores on *DX*; 1.65 R2 and 2.69 BS scores on *QX*). Note that 2-Step system is additionally presented with source summary and input text information compared with E2E and *S-T* systems. Thus, the superiority of 2-Step demonstrates that the source document and source summary are cru-

[5] https://github.com/csebuetnlp/xl-sum
[6] https://github.com/Tiiiger/bert_score
[7] https://github.com/microsoft/UniSumm
[8] https://huggingface.co/facebook/mbart-large-50-many-to-many-mmt

Model	DialogSumX															
	En2Zh				***En2Fr***				***En2Ukr***				***Avg.***			
	R1	R2	RL	BS	R1	R2	RL	BS	R1	R2	RL	BS	R1	R2	RL	BS
S-T	46.32	24.08	39.51	78.36	46.12	23.66	**37.76**	80.43	**36.19**	18.44	**31.80**	**78.30**	42.88	22.06	36.36	79.03
E2E	41.33	20.14	34.74	76.66	39.96	17.81	31.14	77.73	31.42	14.61	26.95	76.33	37.57	17.52	30.94	76.91
E2M	39.12	18.94	33.70	75.45	39.51	16.96	30.92	77.33	30.24	13.52	26.03	76.11	36.29	16.47	30.22	76.30
2-Step	**46.87**	**24.48**	**39.92**	**79.10**	**46.19**	**23.82**	37.65	**80.46**	36.05	**18.46**	31.60	78.24	**43.04**	**22.25**	**36.39**	**79.27**

Model	QMSumX															
	En2Zh				***En2Fr***				***En2Ukr***				***Avg.***			
	R1	R2	RL	BS	R1	R2	RL	BS	R1	R2	RL	BS	R1	R2	RL	BS
T-S-T	31.89	7.82	22.03	68.45	38.74	13.49	24.26	74.19	20.15	5.55	**14.44**	71.57	30.26	8.95	20.24	71.40
E2E	30.74	6.84	21.98	67.81	35.81	11.38	22.24	72.96	16.76	4.52	12.22	69.54	27.77	7.58	18.81	70.10
E2M	30.09	6.59	20.91	67.47	32.51	10.01	20.66	70.90	17.93	4.88	12.92	69.58	26.84	7.26	18.16	69.32
2-Step	**33.20**	**8.43**	**23.12**	**69.36**	**38.91**	**13.52**	**24.37**	**74.27**	**20.51**	**5.73**	14.38	**71.75**	**30.87**	**9.23**	**20.63**	**72.79**

Table 6: Main results on ConvSumX. *S-T* and *T-S-T*: pipeline methods that decompose CLS as mono-lingual summarization and translation tasks; E2E: End2End method that directly generates target summaries; E2M: End2End method that generates source summaries and target summaries sequentially; 2-Step: our method that first generates source summaries, and generates target summaries with mono-lingual summaries as guiding information.

cial in modeling cross-lingual summaries, and are complementary to each other.

Moreover, *S-T* outperforms End2End models. The contradiction between our results and previous findings (Bai et al., 2021; Chen et al., 2022b) can be explained by the fact that the summarizer and translator we use are much stronger and the error propagation problem is less severe. Also, *S-T* can benefit from our high-quality parallel cross-lingual summary pairs (S^{src} and S^{tgt}) as few-shot translation data, while previous work ignores such valuable data and only uses a fixed MT system without fine-tuning (Zhu et al., 2019).

All CLS systems perform better at En2Zh and En2Fr than En2Ukr. The high performance on En2Zh and En2Fr can be explained by that both Zh and Fr are highly-rich resource data on which mBART-50 is pre-trained (Tang et al., 2021), and mBART-50 can easily bridge the alignment between texts in Zh/Fr and En. In contrast, Ukr is a low-resource language, on which the mBART-50 performs poorly. All systems have higher performance on *DX* compared with *QX*, which is because *QX* is more challenging w.r.t the task of query-based summarization for long text and more extreme few-shot setting, and its domain is very different from mBART-50's pre-training data.

We notice that all models perform better on *QX* En2Fr than En2Zh and En2Ukr. A possible reason can be that *QX* contains many professional in-domain words whose word sense can be multiple and very different from its general ones. The sense of these words can be different lexical items, in particular for Zh or Ukr, which are typologically different from En (Chen and Ng, 1989; Budzhak-Jones, 1998). In contrast, Fr and En both use Latin script and are more similar in terms of morphology and lexicon rules (Kirsner et al., 1984; Pacton and Deacon, 2008; Fan et al., 2021) compared with Zh and Ukr. For example, “*discourse*” can be mapped into “论文(academic paper)/讲述(talk)/...” in Zh and “дискусицч (discussion)/дискурс (linguistic discourse)” in Ukr, while “*discours* (discussion/linguistic...)” in Fr.

