CroCoSum: A Benchmark Dataset for Cross-Lingual Code-Switched Summarization

Ruochen Zhang[1], **Carsten Eickhoff**[1,2]
[1]Brown University, [2]University of Tübingen
Dep. of Computer Science, School of Medicine
ruochen_zhang@brown.edu, c.eickhoff@acm.org

Abstract

Cross-lingual summarization (CLS) has attracted increasing interest in recent years due to the availability of large-scale web-mined datasets and the advancements of multilingual language models. However, given the rareness of naturally occurring CLS resources, the majority of datasets are forced to rely on translation which can contain overly literal artifacts. This restricts our ability to observe naturally occurring CLS pairs that capture organic diction, including instances of code-switching. This alteration between languages in mid-message is a common phenomenon in multilingual settings yet has been largely overlooked in cross-lingual contexts due to data scarcity. To address this gap, we introduce CroCoSum, a dataset of cross-lingual code-switched summarization of technology news. It consists of over 24,000 English source articles and 18,000 human-written Chinese news summaries, with more than 92% of the summaries containing code-switched phrases. For reference, we evaluate the performance of existing approaches including pipeline, end-to-end, and zero-shot methods. We show that leveraging existing CLS resources as a pretraining step does not improve performance on CroCoSum, indicating the limited generalizability of current datasets. Finally, we discuss the challenges of evaluating cross-lingual summarizers on code-switched generation through qualitative error analyses. Our collection and code can be accessed at `[anonymous]`.

Keywords: Code-Switching, Cross-Lingual, Summarization, Corpus, Collection, Dataset

1. Introduction

Cross-lingual summarization (CLS) is the task of producing summaries in a target language given source documents in a different language. CLS can help with the rapid dissemination of information across multiple languages in an increasingly globalized context. It is considered more challenging than within-language summarization, as it combines translation and summarization objectives (Wang et al., 2022b). With more multilingual resources (Raffel et al., 2019; Laurençon et al., 2022; Scialom et al., 2020; Hasan et al., 2021) becoming available and the advancement of large multilingual language models (Liu et al., 2020; Lin et al., 2021; Scao et al., 2022), CLS has attracted significant attention in recent years. One key factor that has been limiting the development of CLS is data scarcity. Therefore, current CLS resources (Wang et al., 2022; Zheng et al., 2022; Ladhak et al., 2020; Hasan et al., 2021; Perez-Beltrachini and Lapata, 2021) heavily rely on automatic or manual translation rather than collecting texts organically written in cross-lingual fashion. However, translated texts have been reported to exhibit features different from the original language's composition (Graham et al., 2020). Summaries generated by models trained on such texts may contain instances of "Translationese", such as literal translations of idioms (Wang et al., 2022a). Humans, on the other hand, code-switch between languages, especially when there is no appropriate translation, or when readers are more familiar with the original foreign entity names or expressions. There have been summarization resources addressing the code-switching phenomenon (Mehnaz et al., 2021), but they focus on summarizing from already code-switched source texts.

To study the phenomenon of code-switching in CLS, we introduce CroCoSum, a new benchmark dataset for **Cro**ss-Lingual **Co**de-Switched **Sum**marization containing human-written Chinese-English code-switched summaries of technology-focused news articles in English. The code-switched summaries are gathered from `solidot.org`, an online platform for sharing technology-related news. The summaries are written and posted by real users including technical professionals, open-source enthusiasts and university students. They are then reviewed by the website's editors before being published. Each post contains one or more links pointing to the original news sources. We collect the original news articles from the Internet Archive[1] and only consider those sources as written in English. We then construct the source-target pairs by tracing back to posts referring to those English source articles. Our final dataset contains over 24,000 English source articles and over 18,000 code-switched summaries. More than 92% of the summaries contain code-switched phrases and over 55% of sentences within the summaries contain code-switching spans. A data example is shown in Figure 1(a).

[1]`https://archive.org/`

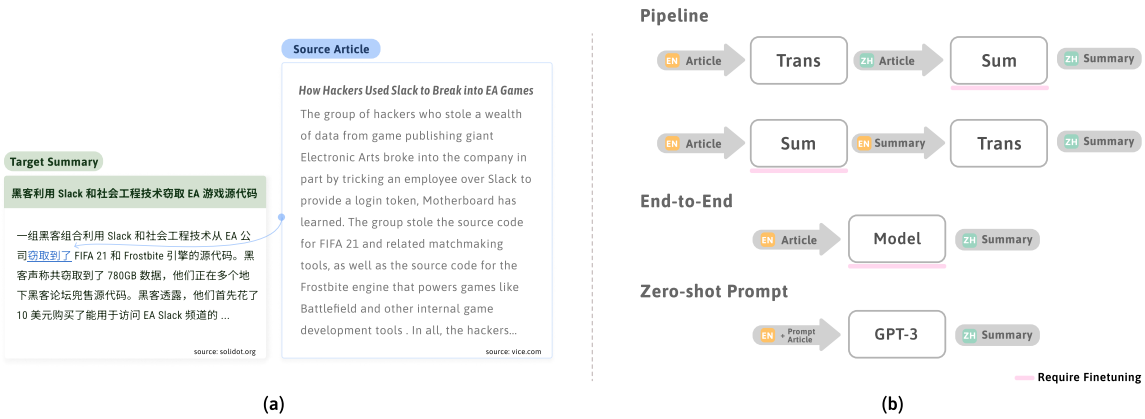

Figure 1: (a) A Data Example of Source Article and Target Summary Pair. (b) Baseline Approaches.

