Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 6932–6937
Marseille, 20-25 June 2022

CATAMARAN: A Cross-lingual Long Text Abstractive Summarization Dataset

Zheng Chen, Hongyu Lin
University of Electronic Science and Technology of China
No.4, Section 2, North Jianshe Road, Chengdu, China
zchen@uestc.edu.cn

Abstract

Cross-lingual summarization, which produces the summary in one language from a given source document in another language, could be extremely helpful for humans to obtain information across the world. However, it is still a little-explored task due to the lack of datasets. Recent studies are primarily based on pseudo-cross-lingual datasets obtained by translation. Such an approach would inevitably lead to the loss of information in the original document and introduce noise into the summary, thus hurting the overall performance. In this paper, we present CATAMARAN, the first high-quality cross-lingual long text abstractive summarization dataset. It contains about 20,000 parallel news articles and corresponding summaries, all written by humans. The average lengths of articles are 1133.65 for English articles and 2035.33 for Chinese articles, and the average lengths of the summaries are 26.59 and 70.05, respectively. We train and evaluate an mBART-based cross-lingual abstractive summarization model using our dataset. The result shows that, compared with mono-lingual systems, the cross-lingual abstractive summarization system could also achieve solid performance.

Keywords: Abstractive summarization, Cross-lingual summarization, Long text summarization

1. Introduction

Abstractive summarization aims to produce a short rephrasing of text that contains compressed and refined information according to a long source document. Existing research in abstractive summarization mostly focuses on monolingual task, the performance of which has been tremendously advanced due to the use of massive pre-trained language models(Devlin et al., 2019; Lewis et al., 2020; Radford et al., 2019). However, cross-lingual abstractive summarization, which is another important sub-field of abstractive summarization, shows hardly explored due to its difficulty and the lack of high-quality datasets(Elhadad et al., 2013; Giannakopoulos, 2013).

Given a source document in one language, the objective of cross-lingual abstractive summarization is to generate a summary in another language, which could facilitate the comprehension of long textual information, especially news stories, in a foreign language. Intuitively, cross-lingual abstractive summarization combines both abstractive summarization and machine translation. Hence, early approaches are based on a simple two-step strategy: first translate the source document then summarize (Leuski et al., 2003; Wan et al., 2010; Ouyang et al., 2019) or first summarize then translate the summary to the target language (Wan et al., 2010; Orăsan and Chiorean, 2008). Despite the widespread use of this two-stage pipeline paradigm, it is not reliable to a considerable degree. Existing machine translation systems, which are powerful yet not precise, might introduce loss of information and mistranslation when translating the original documents or summaries. In addition, running two independent systems also increases the time and memory cost, which makes the two-step approach hardly applicable in the real world (Ladhak et al., 2020).

To the best of our knowledge, the best way to alleviate these problems is to adopt a direct end-to-end cross-lingual abstractive summarization system. However, previous studies of the end-to-end approach are very limited, as there is no available high-quality cross-lingual abstractive summarization dataset. (Zhu et al., 2019) construct a dataset via round-trip translation on large scale monolingual datasets.

	En.	**Zh.**
Mean article length	1133.65	2035.33
Max article length	16123	29885
Mean summary length	26.59	70.05
Max summary length	115	156
Mean title length	9.91	15.79
Mean sentence length	25.57	39.77
Mean number of sentences	44.34	51.26
Compression Ratio	0.96	0.57

Table 1: Characteristics of CATAMARAN

Based on the pseudo cross-lingual dataset, they perform multi-task training that combines machine translation and cross-lingual abstractive summarization to obtain an end-to-end system. Without cross-lingual dataset, (Duan et al., 2019; Shen et al., 2018) propose a zero-shot system based on the teacher-student framework, whose student network still learned from the translated data thus suffers from the inaccuracy of the machine translation system.

To address the scarcity of datasets in cross-lingual abstractive summarization, we introduce CATAMARAN[1], a high-quality Cross-linguAl long Text Abst

[1] https://figshare.com/articles/dataset/CATAMARAN_v1_json/15001371

-ractive suMmARizAtioN dataset that contains about 20,000 parallel Chinese and English human-written news articles and corresponding summaries crawled from New York Times web sites. We train four abstractive summarization systems, two monolingual and two cross-lingual, by finetuning the mBART (Liu et al., 2020) model on our dataset. By comparing the performance of the monolingual and crosslingual systems, we prove the feasibility of obtaining a high-performance end-to-end crosslingual abstractive summarization system with a small yet high-quality human-written crosslingual dataset. We also conduct experiments to demonstrate that a system trained on a real cross-lingual dataset outperforms the ones based on traditional two-stage strategy as well as pseudo cross-lingual approach. Detailed results are shown in Section 3.