We also conduct experiments on pipelined datasets, XSAMSum and XMediaSum (Appendix C). Experimental results show that, with a fine-tuned translator, *S-T* method outperforms best-reported systems on most tasks. Moreover, 2-Step does not show better performance than *S-T*, which can be because 2-Step systems are trained to only translate source summaries instead of comprehending source input text. The high performance of *S-T* emphasizes that cross-lingual summaries in those pipelined datasets do not rely on source input text, which can rather be a translation task. This confirms our motivation that the pipeline annotation protocol has important limitations.

5.2 Human Evaluation

To comprehensively understand CLS systems, we conduct human evaluations of the model outputs, as multi-dimensional assessment offers a more robust and holistic perspective (Zhong et al., 2022b).

Following previous work (Kryscinski et al., 2019; Fabbri et al., 2021), we evaluate generated summaries from the following dimensions: *Fluency* evaluates the quality of generated sentences,

Model		*DX*				*QX*			
		F.	*Coh.*	*Con.*	*R.*	*F.*	*Coh.*	*Con.*	*R.*
S-T	En2Zh	2.60	2.87	2.27	3.30	2.10	2.15	1.95	2.25
	En2Fr	3.23	4.43	3.37	2.50	2.85	3.65	1.60	1.35
	En2Ukr	3.90	3.57	3.20	3.20	3.30	3.25	2.90	3.00
2-S	En2Zh	2.90	3.00	2.50	3.30	2.40	2.45	2.20	2.45
	En2Fr	3.30	4.47	3.47	2.50	3.00	3.65	1.90	1.50
	En2Ukr	3.83	3.70	3.57	3.30	3.35	3.25	3.00	3.05

Table 7: *F.*, *Coh.*, *Con.* and *R.* are *Fluency*, *Coherence*, *Consistency* and *Relevance*. 2-S: 2-Step. Please note that the scores are not comparable between languages.

including grammar and whether it is natural; *Coherence* evaluates the collective quality of generated summaries; *Relevance* evaluates the importance of information in generated summaries; *Consistency* evaluates factual alignment between generated summaries and source input texts. We randomly extract 50 summaries from *S-T* and 2-Step outputs on ConvSumX for each language, and ask native speakers to give scores from 1 to 5. Higher scores indicate higher qualities.

The result is shown in Table 7. Generally, all metrics see low scores, suggesting the challenge of few-shot CLS. Both models see higher scores on *DX* compared with *QX*, which is consistent with our automatic evaluation. Compared with *S-T*, 2-Step achieves similar Relevance scores on all tasks. This is because the input source summary for both models is identical, thus the information in it is the same. However, 2-Step achieves higher Fluency, Coherence, and Consistency scores, which justifies our assumption that source input text information is critical, in particular for consistency.

We present case study of model outputs in Appendix D.

6 Related Work

CLS Corpora Existing CLS corpora construction can be categorized into two main protocols. 1) Pipeline annotation: translating summaries from MLS corpora into other languages and; 2) Automatic alignment: aligning summaries and input texts of different language versions.

Zhu et al. (2019) construct the first large-scale CLS dataset by automatically translating monolingual summaries using MT systems with a round-trip strategy and manual post-editing on test sets. Bai et al. (2021) construct an En2De dataset using the same method. Feng et al. (2022) automatically translate summaries from SAMSum (Gliwa et al., 2019) into Russian, De and Zh. Wang et al. (2022a) manually translate summaries from SAMSum (Gliwa et al., 2019) and MediaSum (Zhu et al., 2021) into De and Zh. Different from them, we propose a new annotation protocol, which helps annotators to comprehend documents quickly and accurately. To our knowledge, we are the first to address such human annotation issues for CLS research and present a new benchmark, ConvSumX.

A different line of work constructs CLS datasets by linking different language versions of online articles, such as Wikipedia (Perez-Beltrachini and Lapata, 2021) and WikiHow (Ladhak et al., 2020). Despite the cheap cost and large scale, there can be misalignment and hallucination problems. For example, Wikipedia articles and their leading paragraphs (pseudo summaries) of the same person in different languages can contain different contents. Also, such a method is limited to resources that contain multi-lingual data, which may not be available for all domains of interest, for example, the conversational text.

CLS Models Early work on CLS focuses on a pipeline paradigm by first summarizing, then translating, or vice versa. However, due to the poor performance of early MT and summarization systems, such methods can often suffer from error propagation. With the advance of deep learning and PLM technologies, recent work deploys end-to-end methods. Zhu et al. (2019), Xu et al. (2020), Bai et al. (2021) and Wang et al. (2022a) propose multi-task learning or pre-training on large in-domain CLS, mono-lingual summarization and translation data. Different from them, we propose a 2-Step method under the same concept of sourcing from source input text with the guidance of source summary, which is free of pre-training on large and thus can be easily adapted to other tasks and languages.