We follow existing CLS approaches (Zhu et al., 2019; Ladhak et al., 2020; Wang et al., 2022), and evaluate baselines including pipeline, end-to-end and zero-shot processing on CroCoSum (See Figure 1(b)). Pipeline methods can be further broken down into translate-then-summarize and summarize-then-translate approaches. We use the Google Translate API as the translation module in the pipeline methods. For the summarizer module and the end-to-end method, we experiment with various pretrained multilingual sequence-to-sequence models such as mT5 (Xue et al., 2020), mBART (Liu et al., 2020) and mBART-50 (Tang et al., 2020). We also prompt GPT-3 (Brown et al., 2020) to generate summaries in a zero-shot manner. Among our baselines, end-to-end finetuning mBART-50 yields the best results. However, we notice a decrease in performance when leveraging other CLS resources as a pretraining step in our best baseline, indicating limited generalizability provided by current CLS resources. Finally, by comparing various automatic metrics, we observe no clear relationship between summarization quality and code-switching complexity, calling for future research designing a more comprehensive evaluation framework.

The novel contributions of this work are threefold: 1) we introduce CroCoSum, the first collection designed for examining the phenomenon of code-switching in cross-lingual summarization, 2) we provide an initial set of benchmark performance measurements of various baseline approaches and architectures (pipeline, end-to-end and zero-shot prompting), 3) we perform a qualitative analysis revealing the common error types in code-switched generation and highlighting opportunities for future investigation.

2. Related Work

Cross-Lingual Summarization Early cross-lingual summarization resources like En2ZhSum and Zh2EnSum (Zhu et al., 2019) have been constructed via machine translation from originally monolingual summarization datasets. Collections such as ClidSum (Wang et al., 2022) crowdsource human translations to obtain cross-lingual resources of higher quality. Additionally, with the prevalence of large-scale web-mined texts, some works (Ladhak et al., 2020; Perez-Beltrachini and Lapata, 2021; Hasan et al., 2021) focus on the number of languages covered and exploit websites that provide multilingual content, such as WikiHow and Wikipedia. For CLS approaches, due to the limited availability of parallel corpora, early works (Wan et al., 2010; Zhang et al., 2016; Ayana et al., 2018; Zhu et al., 2019) use pipeline methods to break CLS into two subtasks: translation and summarization then develop dedicated models for each subtask. With the availability of large-scale parallel corpora and pretrained multilingual language models, more works (Ladhak et al., 2020; Hasan et al., 2021; Wang et al., 2022) experiment with end-to-end approaches with different pretraining techniques, allowing models to directly take source texts in one language and summarize them in another. However, all current CLS resources, whether translated or web-mined, may contain different levels of "Translationese" artifacts due to some dependency on machine-translated texts (Wang et al., 2022a; Ladhak et al., 2020). Kreutzer et al. (2022) mention that automatically crawled and filtered datasets tend to show a lower quality compared to hand-produced collections. Furthermore, unlike human-written summaries crafted for each article, using initial sentences or paragraphs as summaries in one language does not ensure alignment with the source texts sampled from articles in another language. Despite many efforts in resource collection, none of the existing collections acknowledge the phenomenon of code-switching, alternating languages in mid-message, in cross-lingual summaries. CroCoSum, therefore, is different from existing resources in that all

summaries are written and reviewed by humans to meet publishing standards. It provides an ideal testbed in which to observe organic human diction in CLS settings.

Code-switching Since code-switching is observed more frequently in colloquial (rather than formal) texts (Doğruöz et al., 2021; Winata et al., 2022), it is challenging to gather large-scale well-annotated resources to study this phenomenon (Yong et al., 2023). Established code-switching resources are usually collected from social media texts and focus on sequence tagging applications, including, for example, language identification (Das and Gambäck, 2014; Barman et al., 2014), NER (Singh et al., 2018), and POS-tagging (Aguilar et al., 2020). There have also been works that develop objective metrics to describe the level of code-switching complexity (Gambäck and Das, 2014, 2016; Khanuja et al., 2020a). Besides sequence tagging, other works also touch on short-form generation (Mondal et al., 2022) and speech recognition tasks with audio data (Li et al., 2012, 2022; Lovenia et al., 2022). Gupshup (Mehnaz et al., 2021), to the best of our knowledge, is the only collection dedicated to studying code-switching in summarization (Doğruöz et al., 2023). Different from CroCoSum, it introduces code-switched source texts by translating the SAMSum (Gliwa et al., 2019) dataset into Hinglish instead of studying this phenomenon in organically occurring target summaries.

3. CroCoSum

CroCoSum contains 18,557 human-written Chinese-English code-switched summaries and 24,171 English source articles. More than 92% of the summaries, and 55% of sentences in the summaries contain code-switched phrases. In the sections below, we describe our data collection and comparison with existing resources in detail.