2. CATAMARAN DATASET

New York Times is one of largest international news media in the world. It has three versions of website in different languages, which are English, Chinese and Spanish, we focus on constructing English and Chinese parallel cross-lingual dataset in this work.

2.1. Data Collection

Since there are fewer articles on New York Times Chinese website and the news pages of Chinese website contains corresponding entries of parallel bilingual and English news page, which provides great convenience, we take New York Times Chinese website as the start point to launch our process of data collection.

We exclude several news catogries which mainly cover picture News and then determine to crawl data in twelve categories including "International", "China", "Commercial", "Technology", "Science", "Lens", "Health", "Education", "Calture", "Style", "Travel" and "Opinion". In the Chinese website, we can extract parallel bilingual title and content, and Chinese summary. In order to get the English summary, we search the news title on the English website and then find the one matches in the search results as the corresponding English news page.

2.2. Data Filtering

Based on our observation, we find some noise in the raw data as follows.

1. The webpage contains a lot of web links.
2. Bi-lingual news are not properly aligned.
3. The summary is not related with the news content, for example, pieces of morning news' summary are all "Here's what you need to know to start your day".
4. One news article may belong to multiple categories, causing data duplication.

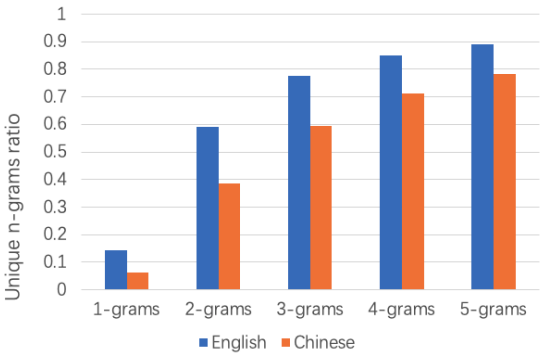

Figure 1: The novel n-grams ratio in CATAMARAN

For addressing these problems, we perform a filtering based on several simple yet effective rules. In terms of situation 1. , we make use of regular expression to recognize the link while counting the number of non-Chinese and non-English characters in the raw data, we will discard data with numerous links or illegal characters but retain ones with a smaller amount of noise in order to enhance robustness of the model; In terms of situation 2. , we check whether empty strings exist in the parallel bilingual sentences, if a sentence in one language is an empty string but its parallel sentence in another language is not, we can conclude that there is a problem with the alignment and discard this piece of data; In terms of situation 3, we intuitively review summaries which bear a high resemblance to each other so as to find and delete data with invalid summary. After data filtering and de-duplication, a total of 18,614 unique cross-lingual article-summary pairs are left.

2.3. Data Properties

2.3.1. Basic Analysis

To observe the characteristics of the dataset more explicitly, we perform some basic statistics for the CATAMARAN, which are shown in Table 1.

2.3.2. Abstractiveness and Compression Ratio

To quantify the level of abstractiveness and the difficulty of summarization, we illustrate the unique n-grams proportion (Sharma et al., 2019) and the compression ratio (Koupaee and Wang, 2018) of CATAMARAN. The abstractiveness of a summary can be measured by calculating the proportion of unique n-grams in the summary that do not exist in the article. The higher the proportion, the more novel paraphrases the summary contains. The compression ratio is the ratio between the average sentence length in the articles and the average length of summaries. Higher compression ratio means summarization task on this dataset is more difficult. CATAMARAN's abstractiveness and compression ratio are shown in Figure 1 and Table 1, respectively.

	ROUGE-1	ROUGE-2	ROUGE-L
MonoSum$_{Zh2Zh}$	33.2	11.27	25.82
CrossSum$_{En2Zh}$	29.18(-4.02)	6.16(-5.11)	21.67(-4.15)
Pseudo CrossSum$_{En2Zh}$	28.65(-4.55)	5.84(-5.43)	21.02(-4.8)
Trans $\rightarrow$ Sum$_{En2Zh}$	25.79(-7.41)	4.83(-6.44)	19.23(-6.59)
Sum $\rightarrow$ Trans$_{En2Zh}$	21.67(-11.53)	3.36(-7.91)	15.37(-10.45)
MonoSum$_{En2En}$	25.38	7.29	19.75
CrossSum$_{Zh2En}$	24.72(-0.66)	5.84(-1.45)	18.79(-0.96)
Pseudo CrossSum$_{Zh2En}$	24.2(-1.18)	5.44(-1.85)	18.51(-1.24)
Trans $\rightarrow$ Sum$_{Zh2En}$	22.62(-2.76)	4.38(-2.91)	16.82(-2.93)
Sum $\rightarrow$ Trans$_{Zh2En}$	22.45(-2.93)	4.08(-3.21)	16.2(-3.55)

Table 2: Comparison of different approaches.