7 Conclusion

We conducted data analysis on 3 typical corpora and showed that the pipeline annotation protocol suffers from errors from both the summarization and translation processes. To address these issues, we proposed that cross-lingual summaries should be sourced from source input text. Based on this principle, we annotated a more faithful CLS benchmark, ConvSumX by relying on both source-language texts and summaries. Based on the same intuition, we proposed a 2-Step method that takes both source text and source summaries as input. Experimental results showed that 2-Step method outperforms strong baselines on ConvSumX, demonstrating that both source-language texts and summaries are crucial in modeling cross-lingual summaries and are complementary to each other. To our knowledge, we are the first to show that summary translation has limitations for CLS, giving a more faithful solution.

Limitations

The limitation of this paper can be stated from three perspectives. First, although using our CLS annotation protocol can label more faithful data, the annotation cost is higher because annotators need to comprehend the full source text instead of only the source summary. Second, ConvSumX only covers 3 typical languages, while languages from different language families and have different morphology and lexical-/syntactic rules require further investigation. Third, although the proposed 2-Step method is effective, we simply concatenate the source input text and mono-lingual summary at the token level as the model input but do not make further exploration. We believe that more smart and sophisticated designs to integrate features from source input text and mono-lingual summary can further improve the CLS performance, which, however, we leave for future work.

Ethics Statement

Data Usage and License ConvSumX is based on two public English conversation summarization datasets, namely DIALOGSUM and QMSum. Both datasets are freely available online under the MIT license, which has no constraint to academic use, modification, and further distribution. We will follow the MIT license to make our data (annotated target summaries/queries and corrected English summaries/queries) freely available online.

Human Annotation The construction of ConvSumX involves human annotation. We hire 4 expert translators as our annotators and editors for each target language. The total cost is around $6,500$ USD, which applies to our annotation (including quiz annotation) and review. The hourly salary is equal. The total annotation time (including training annotation and editing) for Zh, Fr and Ukr is around 96, 96, and 120 hours (according to our annotation cost/hourly salary). Detailed information about our annotators/judges/editors can be found in Appendix A.

Content Safety During our annotation, annotators are explicitly asked to not involve any personal/violent information and to write summaries strictly limited to the scope of source input text. Also, if any violent or uncomfortable information is found in source input text, annotators are asked to report such issues. All data are further reviewed by editors. With careful checking and evaluation, ConvSumX (including source input text) contains no personal/violent content, and is safe to use.

Acknowledgement

We thank reviewers from ACL2023 for their suggestions. We extend our sincere and special thanks to our meta-reviewers for their indispensable and exceptional contributions. We also appreciate Ruochen Xu for insightful discussion and expert translators from Lan-bridge who have played a crucial role in the development of ConvSumX. This work is funded by the Ministry of Science and Technology of China (grant No. 2022YFE0204900) and National Natural Science Foundation of China (grant NSFC No. 62161160339).

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.

Svitlana Budzhak-Jones. 1998. Against word-internal codeswitching: Evidence from ukrainian-english bilingualism. *International Journal of Bilingualism*, 2(2):161–182.

Hsuan-Chih Chen and Man-Lai Ng. 1989. Semantic facilitation and translation priming effects in chinese-english bilinguals. *Memory & cognition*, 17(4):454–462.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. 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, Ruochen Xu, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Yue Zhang. 2022a. Unisumm: Unified few-shot summarization with multi-task pre-training and prefix-tuning. *arXiv preprint arXiv:2211.09783*.

Yulong Chen, Ming Zhong, Xuefeng Bai, Naihao Deng, Jing Li, Xianchao Zhu, and Yue Zhang. 2022b. The cross-lingual conversation summarization challenge. In *Proceedings of the 15th International Conference on Natural Language Generation: Generation Challenges*, pages 12–18, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. 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.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22(107):1–48.

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

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.

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.

Kim Kirsner, Marilyn C Smith, RS Lockhart, ML King, and M Jain. 1984. The bilingual lexicon: Language-specific units in an integrated network. *Journal of verbal learning and verbal behavior*, 23(4):519–539.

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.

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.

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.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. 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 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Yixin Liu, Alexander R Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, et al. 2022. Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation. *arXiv preprint arXiv:2212.07981*.

Sébastien Pacton and S Hélène Deacon. 2008. The timing and mechanisms of children's use of morphological information in spelling: A review of evidence from english and french. *Cognitive Development*, 23(3):339–359.

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. *J. Mach. Learn. Res.*, 21:140:1–140:67.

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.

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

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

Peng Xu, Hamidreza Saghir, Jin Sung Kang, Teng Long, Avishek Joey Bose, Yanshuai Cao, and Jackie Chi Kit Cheung. 2019. A cross-domain transferable neural coherence model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 678–687, Florence, Italy. 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. 2022a. Dialoglm: Pre-trained model for long dialogue understanding and summarization. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 11765–11773. AAAI Press.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022b. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 2023–2038. Association for Computational Linguistics.

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

For human evaluation in § 2, we invite 2 expert translators as judges to conduct human evaluation and analysis of existing CLS corpora. For cross-lingual summary annotation and mono-lingual correction (§ 3), we invite 3 translators as our annotators and 1 as an editor to evaluate human annotation and model outputs (§ 5.2) for each language direction. Additionally, we invite one senior translator as the project manager to monitor the whole annotation progress.