3.1. Dataset Construction

The target summaries in CroCoSum are collected from `solidot.org`, an Chinese online platform for IT professionals and open-source enthusiasts to share technology-related news. Users summarize technology news from international outlets and compose short Chinese summary posts. Due to the highly timely nature of news and tech-focused topics, some English entities and phrases in the original news items are yet to receive formal translations or are preferred in their original form by website writers and readers. To guarantee a high-quality feed, prior to getting published on the platform, each post is encouraged to contain at least one hyperlink pointing to the original news source for credibility (See 1 for an example.) and is reviewed by human editors to ensure clarity. Our initial crawl contains 28,953 post webpages and 51,258 embedded links. We obtain web pages of the embedded links from the Internet Archive and use the newspaper3k[2] package to extract titles and articles. After running language detection[3] on the extracted content, we only retain English sources (80% of all sources). From the remaining websites, we filter out those that contain failed extractions (empty body, login information, javascript and cookie notifications, etc.) based on a manually curated list of cue words and sentences. We then trace back and remove posts containing these deleted websites and posts that do not contain any links, resulting in the final collection of 18,557 posts and 24,171 corresponding English source articles. We construct source-target pairs by matching summary posts with the embedded links to the source articles. If an article contains multiple sources, a list with all source texts is mapped to that article.

To examine the data quality obtained from the automatic filtering process, we recruit 3 annotators with bilingual proficiency and ask them to annotate a random sample of 20 source-target pairs following Perez-Beltrachini and Lapata (2021). Here, we depart from the original authors' annotation scheme of seeking binary answers to two general questions, and instead collect more fine-grained ratings of both the syntactic and semantic dimensions of the data instances. We adopt 4 rubrics suggested by Grusky et al. (2018), which are Fluency(F), Coherence(C), Informativeness(I) and Relevance(R). The pairs were rated using a 5-point Likert scale with 1 being the lowest score. In our annotations, the scores are F: 4.62, C:4.95, I: 3.97 and R: 4.18. Notably, 75% of our samples received a rating of 4 and 5 for the semantic dimensions (I and R), aligning with the quality statistics observed in Perez-Beltrachini and Lapata (2021). Finally, the collected data is partitioned into distinct training (70%), validation (15%) and test (15%) sets.

3.2. Dataset Characterization

CLS Dataset Statistics Table 1 shows a comparison of key measurements between CroCoSum and other existing CLS resources[4] containing English-Chinese source-target pairs. Note that, except En2ZhSum and CroCoSum, all other datasets con-

[2] `https://github.com/codelucas/newspaper`

[3] `https://github.com/Mimino666/langdetect`

[4] XSAMSum and XMediaSum40k are subsets of ClidSum (Wang et al., 2022) with CLS data. The remaining MediaSum424k subset only contains monolingual data.

Type	Dataset	Domain	Size	Source			Target		
				Lang.	Words	Sents	Lang.	Words	Sents
Translation	**En2ZhSum**	News	370,687	En	755.0	40.6	Zh	84.4	3.6
	XSAMSum	Dialogue	16,369	En	97.7	12.1	De/Zh	32.0	2.0
	XMediaSum40k	Dialogue	40,000	En	1661.5	113.5	De/Zh	28.04	1.2
Web-mined	**WikiLingua**	Guides	17,904	Multi	407.4	24.6	Multi	50.3	5.2
	CrossSum	News	4,975	Multi	673.5	33.7	Multi	44.8	1.21
Human-written	**CroCoSum**	News	18,557	En	1079.6	55.4	Zh	225.6	5.98

Table 1: Data statistics of CroCoSum and other CLS datasets. Except for En2ZhSum and CroCoSum, statistics are calculated based on the English-Chinese subsets.

Task	Dataset	Total Sents	Avg Sent Len	Switched		CMI		SP	
				Sents	%	All	Switched	All	Switched
Cross-Lingual									
Summ.	**En2ZhSum**	1,336,155	23.41	90,930	6.81	0.49	2.26	0.14	0.65
	XSAMSum	31,517	24.41	1,429	4.53	0.39	4.29	0.10	1.12
	XMediaSum40k	45,966	24.40	3,140	6.83	0.38	4.27	0.14	1.60
	WikiLingua	92,433	9.75	6,327	6.84	1.19	5.53	0.13	0.61
	CrossSum	6,027	36.96	2,099	34.83	2.10	4.97	0.94	2.23
Code-Switching									
Summ.	**GupShup**$^{\alpha}$	76,330	10.07	43,407	56.87	-	-	-	-
Tweet LID	**EMNLP2014**$^{\alpha}$	999	17.45	322	32.23	4.19	13.01	0.7	2.18
Speech Recog.	**ASCEND**$^{\beta}$	12,314	11.83	3,326	27.01	5.02	18.59	0.62	2.28
Speech Recog.	**SEAME** $^{\beta\gamma}$	11,852	12.69	6,468	54.57	14.11	25.86	1.84	3.37
Summ.	**CroCoSum**	110,534	37.88	61,678	55.75	4.74	5.09	2.18	2.35

$^{\alpha}$ Due to limited access to data, we report statistics of GupShup based on its original paper (Mehnaz et al., 2021) and EMNLP14 based on Gambäck and Das (2016). "-" means statistics cannot be computed from the original paper.
$^{\beta}$ We use transcriptions of the speech utterances, and remove noise tokens like <v_noise> and [UNK] prior to calculation.
$^{\gamma}$ The training split of SEAME is non-public, the statistics are reported on its dev splits.