2.3.3. Key Information Distribution

The distribution of the reference summary's information in the article is also an important indicator of the dataset. If the information in the reference summaries appears in the first sentences of a article, it means that the summarization task is less challenging. According to the cross-lingual character, we make use of semantic similarity to calculate the score between the reference summary and each sentence in the article. We select the top 3 sentences with the highest scores as representatives, so as to determine the key information distribution in the articles. We conduct this metric on Zh2En and En2Zh set. Results are shown in Figure 2.

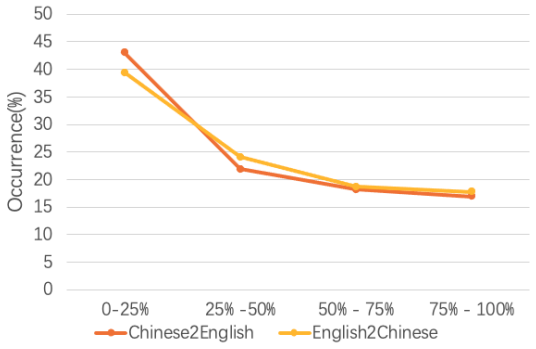

Figure 2: Key information distribution in different parts of articles in CATAMARAN

3. Experiments

To evaluate our dataset, we train 10 abstractive summarization models and conduct comprehensive comparisions. All of these models use mbart-large-cc25 published on HuggingFace by Facebook as starting checkpoint. Our GPUs for training are two RTX 3090. We randomly extract 500 pieces of data as validation set while another 500 pieces of data as prediction set. Other details and results will be revealed in the following subsections.

3.1. Evaluation Models

To demonstrate the feasibility of training a solid cross-lingual summarization system with a high-quality cross-lingual dataset, we train two direct end-to-end cross-lingual summarization models (CrossSum$_{Zh2En}$ and CrossSum$_{En2Zh}$) and two monolingual summarization models (MonoSum$_{Zh2Zh}$ and MonoSum$_{En2En}$). We perform abstractive summarization with these four models on the test set, and then calculate Rouge-1, Rouge-2 and Rouge-L between the generated summaries and the reference.

We also implement some baseline approaches. For the classic Two-Stage Strategy, we implement both "First translate then summarize" and "First summarize then translate". For the Trans$\rightarrow$Sum approaches, we leverage a transformer-to-transformer model to translate the articles, then input it into our fine-tuned monolingual summarization model. For the Sum$\rightarrow$Trans approaches, we directly get the summaries by the monolingual model, then translate the summaries into target language. For pseudo cross-lingual approaches, we follow the implement in (Zhu et al., 2019). Since the cross-lingual dataset are translated from the monolingual data, we call them pseudo cross-lingual approaches.

3.2. Human Evaluation

Since ROUGE(Lin, 2004) cannot accurately and comprehensively represent the quality of summaries(Maynez et al., 2020; Nallapati et al., 2016), we introduce human evaluation. We randomly select 100 pieces of cross-lingual summary from each system, and send each summary to 2 annotators to perform the evaluation. Each summary should be scored for Informative, Fluent and Faithful on a scale from 0-5. Informative is used to evaluate how much salient information or fragments in the articles that the generated summaries contain. Fluent indicates that whether the summary is smooth and fluent without grammatical errors. Faithful shows the degree that how faithful the generated summaries are to the articles, that is to say, is there any inconsistency between the facts and information in summaries and that of the articles.

3.3. Results and Analysis

Table 2 shows the comparison of the summarization models in ROUGE metrics. By comparing the perfor-

	Informative		Fluent		Faithful		Average	
	Zh2En	En2Zh	Zh2En	En2Zh	Zh2En	En2Zh	Zh2En	En2Zh
Summarize → Translate	2.97	2.85	2.64	2.26	2.87	2.13	2.83	2.41
Translate → Summarize	3.15	3.22	2.92	2.69	3.01	3.21	3.03	3.04
Pseudo CrossSum	3.63	3.87	3.71	3.54	3.37	**3.44**	3.57	3.62
CrossSum(ours)	**3.68**	**3.92**	**3.95**	**3.81**	**3.53**	3.39	**3.72**	**3.71**

Table 3: Human evaluation results on cross-lingual summarization based different approaches

mance of the these systems, we can see that the performance of CrossSum is the closest to the MonoSum, which suggests that direct cross-lingual summarization systems can perform solidly without error accumulation. As a relatively novel approach, the performance of pseudo cross-lingual summarization is close to that of CrossSum, however, it still has a certain gap with CrossSum due to the loss of information. The earlier traditional systems show a disparity in performance due to error propagation. In the two traditional systems, Trans → Sum significantly outperforms Sum → Trans. We suspect that since the article is a long text, identical information may appear in several different parts of the article, making the model not very sensitive for noise. In contrast, the summary is short and has a larger density of information, which means key information and words often only appear once, therefore, the loss of information caused by translation is more fatal to the summary.