All expert translators are from Lan-bridge, a qualified institution for translation service[9], recognized by the ISO[10]. All annotators, editors and judges are native in the target language (i.e., Chinese, French or Ukrainian), and professional in English. They are competent in translation, linguistic research and related information processing. They also have a good understanding of the textual background of certain culture, technology and domain. Our annotators and editors either got graduate certificates in translation major or got graduate certificates in other fields but have more than 2 years of full-time professional experience in translating. Besides the above requirements, the manager has more than 5-year experience in multi-lingual translation projects that cover the language directions as described in this paper.

B Analysis and Cases of Translation Errors

As shown in Table 9 and Table 10, we present cases of each error type as discussed in § 2.2.

In Table 9, "*Sheen*" refers to an actor, while annotators translate it into "高光/*Highlight*", and the term "*queer group*" into "同性恋群体/*gay group*". Although "*queer*" has a meaning of "*gay*", the proper translation should be "酷儿群体". Also, in the Coreference case, "*the date*" refers to the day when "*go do groceries together*", which is incorrectly translated into "聚会的日期/*the date of party*". In Table 10 Sentence Relation, annotators confuse the meaning and relation between two sentences, and the translation is completely incorrect at the sentence semantic level.

Model	***XSAMSum***							
	En2Zh				***En2De***			
	R1	R2	RL	BS	R1	R2	RL	BS
Summ-Trans*	42.1	17.2	35.1	**77.6**	**48.0**	**23.1**	**40.3**	**76.3**
Trans**-Summ*	40.0	14.9	32.6	76.6	43.5	17.8	35.1	74.1
mBART$_{\text{E2E}}$	39.6	15.5	32.9	76.6	43.4	17.6	35.9	74.0
mDBART$_{\text{E2E}}$	-	-	-	-	-	-	-	-
S-T*†	36.0	12.3	29.2	74.1	43.3	16.5	34.5	73.4
S-T†	**43.8**	**18.7**	**35.9**	**77.6**	46.2	20.0	37.6	75.1
2-Step†	43.5	**18.7**	35.8	**77.6**	46.2	20.2	37.6	75.1

Model	***XMediaSum40K***							
	En2Zh				***En2De***			
	R1	R2	RL	BS	R1	R2	RL	BS
Summ-Trans*	24.8	8.6	22.2	66.8	23.9	9.9	21.2	62.0
Trans**-Summ*	24.1	8.2	21.4	65.9	20.9	8.2	18.5	60.4
mBART$_{\text{E2E}}$	23.8	7.8	21.0	66.0	20.6	7.7	18.2	60.4
mDBART$_{\text{E2E}}$	28.7	11.1	25.7	68.4	26.7	**12.1**	**24.0**	**63.8**
S-T*†	24.2	6.7	20.1	65.8	24.1	8.8	21.0	61.2
S-T†	29.6	11.1	25.9	**68.5**	27.3	11.7	**24.0**	63.6
2-Step†	**29.7**	**11.2**	**26.0**	**68.5**	**27.4**	11.8	**24.0**	63.6

Table 8: Experimental results on CLIDSUM dataset. We show the best-reported pipeline and End2End method from (Wang et al., 2022a). †: our results. *: translator that is not fine-tuned. For a fair comparison, we use BART-large (Lewis et al., 2020) for mono-lingual summarization and mBART as for cross-lingual generation. mDBART$_{\text{E2E}}$: mDialBART$_{E2E}$.

All those translation cases together with summarization case (Figure 1) suggest the pipeline annotation can contain a large number of errors.

C Experiment on Pipelined Datasets

We conduct experiments on two pipelined datasets, namely XSAMSum and XMediaSum from the CLIDSUM benchmark and compare our pipeline and 2-Step methods with best-reported systems from (Wang et al., 2022a):

***Summ-Trans* Pipeline** They use BART(D_{all}) for mono-lingual summarization (Feng et al., 2021), and Google Translate [11] for summary translation.

***Trans-Summ* Pipeline** They use Google Translate to first generate cross-lingual dialogues, and then use mBART-50 to generate target language summaries.

mBART$_{E2E}$ They directly fine-tune an mBART-50 on cross-lingual $\langle S^{src}, S^{tgt}\rangle$ pairs, which is also an E2E baseline in our § 4.

mDialBART$_{E2E}$ They further pre-train an mBART-50 model using multiple tasks, including action filling, utterance permutation, mono-lingual

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

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

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

summarization and machine translation, on MediaSum (Zhu et al., 2021).

For more fair comparison, we fine-tune BART-large on corresponding mono-lingual data for mono-lingual summarization and fine-tune mBART-50 for translation and our 2-Step model. The result is shown in Table 8.