Table 2: Code-switching metrics of CroCoSum and other CLS datasets.

tain multiple languages among either their sources or targets, but the statistics are only calculated on their English-Chinese subset for a more accurate comparison. Additionally, since CroCoSum could contain more than one source article per summary, we concatenate multiple sources for such examples before calculation. Besides general dataset descriptions like construction types, domains, dataset sizes and languages, we provide the average number of words and sentences (segmented by stanza[5]) in English source texts and Chinese target summaries. We observe that, compared with CLS datasets in the news domain (on average 600-700+ words per sample), CroCoSum contains much more expansive source texts (1,000+ words per sample). Summaries in CroCoSum are also much longer than those in other CLS datasets (225.6 words vs. 20-80 words on average per summary).

Code-Switching Complexity We also investigate the code-switching frequency of CroCoSum in comparison with other code-switching datasets. Because there is no existing code-switched resource for CLS, we extend our comparison to the loosely related summarization dataset GupShup (Mehnaz et al., 2021) that focuses on summarizing from Hindi-English code-switched source texts. For a more comprehensive analysis, we also select datasets that contain Chinese-English code-switched texts but across different tasks such as language identification (LID) in tweets (Solorio et al., 2014) and speech recognition (Lyu et al., 2010; Lovenia et al., 2022). Although current CLS resources do not study the code-switching phenomenon, they could potentially contain code-switched tokens in their target summaries. Therefore, we include them in our comparison as well. We use the metrics suggested by Gambäck and Das (2016) below to measure the level of code-switching complexity.

Code-Mixing Index (CMI) is the fraction of language-dependent tokens not belonging to the matrix language (the most frequent language in the sentence) in the utterance. CMI for a sentence x can be computed as

$$\text{CMI}(x) = \frac{(N(x) - max_{L_i \in L}\{t_{L_i}\}(x))}{N(x)}$$

[5] https://github.com/stanfordnlp/stanza

where $N(x)$ refers the number of language-independent tokens[6] in sentence x and t_{L_i} refer to tokens in language L_i. For monolingual sentences, CMI is 0. Higher CMIs indicate more code-switched tokens.

Intra-Sentence Switch Points (SP) are the number of word boundaries within a sentence for which the words on either side are in different languages.

Both metrics measure sentence-level switching. We report dataset-level statistics in Table 2 by taking an average of sentences across all examples (All) and in those with code-switched words (Switched). Additionally, we provide the total number of sentences (Total Sents), average sentence length (Avg Sent Len), the number of code-switched sentences (Switched-Sents) and their percentage (Switched-%). Notice that among all summarization datasets, cross-lingual ones report metrics based on their target summaries, while GupShup bases them on its source documents.

We observe that CroCoSum offers the highest percentage of code-switched sentences, overall CMI, and SP among all CLS datasets. To our surprise, CLS datasets such as WikiLingua contain high CMI for their code-switched summaries. We hypothesize that this is due to their shorter summary length which makes code-switched tokens relatively more prominent. However, its scores in other code-switching metrics are significantly lower compared to CroCoSum.

When comparing to code-switching datasets, we note that in terms of CMI over code-switched examples, CroCoSum shows a lower score compared to speech corpora like ASCEND and SEAME. We assume that this stems from ASCEND/SEAME's sentences of colloquial text being on average shorter and of less formal diction than what is observed in the curated news domain. Yet our dataset still has a similar or higher percentage of code-switched sentences and SP.

To summarize, CroCoSum is the only CLS dataset that studies code-switching in human-written target summaries. It has longer source and target texts compared to other CLS datasets in the news domain, and a similar level of code-switching complexity compared to existing code-switching datasets in other tasks.

4. Experiments and Evaluations

In this section, we describe the details of our baseline approaches and metrics used for evaluation.

[6]Tokens that are shared by languages. For example, numerical digits.

4.1. Baselines

Similar to previous works (Nguyen and Daumé III, 2019; Ladhak et al., 2020; Wang et al., 2022), our baselines include pipeline methods that decompose the CLS task into machine translation and summarization as well as end-to-end approaches with direct cross-lingual supervision. Given the recent convincing performance of prompt learning with large language models, we also experiment with zero-shot prompting of GPT-3 to understand the dataset's difficulty and the necessity of training dedicated models.

Pipeline Pipeline methods decompose the CLS task into summarization and machine translation subtasks. The main reason behind employing such two-step processes in earlier works was a lack of cross-lingual resources at the time (Ladhak et al., 2020). Depending on the ordering of subtasks, methods can be further broken down into translate-then-summarize and summarize-then-translate approaches. We choose Google's translation API as Wang et al. (2022) find it to perform best in pipeline methods. We use it via the Translators[7] library as our translation module. In the summarization module, we finetune the same multilingually pretrained models described below in the end-to-end baseline.

- Translate-then-Summarize (Trans-Sum). We first translate all English source texts into Chinese. Then we finetune different models on the summarization task with the translated source texts and original Chinese code-switched summaries in our training set.
- Summarize-then-Translate (Sum-Trans). We finetune multilingual models with English source texts and English summaries translated from the Chinese ground truths. At inference time, the generated English summary is translated back into Chinese for final evaluation.