In Table 3, although CrossSum is slightly worse than Pseudo CrossSum on Faithful, CrossSum still works best. In terms of the overall performance, Pseudo CrossSum is about 3% worse than CrossSum, while Trans → Sum and Sum → Trans are about 18% and 29% lower, respectively. This suggests that a high-quality cross-lingual abstractive summarization dataset can greatly boost the performance of the cross-lingual summarization system without a requirement of a large scale of data.

3.4. Data Samples

Here is a sample of generated result in experiments.

English Article: VANCOUVER, British Columbia — Meng Wanzhou, a top executive of the Chinese technology company Huawei, was granted bail of 10 million Canadian dollars, or about $7.5 million, while awaiting extradition to the United States from Canada, a judge ruled on Tuesday.The decision came on the third day of a bail hearing for Ms. Meng, who is also a daughter of the Huawei founder Ren Zhengfei, in a case that has complicated the relationship between China and the United States."I am satisfied that on the particular facts of this case, including the fact that Ms. Meng is a well-educated businesswoman who has no criminal record and of whom several people have attested to her good character, the risk of her non-attendance in court can be reduced to an acceptable level," Justice William Ehrcke said in his ruling.Ms. Meng and her husband will be responsible for a 7 million dollar cash deposit for bail, with the remaining 3 million dollars coming from her acquaintances. She will be released upon completion of court paperwork.Ms. Meng will be subject to 24-hour physical and electronic surveillance, which she plans to pay for herself. Physical surveillance will include two security guards and a driver, while electronic surveillance will include GPS monitoring. She will also be under curfew between 11 p.m. and 6 a.m., the judge said...

Chinese Article: 加拿大温哥华——加拿大一名法官于周二裁定,中国科技公司华为高管孟晚舟获准在等待美国引渡期间,以1000万加元(约合5100万人民币)获得保释。这一决定是在保释听证会的第三天做出的。孟晚舟是华为创始人任正非的女儿,此案令中美关系进一步复杂化。"此案的具体情况令我满意,孟女士是一位受过良好教育的职业女性,没有犯罪记录,有几个人证明了她品格良好,她不出庭的风险可以降低到一个可接受的水平,"法官威廉·厄尔克(William Ehrcke)在裁决中说。孟晚舟和丈夫将负责缴纳700万加元现金作为保释金,其余300万加元来自她的熟人。她将在完成法庭文书工作后获释。孟晚舟将接受24小时的有形监控和电子无形监控,她计划自己支付监控费用。有形监控包括两名保安和一名司机,电子监控包括GPS监控。法官说,晚11点到早6点之间她还被禁止外出...

English Reference Summary: Meng Wanzhou, a Huawei executive who American authorities want to extradite from Canada on fraud charges, was granted bail of 10 million dollars on Tuesday.

Chinese Reference Summary: 孟晚舟以1000万加元获保释,将接受24小时监控,2月6日将再次出庭。若美国提出正式引渡请求并获得批准,孟晚舟有几个上诉选择,整个过程可能持续数月。

English Summary Generated by MonoSum: The decision came on the third day of a bail hearing for Meng Wanzhou, a top executive of the Chinese technology company Huawei, in a case that has complicated the relationship between China and the United States.

Chinese Summary Generated by MonoSum: 孟晚舟将接受24小时的有形监控和电子无形监控,她计划自己支付监控费用。她将在完成法庭文书工作后获释。孟晚舟是华为创始人任正非的女儿,此事令中美关系进一步复杂化。

English Summary Generated by CrossSum: A Canadian court ruled that Meng Wanzhou, the daughter of the founder of Huawei, had been allowed to stay indefinitely in the United States while waiting for a release.

Chinese Summary Generated by CrossSum: 加拿

大法官裁定孟晚舟被保释金1000万加元,但仍需缴纳巨额保释金。孟的律师称其被捕是出于政治动机,并称其已向加拿大提交引渡申请。

English Summary Generated by Pseudo CrossSum: A Canadian judge ruled on Tuesday that the daughter of the founder of the Chinese tech company Huawei, Liu Ying-jeou, was granted bail for the time she waited for the United States extradition.