We see that without fine-tuning, our pipeline method (*S-T**) shows low performance. However, equipped with fine-tuned mBART as translator, *S-T* outperforms all previous methods on all tasks (e.g., *S-T* outperforms the best-reported mDialBART on En2Zh XMediaSum by 1 R1) except for the *S-T* pipeline on En2De XSAMSum, which can be because that Google Translate is better than mBART-50 at En2De translation. However, our 2-Step method, which is explicitly presented with source dialogue information and outperforms *S-T* on *ConvSum*, only shows comparable or even worse results compared with *S-T* on XSAMSum and XMediaSum. The contradiction of such model performance on CLIDSUM can be explained by that such pipelined datasets focus more on how to directly translate the mono-lingual summary, while adding more source dialogue information is less useful and sometimes harmful.

D Case Study

To qualitatively demonstrate the advantage of 2-Step method, we present cases from *S-T* and 2-Step on ConvSumX. Figure 3 shows a sample from DialogSumX and Figure 4 shows another sample from QMSumX.

As shown in Figure 3, both summaries generated by 2-Step and *S-T* contain errors ("*donson*", which should be "*dawson*"). However, compared with *S-T* ("下台", which means "*step down*" or "*lose power*"), 2-Step method can improve the faithfulness of summaries ("发了一份备忘录, which means "*send a memo*"). Similarly, as shown in Figure 4, *S-T* method suffers from generating unnatural language(e.g., a string of *d*-s in En2Zh case) and it has trouble generating some not-commonly used words (e.g., the counterpart word of *cepstra* in 3 target languages), while 2-Step method can significantly ameliorate such problems.

Moreover, we also find that 2-Step method not only focus on "translating" summaries from source language to target language, but also rewriting them by referring to the original input text. In the En2Zh example in Figure 4, 2-Step method properly generates "告诉" (which has a sense of "*tell*") for the word "*inform*", which is more natural considering its textual background. In contrast, *S-T* method simply translates it into "通知", which is more of the sense "*announce/notify*", and is not natural in Mandarin.

These cases indicate that source-language texts are essential in cross-lingual summarization tasks, which further demonstrates our conclusion.

Word Sense	
Input Text	BLITZER: T̈wo and a Half Menm̈inus one. Charlie Sheen has been fired. Warner Brothers Television terminated Sheen's contract with the sitcom hit today. CNN's Jeanne Moos has more now on the Sheen saga and the backlash. JEANNE MOOS, CNN CORRESPONDENT (voice-over): It's the Sheening of America. CHARLIE SHEEN, ACTOR: Welcome to Sheen's Corner. MOOS: He's on every corner. BILL HADER, CAST MEMBER, NBC'S S̈ATURDAY NIGHT LIVE:̈ Live from New York, it's Saturday night. MOOS: Sirius Radio devoted an entire channel to him for a day. ANNOUNCER: Tiger Blood Radio. MOOS: Spike TV will feature Sheen's greatest antics in Taiwanese animation. He's even alienated witches for misusing the word ẅarlock.ÜNIDENTIFIED MALE: We bind you. UNIDENTIFIED FEMALE: We bind you. MOOS: So a couple of witches in Massachusetts performed a magical intervention. UNIDENTIFIED FEMALE: We need to come and cleanse your house. MOOS: But Sheen's very own Web casts are what tipped the scale. SHEEN: The tag line is ẗorpedoes of truth.M̈OOS (on camera): Well, how's this for a torpedo of truth? It seems the shine has come off Charlie Sheen. (voice-over) In one Web cast he showed off a tattoo on his wrist of his slogan ẅinning,änd said hi to his kids. SHEEN: Daddy loves you. And if you're watching, tell Mom to leave the room. It's on. MOOS: One of his goddesses perched on his lap. Sheen was literally playing with fire as viewers wait for him to combust. SHEEN: It's kind of an eerie image. I'm burning my own face, but I can't feel the MOOS: As one poster on TMZ put it, P̈arents, make your kids watch this. If that doesn't scare them away from drugs, nothing will.(̈on camera) You know the joke has become a little too sick when a comedian refuses to tell any more jokes about Charlie Sheen. (voice-over) Craig Ferguson spoke of how the English insane asylum named Bedlam provided entertainment back in the 18th Century. CRAIG FERGUSON, TALK SHOW HOST: They would pay a penny, and they would look through the peepholes of the cells, and they would look at the lunatics. And looking at the Charlie Sheen thing unfold, and I'm thinking oh, man. MOOS: But Ferguson wasn't kidding. No more Charlie Sheen jokes. Sheen himself has become a verb. The creators of S̈outh Parküsed it to describe the state they got themselves in when they once dressed in drag for the Oscars. UNIDENTIFIED MALE: We were just Sheening our heads off. MOOS: From our couches, we judge who does the best Sheen. Is it S̈NL?̈ HADER: Sorry, middle America. Losers, winning, bye-bye. MOOS: Or Jimmy Fallon? JIMMY FALLON, TALK SHOW HOST: Winning, Adonis DNA. I'm a bitching rock star, blood of a tiger. I'm like Zeus in a Speedo. MOOS: But something stinks when we don't know if it's OK to laugh and winning is a losing proposition. FRED ARMISEN, CAST MEMBER, NBC'S S̈ATURDAY NIGHT LIVE:̈ Winning! HADER: Winning. MILEY CYRUS, SINGER/ACTRESS: Winning. HADER: Winning. MOOS: Jeanne Moos, CNN... SHEEN: Winning, winning, winning! UNIDENTIFIED MALE: Winning, winning, winning! UNIDENTIFIED MALE: Winning, winning, winning! SHEEN: ... New York. BLITZER: Thanks, Jeanne. Thanks very much for watching. I'm Wolf Blitzer in THE SITUATION ROOM. J̈OHN KING USAs̈tarts right now.
En Summary	Sheen Fired from Hit Sitcom
Zh Summary	热门情景喜剧的高光时刻 (Hightlight Moment of Hit Sitcom)
Terminology	
Input Text	Mika: I wanted to ask you to stop supporting the queer group Ann: why? I think they do great things Mika: they discriminated Molly horribly Ann: why? how? Mika: they refused to include her in the panel about sexuality Tom: did they give a reason? Mika: they said her research doesn't match the topic of the panel, what is a bullshit Mika: I believe it's only because she is straight Tom: hmm...
En summary	The queer group discriminated Molly - they refused to include her in the panel about sexuality.
Zh summary	同性恋团体歧视莫莉——他们拒绝让她参加关于性的小组讨论。 (The gay group discriminated Molly - they refused to include her in the panel about sexuality.)
Coreference	
Input Text	Wanda: Let's make a party! Gina: Why? Wanda: beacuse. I want some fun! Gina: ok, what do u need? Wanda: 1st I need too make a list Gina: noted and then? Wanda: well, could u take yours father car and go do groceries with me? Gina: don't know if he'll agree Wanda: I know, but u can ask :) Gina: I'll try but theres no promisess Wanda: I know, u r the best! Gina: When u wanna go Wanda: Friday? Gina: ok, I'll ask"
En summary	Wanda wants to throw a party. She asks Gina to borrow her father's car and go do groceries together. They set the date for Friday.
Zh summary	旺达想办个聚会。她问吉娜借她父亲的车，两人一起去买聚会用的东西。他们把聚会的日期定在了周五。 (Wanda wants to throw a party. She asks Gina to borrow her father's car and go do groceries together. They set the date of party for Friday.)