End-to-End The end-to-end method requires models to learn translation and summarization at the same time in a supervised manner. More specifically, the model takes in articles in the source language (English) and is expected to generate a summary in the target language (Chinese) directly. We adopt the following multilingually pretrained models as our cross-lingual summarizers.

- mT5 (Xue et al., 2020) is a multilingual variant of T5 (Raffel et al., 2020) that was pretrained in 101 languages unsupervisedly. We use mT5-base with 580M parameters to be closer in size to the other models below.
- mBART (Liu et al., 2020) is a sequence-to-sequence model using denoising objectives for

[7]https://github.com/uliontse/translators

	R1	R2	RL	BS	Sents.	Words	Sents$_{cs}$ %	CMI$_{all}$	CMI$_{cs}$	SP$_{all}$	SP$_{cs}$
GT$^{\alpha}$	-	-	-	-	5.96	225.05	55.28	4.75	5.11	2.19	2.36
Pipeline - Trans-Sum											
mT5	26.44	11.80	24.33	51.10	4.34	158.15	-1.95	+0.06	+1.01	**+0.08**	+0.53
mBART	34.97	16.39	30.72	57.70	5.80	240.89	+2.36	-0.10	-0.13	+0.19	+0.19
mBART50	34.91	16.77	30.85	57.77	5.67	228.29	+2.62	**+0.03**	**+0.01**	+0.20	+0.20
Pipeline - Sum-Trans											
mT5	24.93	11.43	23.08	47.08	7.03	174.98	-15.81	-0.67	+0.11	-0.99	-0.83
mBART	35.08	16.52	30.82	54.90	7.64	233.26	-7.12	-0.23	-0.24	-0.51	-0.54
mBART50	35.27	17.01	31.26	54.94	6.97	195.57	-8.08	-0.27	-0.26	-0.63	-0.67
End-to-End											
mT5	31.62	15.87	28.85	53.79	4.56	165.27	**+0.37**	+0.74	+1.28	+0.28	+0.52
mBART	38.44	19.94	34.01	58.67	5.33	193.77	+4.89	+0.86	+0.81	+0.18	**+0.14**
mBART50	**38.73**	**20.34**	**34.35**	**58.81**	5.22	189.08	+7.41	+1.31	+1.26	+0.29	+0.25
Zero-Shot											
GPT-3	26.50	12.33	24.00	52.08	4.74	153.53	-16.37	-1.62	-0.75	-0.95	-0.62

$^{\alpha}$ GT refers to the ground truth summaries in the test set. ROUGE and BERTScore are omitted since there's no prediction to compare to.

Table 3: Experimental results of different CLS baseline approaches.

neural machine translation. We use mBART25 which was trained on a 25-language monolingual corpus, and contains 610M parameters.

- mBART-50 (Tang et al., 2020) is an extension to mBART, adding tokens for additional languages in its embedding layer, and pretraining on a total of 50 languages. It is of the same size as mBART25.

Zero-shot Prompting Different from the methods above, the zero-shot method requires no training and relies on the models' generalizability to unseen tasks through manually crafted prompts. In this approach, we format each source-target pair in the test set as "Can you summarize the English article below in Chinese? ‹English source text›" and feed it into GPT-3 (Brown et al., 2020). It is expected to generate Chinese summaries following the prompts without task-specific training.

In the pipeline and end-to-end baselines, using the train set of CroCoSum, we finetune the summarization models based on their implementations in the transformers library (Wolf et al., 2020) for 15 epochs on a single RTX 3090 GPU and select the best checkpoint for final evaluation.

4.2. Evaluation Metrics

In Table 3, we report F1 scores of ROUGE-{1,2, L} (R1/R2/RL) (Lin, 2004) and BERTScore (BS) (Zhang et al., 2019) to compare the lexical and semantic similarities between the predicted and ground truth summaries. We also compute basic statistics such as the average number of sentences (Sents.) and word counts (Words). Differences in code-switched sentence percentage (Sents$_{cs}$ %) and code-switching metrics (CMI$_{all/cs}$, SP$_{all/cs}$) with respect to gold summaries in the test set are also reported for more comprehensive analysis.

Method	R1	R2	RL
FT$_{\text{CroCo}}$	38.73	20.34	34.35
PT$_{\text{Wiki}}$ + **FT**$_{\text{CroCo}}$	37.25	19.09	33.05
PT$_{\text{Cross}}$ + **FT**$_{\text{CroCo}}$	37.85	19.36	33.47
PT$_{\text{Wiki + Cross}}$ + **FT**$_{\text{CroCo}}$	37.19	18.81	32.79
FT$_{\text{Wiki + Cross + CroCo}}$	37.91	19.58	33.82

Table 4: Result Comparison of No Data Augmentation vs. Additional CLS Pretraining. **PT** means pretrain and **FT** means finetune.

