CrossSum: Beyond English-Centric Cross-Lingual Abstract Text Summarization for 1500+ Language Pairs

Tahmid Hasan¹, Abhik Bhattacharjee¹, Wasi Uddin Ahmad², Yuan-Fang Li³, Yong-Bin Kang³, Rifat Shahriyar¹

Bangladesh University of Engineering and Technology (BUET)¹, University of California, Los Angeles², Monash University³, Swinburne University of Technology⁴
tahmidhasan@cse.buet.ac.bd, abhik@ra.cse.buet.ac.bd, wasiahmad@ucla.edu, yuanfang.li@monash.edu, ykang@swin.edu.au, rifat@cse.buet.ac.bd

Abstract

We present CrossSum, a large-scale dataset comprising 1.65 million cross-lingual article-summary samples in 1500+ language-pairs constituting 45 languages. We use the multilingual XL-Sum dataset and align identical articles written in different languages via cross-lingual retrieval using a language-agnostic representation model. We propose a multi-stage data sampling algorithm and fine-tune mT5, a multilingual pretrained model, with explicit cross-lingual supervision with CrossSum and introduce a new metric for evaluating cross-lingual summarization. Results on established and our proposed metrics indicate that models fine-tuned on CrossSum outperforms summarization+translation baselines, even when the source and target language pairs are linguistically distant. To the best of our knowledge, CrossSum is the largest cross-lingual summarization dataset and also the first-ever that does not rely on English as the pivot language.

1 Introduction

Cross-lingual summarization is the task of generating a summary in a target language given a source text in another language. The task is challenging is the sense that it combines summarization and translation in one single task. Earlier approaches to cross-lingual summarization thus employed a translation-summarization (Leuski et al., 2003) or summarization-translation pipeline (Wan et al., 2010) to generate the summary in the target language. Not only computationally expensive having to use separate models for each task, these approaches suffer from error-propagation (Zhu et al., 2019) from one model to another, resulting in the degradation of performance. They also assume

Input Article: [... ] (Dexamethasone was tested as part of a global clinical trial to test the effectiveness of various existing therapies against the new coronavirus.) [... ]

Summary:  

Scientists say a cheap and readily available drug called dexamethasone will help save the lives of critically ill patients with coronavirus.

Figure 1: A sample article-summary pair from CrossSum, the article is written in Japanese and the summary is in Bengali. We translate the texts in English for better understanding. Word and phrases of the article relevant to the summary are color-coded.

The availability of translation model for the source and target languages, which may not hold for low-resource languages. The success of sequence-to-sequence (seq2seq) models (Cho et al., 2014; Sutskever et al., 2014) over the last decade and the recent advances in Transformer-based models (Vaswani et al., 2017; Rothe et al., 2020) have aided in the emergence of end-to-end methods that can jointly summarize and translate to produce cross-lingual summaries with one single model (Zhu et al., 2019). The availability of cross-lingual summarization datasets (Ladhak et al., 2020; Perez-Beltrachini and Lapata, 2021) has also helped this task gain popularity in the last few years. However, these datasets often exhibit certain limitations, they are cover only a few languages, have few samples for training and evaluation, or use English as the pivot language,
thereby limiting the applicability of cross-lingual summarization to a wide variety of languages.

To democratize cross-lingual summarization beyond the high-resource languages, in this work, we introduce CrossSum, a large-scale cross-lingual abstractive summarization dataset containing 1.65 million article-summary samples in 1500+ language-pairs, aligned from the multilingual XL-Sum (Hasan et al., 2021) dataset. CrossSum is the first publicly available cross-lingual abstractive summarization dataset for a large number of language pairs. We design a multistage sampling algorithm for successful training of multilingual models on cross-lingual text generation task. We propose a new automatic metric for evaluating cross-lingual summaries when reference summaries in the target language may not be available. For the very first time, we perform cross-lingual summarization on a broad and diverse set of languages without relying on English as the standalone pivot language. We achieve strong results on non-English pivots and even on language pairs, outperforming summarization-translation baselines. In summary, we make the following contributions:

- We introduce and release CrossSum.
- We design a multistage sampling algorithm for cross-lingual text generation training.
- We propose a new automatic metric for cross-lingual summary evaluation.
- We perform training on cross-lingual summarization achieving state-of-the-art results on non-English pivots and distant language pairs.

We are releasing the dataset, alignment and training scripts, and trained models in the hope that these resources will encourage the community to push the boundaries of cross-lingual abstractive summarization beyond the English and other high-resources.