Table 9: Error case (a).

Sentence Relation	
Input Text	BLITZER: WOLF BLITZER, CNN ANCHOR: Happening now, neck and neck in Indiana. New evidence Barack Obama and Hillary Clinton are in for another fierce battle. Meantime, Obama is dealing with a familiar distraction, the words of his former pastor. We'll tell you what's going on today. John McCain makes a provocative claim about Barack Obama. The Republican suggests Obama is the candidate of the Islamic militant group Hamas. We're looking into this story right now. And President Bush wants to show you the money. We're going to tell you what he's telling taxpayers and why he hopes it will send them to the stores. I'm Wolf Blitzer. You're in THE SITUATION ROOM. Barack Obama wanted to talk to Indiana voters about the soaring gas prices that make their lives tougher every single day, but today the Democrat found he couldn't ignore an ongoing source of controversy. That would be his former pastor, the Reverend Jeremiah Wright. After clamming up and lowering his profile, Wright is now speaking out publicly about the impact on – and it's having an impact, potentially, at least, on the Obama campaign. Let's go right to CNN's Jessica Yellin. She's watching the story for us. It's a familiar problem over these past few weeks for the senator, Jessica. JESSICA YELLIN, CNN CONGRESSIONAL CORRESPONDENT: It really has been, Wolf. Today it seems Barack Obama was trying yet again to put that Reverend Wright controversy behind him. He fielded a question about the latest statement from his former pastor. SEN. BARACK OBAMA (D-IL), PRESIDENTIAL CANDIDATE: I understand that he might not agree with me on my assessment of his comments. That's to be expected. So, you know, he is obviously free to express his opinions on these issues. You know, I've expressed mine very clearly. I think that what he said in several instances were objectionable. And I understand why the American people took offense. And, you know, and as I indicated before, I took offense. YELLIN (voice-over): Barack Obama speaking out on new comments by his former pastor. REV. JEREMIAH WRIGHT, BARACK OBAMA'S FMR. PASTOR: And put constantly other and over again... YELLIN: The Reverend Jeremiah Wright, in an interview airing on PBS Friday night, stands by past sermons that became a political firestorm. WRIGHT: ... controlled by rich white people. YELLIN: Wright said his words regarding the 9/11 attacks and race relations were taken out of context. He also reacts to Obama's criticism of him. WRIGHT: He's a politician. I'm a pastor. We speak to two different audiences. And he says what he has to say as a politician. I say what I have to say as a pastor. Those are two different worlds. I do what ...
En Summary	Obama's Ex-Pastor Reacts to Criticism; McCain: Obama Favored by Hamas
Zh Summary	奥巴马前总统回应批评麦凯恩：奥巴马受哈马斯青睐 (The Ex-President Obama Responds to Criticism of McCain: Obama is Favored by Hamas)
Others	
Input Text	Elliot: i can't talk rn, i'm rly busy Elliot: can i call u back in about 2 hours? Jordan: Not really, I'm going to a funeral. Jordan: I'll call you tonight, ok? Elliot: sure Elliot: whose funeral is it? Jordan: My colleague's, Brad. Jordan: I told you about him, he had a liver cancer. Elliot: i'm so sorry man, i hope u're ok Elliot: i'll call u at 8 pm
En summary	Elliot can't talk to Jordan now, he's busy. He'll call him back at 8 pm. Jordan is going to Brad's funeral. He had liver cancer.
Zh summary	艾略特现在没空和乔丹说话，他很忙。晚上6点他会给乔丹回电话。乔丹要去参加布拉德的葬礼，布拉德因肝癌去世了。 (Elliot can't talk to Jordan now, he's busy. He'll call him back at 8 pm. Jordan is going to Brad's funeral. He died of liver cancer.)