5. Results and Analysis

Automatic Metrics Among all baseline methods, end-to-end finetuning generally attains the best performance in terms of both ROUGE and BERTScore. mBART50, specifically, works best compared to all other base models. Our results seem to contradict what Ladhak et al. (2020) and Wang et al. (2022) have reported, namely pipeline methods perform better than end-to-end finetuning methods. However, note that their dataset is able to supply gold monolingual article-summary pairs for training the summarizers by exploiting web-mined parallel resources or human translations while ours relies on

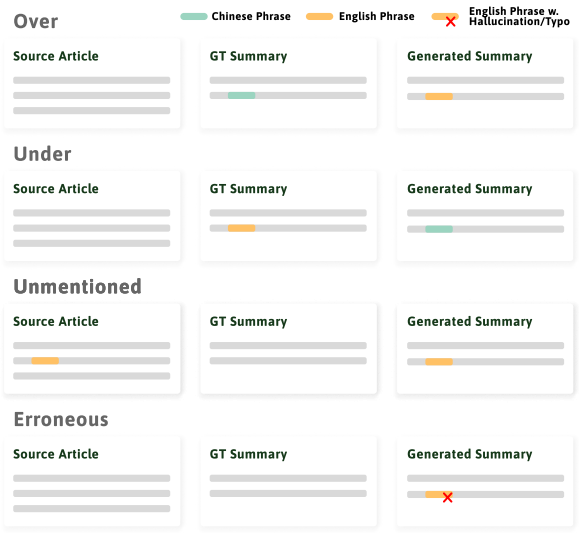

Figure 2: Illustration of Error Types.

silver pairs from the Google Translation API. This makes our pipeline methods more prone to error propagation, and therefore may lead to worse performance. This discrepancy in results suggests that end-to-end methods could provide promising results when there are no monolingual resources for training individual submodules.

We find zero-shot prompting with GPT-3's performance comparable to pipeline approaches with finetuned mT5. This result is encouraging as GPT-3 is only exposed to a small number of non-English tokens mixed in its English pretraining data, and it shows that simple prompts can elicit useful knowledge for unseen tasks without any specific training.

For different multilingually pretrained base models, we find that mBART50 and mBART have similar performance across different baselines and consistently outperform mT5. We hypothesize that this distinction may stem from the fact that mBART and mBART50 have already been pretrained for translation tasks, whereas mT5 has only been pretrained unsupervisedly, requiring additional effort to establish alignments between languages.

Data Augmentation To explore whether we can further boost our best-performing baseline, we leverage English-Chinese CLS pairs in WikiLingua and CrossSum in an additional pretraining step for mBART50 before finetuning it on CroCoSum. We only select the web-mined datasets instead of those created by translation as they contain more natural texts, and minimize the occurrences of translationese in the datasets. We experiment with three settings: 1) pretrain on WikiLingua or CrossSum then finetune on CroCoSum 2) pretrain on the shuffled set of WikiLingua and CrossSum, then finetune on CroCoSum 3) finetune on the shuffled set of WikiLingua, CrossSum and CroCoSum. Results in Table 4 show that pretraining with additional CLS pairs does not improve the model's performance on CroCoSum, but rather results in a slight score decrease, indicating limited generalizability provided by existing CLS datasets and motivating the need for more diverse CLS resources.

Code-switching Metrics Besides automatic metrics of summarization quality, we also compute code-switching metrics for the ground truth summaries and model generations using code-switched percentage, CMI and SP in Table 3. We assume the smaller the difference in metrics, the more closely the model prediction resembles how humans choose to code-switch in their writing. Yet, from the results, we note that the smallest differences are obtained by different baseline approaches as well as different pretrained base models. While the end-to-end finetuned mBART50 excels in automatic metrics, its level of code-switching is not the closest to the ground truth in any given code-switching metric. This discrepancy calls for more in-depth error analysis as current ngram-based auto metrics, limited to monolingual texts, fail to identify semantically correct instances like over/under-switched cases as described in Section 6 below.

6. Qualitative Analysis

To further investigate the challenges in generating code-switched summaries under CLS settings, we randomly select 100 test generations provided by our best-performing baseline and manually compare them to the corresponding ground truth summaries. We propose four error types to categorize the differences in code-switching tokens between model predictions and human-written summaries. See Figure 2 for a graphical representation of the error types. Each prediction may contain zero, one or more of the errors below.

Over Switched Phrases In 30 out of the 100 examples, the generated summaries contain English phrases that should have been Chinese according to the ground truth. We find that the generations tend to follow the three patterns below:

Under Switched Phrases 8 of the 100 generated summaries contain phrases in the target language which the ground truths chose to code-switch into the source language.

Unmentioned Code-switched Phrases 47 summaries contained English phrases that exist in the source texts but not in the ground truth summaries.

As the inputs are lengthy, the ground truths and predictions may focus on different aspects during summarization. In the example below, the prediction details the information source whereas the ground truth omits this information.

Erroneous Code-switched Phrases In 43 examples, the model generates misspelled English phrases or those that contradict, or are irrelevant to the source text. For example, it wrongly attributes World Wide Web inventor Tim Berners-Lee, to be the CTO instead of Chris Urmson, who actually holds that position.

As an additional point of reference, we collect human assessments of Fluency, Coherence, Informativeness and Relevance for 20 summary generations to the same sets of articles sampled and described in Sec 3.1: F 3.32, C 3.77, I 3.40, R 3.23. Compared to the scores on human reference summaries, these scores are significantly lower, indicating a lesser quality than human references.