Chinese Summary Generated by Pseudo CrossSum: 孟晚舟在加拿大被捕,她和丈夫将承担7万美元的保释金,其余3万美元来自朋友。孟的律师称,孟晚舟没有犯罪记录,但有好斗的一面。孟将面临24小时的监视。

English Summary Generated by Trans → Sum: The daughter of the founder of China Technology Corporation was allowed to bail at $10 million while awaiting extradition from the United States.",

Chinese Summary Generated by Trans → Sum: 孟万州是Huawei创始人Ren Zhengfei的女儿,被控欺骗金融机构,并导致它们违反对伊朗的制裁。她和丈夫将负责700万美元的现金保释金,其余300万美元来自她熟人。

English Summary Generated by Sum → Trans: Mon-Teong will be subject to 24 hours of physical and electronic invisible surveillance, and she plans to pay for it herself. She will be released after completing her court paperwork. Mon-Teong is the daughter of China's founder, and this complicates China-US relations. **Chinese Summary Generated by Sum → Trans**: 这一决定是在中国技术公司华威公司顶尖执行官孟万州保释听证会的第三天作出的,该案使中国和美国之间的关系复杂化。

4. Conclusion

In this paper, we present CATAMARAN, the first high-quality cross-lingual long text news abstractive summarization dataset. It could significantly boost the performance of cross-lingual summarization model compared with the ones based on previous approaches through our evaluation. Besides, we show that the performance of cross-lingual summarization systems trained on a genuine cross-lingual dataset is the closest one to that of the monolingual systems. Considering the vital role that cross-lingual summarization plays, our next plan is to build a larger high-quality cross-lingual summarization dataset by collecting data resources from various web sites.

5. References

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

Duan, X., Yin, M., Zhang, M., Chen, B., and Luo, W. (2019). Zero-shot cross-lingual abstractive sentence summarization through teaching generation and attention. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3162–3172, Florence, Italy, July. Association for Computational Linguistics.

Elhadad, M., Miranda-Jiménez, S., Steinberger, J., and Giannakopoulos, G. (2013). Multi-document multilingual summarization corpus preparation, part 2: Czech, Hebrew and Spanish. In *Proceedings of the MultiLing 2013 Workshop on Multilingual Multi-document Summarization*, pages 13–19, Sofia, Bulgaria, August. Association for Computational Linguistics.

Giannakopoulos, G. (2013). Multi-document multilingual summarization and evaluation tracks in ACL 2013 MultiLing workshop. In *Proceedings of the MultiLing 2013 Workshop on Multilingual Multi-document Summarization*, pages 20–28, Sofia, Bulgaria, August. Association for Computational Linguistics.

Koupaee, M. and Wang, W. Y. (2018). Wikihow: A large scale text summarization dataset.

Ladhak, F., Durmus, E., Cardie, C., and McKeown, K. (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, November. Association for Computational Linguistics.

Leuski, A., Lin, C.-Y., Zhou, L., Germann, U., Och, F. J., and Hovy, E. (2003). Cross-lingual c*st*rd: English access to hindi information. *ACM Transactions on Asian Language Information Processing*, 2(3):245–269, September.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (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, July. Association for Computational Linguistics.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *ACL 2004*.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation.

Maynez, J., Narayan, S., Bohnet, B., and McDonald, R. T. (2020). On faithfulness and factuality in abstractive summarization. *ArXiv*, abs/2005.00661.

Nallapati, R., Zhou, B., dos Santos, C. N., Çaglar Gülçehre, and Xiang, B. (2016). Abstractive text summarization using sequence-to-sequence rnns and beyond. In *CoNLL*.

Orăsan, C. and Chiorean, O. A. (2008). Evaluation of a cross-lingual Romanian-English multi-document summariser. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May. European Language Resources Association (ELRA).

Ouyang, J., Song, B., and McKeown, K. (2019). A robust abstractive system for cross-lingual summa-

rization. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2025–2031, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners.

Sharma, E., Li, C., and Wang, L. (2019). Bigpatent: A large-scale dataset for abstractive and coherent summarization.

Shen, S.-q., Chen, Y., Yang, C., Liu, Z.-y., and Sun, M.-s. (2018). Zero-shot cross-lingual neural headline generation. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 26(12):2319–2327, December.

Wan, X., Li, H., and Xiao, J. (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, July. Association for Computational Linguistics.

Zhu, J., Wang, Q., Wang, Y., Zhou, Y., Zhang, J., Wang, S., and Zong, C. (2019). Ncls: Neural cross-lingual summarization.