2 The CrossSum Dataset

Language-agnostic text sequence representations (Artetxe and Schwenk, 2018) provide a way to identify identical contents across languages. For instance, the Language-agnostic BERT Sentence Representation (LaBSE) (Feng et al., 2020) model achieved state-of-the-art results on the BUCC (Zweigenbaum et al., 2017) and Tatoeba (Artetxe and Schwenk, 2019) cross-lingual text mining tasks.

It is possible to automatically curate a large-scale cross-lingual summarization dataset employing this cross-lingual retrieval strategy, with the assumption that a large multilingual summarization dataset be available where different languages have identical contents. Two contemporary works have compiled large-scale multilingual summarization datasets in a large number of languages, namely XL-Sum (1.35M samples in 45 languages) (Hasan et al., 2021) and MassiveSumm (Varab and Schluter, 2021) (28.8M samples in 92 languages).

While it may seem that MassiveSumm should be the ideal choice for cross-lingual dataset curation, the dataset is not publicly available\(^1\). Since public availability is essential for promoting open research, we opted for the other alternative, XL-Sum, which is distributed under a non-commercial research license. XL-Sum has another benefit: all articles are crawled from a single source, BBC News. Since BBC publishes similar news contents in different languages, it would increase the quality the cross-lingual mining yields.

We first encoded all summaries using LaBSE (Feng et al., 2020) (which outputs a unit vector embedding for an input text sequence). Using these embeddings, for each summary in a language, we searched its nearest neighbors in the other languages using the FAISS library (Johnson et al., 2017a). We used cosine similarity as the metric for neighbor search. To reduce the number of incorrect alignments, we further set a minimum similarity score of 0.735 before accepting a matched alignment.\(^2\) As validation, we performed a small human evaluation with three language pairs and found over 90% alignments ($\kappa = 0.7 - 0.8$) to be correct. Since Hasan et al. (2021) reported summaries around this percentage to be good-quality in XL-Sum, we did not change the thresholds. Figure 2 shows the article-summary statistics for all language-pairs in the CrossSum dataset. To avoid any potential data leakage and to allow us to use the language-agnostic summary evaluation metric (to be discussed in Section 3.2), we ensure identical articles do not occur in train set of one language and dev/test set of another.

\(^1\)https://github.com/danielvarab/massive-summ
\(^2\)If the HTML pages of the matched articles have other identical information, e.g., presence of the same image, we lower the threshold to 0.666.
Figure 2: A bubble plot containing the article-summary statistics of the CrossSum dataset. Languages in the axes are sorted by the number of Wikipedia entries for those languages to show a sequential contrast from high- to low-resource languages. The radius of the bubbles depict the number of article-summary pairs for the corresponding language pair. We considered a language pair to be low-resource in CrossSum if the number of samples is below 500, mid-resource for 500 to less than 5000, and high-resource for pairs exceeding 5000.

3 Training & Evaluation Methodologies

In this section, we discuss the multistage sampling strategy for training cross-lingual text generation models and our proposed metric for evaluating model-generated summaries.

3.1 Multistage Language Sampling

From figure 2, we can see that the CrossSum dataset is heavily imbalanced in terms of samples for different language pairs and thus training directly without upsampling low-resource languages may result in degraded performance for them. Conneau et al. (2020) upsamled low-resource languages using a probability smoothing for multilingual language model pretraining. They used a batch size of 256 and ensured all data points of a batch be sampled from one language. However, if we did the same for the language pairs in CrossSum, many batches would have duplicate samples since many language pairs do not have more than 256 samples and at the same time, many language pairs would not be sampled during training (since we are limited by computational resources) for lack
of enough training steps. To address this issue, we adapt their upsampling algorithm to perform multistage upsampling and ensure either the source or the target texts of a batch are sampled from the same language.

Let $L_1, L_2, \cdots, L_n$ be the languages of a cross-lingual source-target dataset. Let $c_{ij}$ be the number of training samples where the source is from $L_i$ and target from $L_j$. We compute the probabilities of the source languages by

$$p_i = \frac{\sum_{j=1}^{n} c_{ij}}{\sum_{j=1}^{n} \sum_{k=1}^{n} c_{jk}} \forall i \in \{1, 2, \cdots, n\} \tag{1}$$

We then use an exponent smoothing factor $\alpha$ and normalize the probabilities

$$q_i = \frac{p_i^\alpha}{\sum_{j=1}^{n} p_j^\alpha} \forall i \in \{1, 2, \cdots, n\} \tag{2}$$

Given the source language $L_i$, we now compute the probabilities of its target languages.

$$p_{ji} = \frac{c_{ij}}{\sum_{k=1}^{n} c_{ik}} \forall j \in \{1, 2, \cdots, n\} \tag{3}$$

We again smooth $p_{ji}$ by a factor of $\beta$ and obtain the normalized probabilities

$$q_{ji} = \frac{p_{ji}^\beta}{\sum_{k=1}^{n} p_{kj}^\beta} \forall j \in \{1, 2, \cdots, n\} \tag{4}$$

We analogously compute $p_j$ and $p_{lj}$ and using them, describe the training algorithm with multistage sampling in Algorithm 1.