As shown in our qualitative analysis, CroCoSum reveals various complications in evaluating code-switched summary generation. Especially for the first three cases, when models produce factual statements either in languages different from or omitted by the ground truth, a more comprehensive semantic evaluation that involves human judgment on naturalness and relevance is required.

7. Conclusion

In this paper, we introduce CroCoSum, the first collection of organic cross-lingual code-switched text summarization. CroCoSum distinguishes itself by exhibiting a significantly higher code-switching frequency when compared to existing CLS datasets, while still demonstrating comparable complexity to other non-summarization code-switching datasets. We provide benchmark performances of current CLS baseline approaches and an in-depth analysis highlighting the challenges of evaluating code-switched summaries using existing metrics.

Limitations

In our baseline experiments, observations are based on the model sizes allowed by our local compute resources. A more exhaustive analysis can be obtained by experimenting with greater baseline variation, including different model sizes, prompt templates, and few-shot experiments given a more generous compute budget.

Additionally, the scope of code-switching presented in our paper is restricted to a range of topics covered by the data source. The predominant presentation of code-switching occurs in the form of named entities, such as scientific terminologies and product names. Using code-switched terms, rather than their official Chinese equivalents, is a common practice among the authors and the intended audience. This writing style is favored because it facilitates rapid dissemination of news delivery and promotes more straightforward understanding. This focus on named entity mentions, however does not trivialize the problem. The decision between translation and literal copying of phrases and names follows intricate linguistic and cultural rules and is far from arbitrary.

Finally, given the diverse cultural contexts in which code-switching may occur, it is important to acknowledge that CroCoSum, and outputs from models finetuned on CroCoSum, may not fully encapsulate the complexities of actual code-switching practices among different linguistic and cultural backgrounds.

8. Bibliographical References

Ayana, Shi-qi Shen, Yun Chen, Cheng Yang, Zhiyuan Liu, and Mao-song Sun. 2018. Zero-shot cross-lingual neural headline generation.

IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26(12):2319–2327.

Utsab Barman, Amitava Das, Joachim Wagner, and Jennifer Foster. 2014. Code mixing: A challenge for language identification in the language of social media. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 13–23, Doha, Qatar. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Amitava Das and Björn Gambäck. 2014. Identifying languages at the word level in code-mixed Indian social media text. In *Proceedings of the 11th International Conference on Natural Language Processing*, pages 378–387, Goa, India. NLP Association of India.

A. Seza Doğruöz, Sunayana Sitaram, Barbara E. Bullock, and Almeida Jacqueline Toribio. 2021. A survey of code-switching: Linguistic and social perspectives for language technologies. 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 1654–1666, Online. Association for Computational Linguistics.

A Seza Doğruöz, Sunayana Sitaram, Barbara E Bullock, and Almeida Jacqueline Toribio. 2023. A survey of code-switching: Linguistic and social perspectives for language technologies. *arXiv preprint arXiv:2301.01967*.

Björn Gambäck and Amitava Das. 2014. On measuring the complexity of code-mixing. In *Proceedings of the 11th international conference on natural language processing, Goa, India*, pages 1–7.

Björn Gambäck and Amitava Das. 2016. Comparing the level of code-switching in corpora. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1850–1855, Portorož, Slovenia. European Language Resources Association (ELRA).

Yvette Graham, Barry Haddow, and Philipp Koehn. 2020. Statistical power and translationese in machine translation evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 72–81, Online. Association for Computational Linguistics.

Tahmid Hasan, Abhik Bhattacharjee, Md Saiful Islam, Kazi Samin, Yuan-Fang Li, Yong-Bin Kang, M Sohel Rahman, and Rifat Shahriyar. 2021. Xl-sum: Large-scale multilingual abstractive summarization for 44 languages. *arXiv preprint arXiv:2106.13822*.

Simran Khanuja, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, and Monojit Choudhury. 2020a. GLUECoS: An evaluation benchmark for code-switched NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3575–3585, Online. Association for Computational Linguistics.

Simran Khanuja, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, and Monojit Choudhury. 2020b. GLUECoS: An Evaluation Benchmark for Code-Switched NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3575–3585, Online. Association for Computational Linguistics.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, et al. 2022. The bigscience roots corpus: A 1.6 tb composite multilingual dataset. In

Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track.

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.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. 2021. Few-shot learning with multilingual language models. *arXiv preprint arXiv:2112.10668*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Sneha Mondal, Shreya Pathak, Preethi Jyothi, and Aravindan Raghuveer. 2022. Cocoa: An encoder-decoder model for controllable code-switched generation. page 14.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv e-prints*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. Mlsum: The multilingual summarization corpus. *arXiv preprint arXiv:2004.14900*.

Vinay Singh, Deepanshu Vijay, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018. Named entity recognition for Hindi-English code-mixed social media text. In *Proceedings of the Seventh Named Entities Workshop*, pages 27–35, Melbourne, Australia. Association for Computational Linguistics.

Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Ghoneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang, and Pascale Fung. 2014. Overview for the first shared task on language identification in code-switched data. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 62–72, Doha, Qatar. Association for Computational Linguistics.

Thamar Solorio and Yang Liu. 2008. Part-of-Speech tagging for English-Spanish code-switched text. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 1051–1060, Honolulu, Hawaii. Association for Computational Linguistics.