Table 10: Error Case (b).

DialogSumX		
Input Text		#Person1#: Ms. Dawson, I need you to take a dictation for me. #Person2#: Yes, sir... #Person1#: This should go out as an intra-office memorandum to all employees by this afternoon. Are you ready? #Person2#: Yes, sir. Go ahead. #Person1#: Attention all staff... Effective immediately, all office communications are restricted to email correspondence and official memos. The use of Instant Message programs by employees during working hours is strictly prohibited. #Person2#: Sir, does this apply to intra-office communications only? Or will it also restrict external communications? #Person1#: It should apply to all communications, not only in this office between employees, but also any outside communications. #Person2#: But sir, many employees use Instant Messaging to communicate with their clients. #Person1#: They will just have to change their communication methods. I don't want any - one using Instant Messaging in this office. It wastes too much time! Now, please continue with the memo. Where were we? #Person2#: This applies to internal and external communications. #Person1#: Yes. Any employee who persists in using Instant Messaging will first receive a warning and be placed on probation. At second offense, the employee will face termination. Any questions regarding this new policy may be directed to department heads. #Person2#: Is that all? #Person1#: Yes. Please get this memo typed up and distributed to all employees before 4 pm.
UniSumm		#person1# asks ms. dawson to take a dictation. #person1# tells ms. dawson that all office communications are restricted to email correspondence and official memos, and it applies to internal and external communications.
S-T	Zh	#person1#让donson小姐下台(*asks miss donson to step down*)。#person1#告诉donson小姐(*miss donson*),所有办公室的通讯只限于电子邮件和官方备忘录,这适用于内部和外部的通讯。
	Fr	#person1# demande à mme dawson de prendre une dictation. #person1# dit à mme dawson que toutes les communications de bureau sont limitées à la correspondance électronique et aux notes (*note*) officielles, et cela s'applique aux communications internes et externes.
	Ukr	#person1# просить місіс донсон зробити проказ (*asks mrs donson to make leprosy*). #person1# каже місіс донсон, що всі офісні зв'язки обмежені електронною korespondenцією та офіційними мемо'ю, і це стосується внутрішніх та зовнішніх зв'язків.
2-Step	Zh	#person1#让donson夫人发了一份备忘录 (*asks ms donson to send a memo*)。#person1#告诉donson夫人(*ms donson*),所有办公室的通讯只限于电子邮件和官方备忘录,这适用于内部和外部的通讯。
	Fr	#person1# demande à mme dawson de prendre une dictation. #person1# dit à mme dawson que toutes les communications du bureau sont limitées à la correspondance électronique et aux mémorandums (*memoranda*) officiels, et cela s'applique aux communications internes et externes.
	Ukr	#person1# просить місіс Дорсон зробити диктацію (*asks mrs donson to dictate*). #person1# розповідає місіс Дорсон, що всі офісні комунікації обмежені електронною korespondenцією та офіційними меморандумами, і це стосується внутрішніх і зовнішніх комунікцій.
Gold	En	Ms. Dawson takes a dictation for #Person1# about prohibiting the use of Instant Message programs in the office. They argue about its reasonability but #Person1# still insists.
	Zh	Dawson女士为#Person1#做了一份关于禁止在办公室使用即时消息程序的口述记录。他们争论了这道命令的合理性，但#Person1#仍然坚持这样。
	Fr	Mme Dawson prend des notes pour #Person1# concernant l‘interdiction d’utiliser des programmes de messagerie instantanée au bureau. Ils (elles) disctutent de si cela est raisonnable mais #Person1# continue d'insister.
	Ukr	Пані Доусон розпоряджається #Person1# заборонити використання програм обміну миттєвими повідомленнями в офісі. Вони сперечаються щодо її доцільності, але #Person1# все ще наполягає.

Figure 3: Case (a): cross-lingual summaries generated by *S-T* and 2-Step method, and human annotated summaries. We show their differences (not all errors) in red and their English translation in bracketed *italics*.

QMSumX		
Input Text		phd d: Yeah. Uh, yeah. So there is this. And maybe we well we find some people so that uh, agree to maybe work with us, and they have implementation of VTS techniques so it's um, Vector Taylor Series that are used to mmm, uh f to model the transformation between clean cepstra and noisy cepstra. So. Well, if you take the standard model of channel plus noise, uh, it's it's a nonlinear eh uh, transformation in the cepstral domain. professor c: Mm - hmm. …
Query	Zh	该团队是如何评价关于格林纳达（Grenada）的人的？
	Fr	Qu'a dit l'équipe à propos de la personne à Grenade ?
	Ukr	Що сказала команда про особу в Гренаді?
Uni	Zh	phd d informed the group that someone in grenada was working on vts techniques. the vts technique was to model the transformation between clean cepstra and noisy cepstra. the team thought that it could get rid of the noise and the channel effect. the idea was originally proposed by the cmu.
	Fr	there is a person working in grenada working on vts techniques. vts is a nonlinear transformation between clean cepstra and noisy cepstra. it can get rid of the noise and the channel effect. the team was trying to find someone who would implement the vts technique.
	Ukr	phd d informed the team that they were looking for people who would work with them on implementing vts techniques. vts is a nonlinear transformation between clean cepstra and noisy cepstra. this could get rid of the noise and the channel effect. the team name was jose carlos segura.
S-T	Zh	博士(dddddddddddd通知了团队)*(Ph D dddddddddddd announced/notified the team)*,格纳da*(genada)*的某人正在开发vts技术。vts技术是用于模拟 cleancpstra和噪音cpstra之间的转换*(simulate the transformation between clean cpstra and noisy cpstra)*。团队认为,该技术可以消除噪音和频道效应。这个想法最初由cmu提出。
	Fr	il y a une personne qui travaille à Grenade et qui travaille sur les techniques vts. vts est une transformation non linéaire entre une céphalote propre et une céphalote bruise. *(vts is a nonlinear transformation between a clean cephalote and a bruise cephalote.)* il peut se débarrasser du bruit et de l'effet de chaîne. l'équipe essayait de trouver quelqu'un qui appliquerait la technique vts.
	Ukr	доктор наук повідомив команді, що вони шукають людей, які працюватимуть з ними над запровадою методів vts. vts - це неlineарна трансформація між чистою та шумною кавстрою. *(cavstra)* це могло б позбутися шуму та ефекту каналів. команда назвалася jose carlos segura.
2-Step	Zh	博士(d)告诉了该团队(*PhD d informed the team)*,格林纳达*(grenada)*有人正在研究Vts技术。该技术是将清洁 cepstra和噪音感知的 cepstra转换成模型*(turn clean cepstra and noisy cepstra into model)*。该团队认为该方法可以消除噪音和频道效应。这个想法最初提出的是cmu。
	Fr	il y a une personne travaillant à Grenade qui travaille sur les techniques de vts. vts est une transformation non linéaire entre un cepstra propre et un cepstra émetteur de bruit. *(vts is a nonlinear transformation between a clean cepstra and a noise-emitting cepstra.)*il peut se débarrasser du bruit et du effet de chaîne. l'équipe a essayé de trouver quelqu'un qui appliquerait la technique de vts.
	Ukr	доктор філософії вповідав команді, що вони шукають людей, які будуть працювати з ними над реалізаціюм методів vts. vts - це неlineарна трансформація між чистою серстрою та шумною серстрою*(cepstra)* . ця команда могла позбутися шуму та вплива каналівкоманда назвалася jose carlos segura.
Gold	En	PhD D brought up a VTS technique to do voice-unvoice which was developed by Jose Carlos Segura, who is a person from Grenada. The professor did not know him. PhD C added that the inspiration for the VTS came from CMU.
	Zh	博士生D提出了一种VTS技术来识别清音和浊音，这是由来自格林纳达的Jose Carlos Segura开发的。教授不认识他。博士生C补充道VTS的灵感来自卡内基梅隆大学（CMU）。
	Fr	Doctorat D a évoqué une technique VTS pour faire du vocal-non vocal qui a été développée par Jose Carlos Segura, originaire de Grenade. Le professeur ne le connaissait pas. Doctorat C a ajouté que l'inspiration pour le VTS venait de la CMU.
	Ukr	Доктор філософії D виніс на обговорення техніку визначення голос-нема голосу VTS, яку розробив Хосе Карлос Сегура, людина з Гренади. Професор його не знав. але натхнення для VTS надійшло від CMU. Доктор філософії C додав, що натхнення для VTS надійшло від CMU.

Figure 4: Case (b): cross-lingual summaries generated by *S-T* and 2-Step method, and human annotated summaries. We show their differences (not all errors) in red and their English translation in bracketed *italics*.