Victor Soto and Julia Hirschberg. 2018. Joint part-of-speech and language ID tagging for code-switched data. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 1–10, Melbourne, Australia. Association for Computational Linguistics.

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

Ishan Tarunesh, Syamantak Kumar, and Preethi Jyothi. 2021. From machine translation to code-switching: Generating high-quality code-switched text. 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 3154–3169, Online. Association for Computational Linguistics.

Xiaojun Wan, Huiying Li, and Jianguo Xiao. 2010. Cross-language document summarization based on machine translation quality prediction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 917–926, Uppsala, Sweden. Association for Computational Linguistics.

Jiaan Wang, Fandong Meng, Tingyi Zhang, Yunlong Liang, Jiarong Xu, Zhixu Li, and Jie Zhou. 2022a. Understanding translationese in cross-lingual summarization. *arXiv preprint arXiv:2212.07220*.

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.

Genta Indra Winata, Alham Fikri Aji, Zheng-Xin Yong, and Thamar Solorio. 2022. The decades

progress on code-switching research in nlp: A systematic survey on trends and challenges. *arXiv preprint arXiv:2212.09660*.

Genta Indra Winata, Samuel Cahyawijaya, Zihan Liu, Zhaojiang Lin, Andrea Madotto, and Pascale Fung. 2021. Are multilingual models effective in code-switching? In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 142–153, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Zheng-Xin Yong, Ruochen Zhang, Jessica Zosa Forde, Skyler Wang, Samuel Cahyawijaya, Holy Lovenia, Genta Indra Winata, Lintang Sutawika, Jan Christian Blaise Cruz, Long Phan, et al. 2023. Prompting multilingual large language models to generate code-mixed texts: The case of south east asian languages. *arXiv e-prints*, pages arXiv–2303.

Jiajun Zhang, Yu Zhou, and Chengqing Zong. 2016. Abstractive cross-language summarization via translation model enhanced predicate argument structure fusing. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(10):1842–1853.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

9. Language Resource References

Aguilar, Gustavo and Kar, Sudipta and Solorio, Thamar. 2020. *LinCE: A Centralized Benchmark for Linguistic Code-switching Evaluation*. European Language Resources Association.

Gliwa, Bogdan and Mochol, Iwona and Biesek, Maciej and Wawer, Aleksander. 2019. *SAMSum Corpus: A Human-annotated Dialogue Dataset for Abstractive Summarization*. Association for Computational Linguistics.

Grusky, Max and Naaman, Mor and Artzi, Yoav. 2018. *Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies*.

Tahmid Hasan and Abhik Bhattacharjee and Wasi Uddin Ahmad and Yuan-Fang Li and Yong-bin Kang and Rifat Shahriyar. 2021. *CrossSum: Beyond English-Centric Cross-Lingual Abstractive Text Summarization for 1500+ Language Pairs*.

Ladhak, Faisal and Durmus, Esin and Cardie, Claire and McKeown, Kathleen. 2020. *WikiLingua: A New Benchmark Dataset for Cross-Lingual Abstractive Summarization*. Association for Computational Linguistics.

Li, Chengfei and Deng, Shuhao and Wang, Yaoping and Wang, Guangjing and Gong, Yaguang and Chen, Changbin and Bai, Jinfeng. 2022. *TALCS: An Open-Source Mandarin-English Code-Switching Corpus and a Speech Recognition Baseline*.

Li, Ying and Yu, Yue and Fung, Pascale. 2012. *A mandarin-english code-switching corpus*.

Lovenia, Holy and Cahyawijaya, Samuel and Winata, Genta and Xu, Peng and Xu, Yan and Liu, Zihan and Frieske, Rita and Yu, Tiezheng and Dai, Wenliang and Barezi, Elham J. and Chen, Qifeng and Ma, Xiaojuan and Shi, Bertram and Fung, Pascale. 2022. *ASCEND: A Spontaneous Chinese-English Dataset for Code-switching in Multi-turn Conversation*. European Language Resources Association.

Lyu, Dau-Cheng and Tan, Tien-Ping and Chng, Eng Siong and Li, Haizhou. 2010. *Seame: a mandarin-english code-switching speech corpus in south-east asia*.

Mehnaz, Laiba and Mahata, Debanjan and Gosangi, Rakesh and Gunturi, Uma Sushmitha and Jain, Riya and Gupta, Gauri and Kumar, Amardeep and Lee, Isabelle G and Acharya, Anish and Shah, Rajiv. 2021. *GupShup: Summarizing open-domain code-switched conversations*.

Nguyen, Khanh and Daumé III, Hal. 2019. *Global Voices: Crossing Borders in Automatic News Summarization*. Association for Computational Linguistics.

Laura Perez-Beltrachini and Mirella Lapata. 2021. *Models and Datasets for Cross-Lingual Summarisation*.

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

Zheng, Shaohui and Li, Zhixu and Wang, Jiaan and Qu, Jianfeng and Liu, An and Zhao, Lei and Chen, Zhigang. 2022. *Long-Document Cross-Lingual Summarization*.

Zhu, Junnan and Wang, Qian and Wang, Yining and Zhou, Yu and Zhang, Jiajun and Wang, Shaonan and Zong, Chengqing. 2019. *NCLS: Neural Cross-Lingual Summarization*. Association for Computational Linguistics.