Note that our proposed algorithm can be applied to any cross-lingual seq2seq task where both the source and target languages are imbalanced.

### 3.2 Evaluating Summaries Across Languages

A sufficient number of reference samples are mandatory for reliable evaluation of model-generated summaries. However, for many Cross-Sum language-pairs, especially low-resource ones, even the training sets are very small, let alone their test sets. Since we respected the original XL-Sum train-dev-test splits, being able to somehow evaluate using reference summaries written in a different language (the language of the source text for our case) would allow evaluation on a broad range of languages, especially for which there are inadequate references in the target language. Embedding-based similarity metrics (Zhang et al., 2019; Zhao et al., 2019) have gained popularity in the last few years. We draw inspiration from them and design a similarity metric that does not rely on lexical overlap between the generated and reference texts. As a result, this new metric can effectively measure similarity across languages in a language-independent manner. We consider three important factors for our metric:

1. **Meaning Similarity**: The generated summary and the reference summary should convey the same meaning irrespective of their language. Just like our alignment procedure from Section 2, we use LaBSE embeddings to compute the meaning similarity between the generated text ($s_{gen}$) and reference summary ($s_{ref}$):

$$MS(s_{gen}, s_{ref}) = \cos(\text{emb}(s_{gen}), \text{emb}(s_{ref}))$$

Here $\text{emb}(s)$ denotes the embedding vector output of LaBSE for input text $s$.

2. **Language Confidence**: The metric should identify, with high confidence, that the summary

| Input: $D_{ij} \forall i, j \in \{1, 2, \cdots, n\}$: training data with source/target languages $L_i/L_j$; $c_{ij} \leftarrow |D_{ij}| \forall i, j \in \{1, 2, \cdots, n\}$; $m$: number of mini-batches. |
| Compute $p_i, p_j, p_{ji}, p_{lj}$ using $c_{ij}$ |
| **while Model Not Covered do** |
| batch $\leftarrow \phi$ |
| Sample $r \sim \text{Unif}(0, 1)$ |
| if $r > 0.5$ then |
| Sample $L_i \sim p_i$ |
| for $i \leftarrow 1$ to $m$ do |
| Create mini-batch $mb$ from $D_{ij}$ |
| batch $\leftarrow$ batch $\cup \{mb\}$ |
| end |
| else |
| Sample $L_j \sim p_j$ |
| for $j \leftarrow 1$ to $m$ do |
| Create mini-batch $mb$ from $D_{ij}$ |
| batch $\leftarrow$ batch $\cup \{mb\}$ |
| end |
| Optimize model using batch |
| end |

**Algorithm 1**: A pseudocode of the multistage sampling algorithm.
We aim to train one model to generate summaries. As such, for this component, we use the Fasttext language-ID Classifer (Joulin et al., 2016) to obtain the language probability distribution of the generated summary and define the Language Confidence (LC) as:

\[
LC(s_{gen}, s_{ref}) = \begin{cases} 
1, & \text{if } L_{s_{ref}} = \arg\max P(L_{s_{gen}}) \\
P(L_{s_{gen}} = L_{s_{ref}}), & \text{otherwise}
\end{cases}
\]

3. Length Penalty: Generated summaries should not be unnecessarily long and the metric should penalize long summaries. While model-based metrics may indicate how similar a generated summary is to its reference and what its language is, it is not clear how they can be used to determine the brevity of a summary. As such, we adapt the BLEU (Papineni et al., 2002) brevity penalty to measure the length penalty of generated summaries:

\[
LP(s_{gen}, s_{ref}) = \begin{cases} 
1, & \text{if } |s_{gen}| \leq |s_{ref}| + c \\
\exp(1 - \frac{|s_{gen}|}{|s_{ref}| + c}), & \text{otherwise}
\end{cases}
\]

The languages of \(s_{gen}\) and \(s_{ref}\) may not be the same and identical texts may vary in length across languages. Hence, we used a length offset \(c\) to avoid penalizing generated summaries slightly longer than the references. By examining the standard deviation of mean summary lengths of the languages, we set \(c = 6\).

We finally define our metric, Language-agnostic Summary Similarity (LaSE) as:

\[
LaSE(S_{s_{gen}}, S_{s_{ref}}) = MS(s_{gen}, s_{ref}) \times LC(s_{gen}, s_{ref}) \times LP(s_{gen}, s_{ref})
\]

4 Experiments & Benchmarks

We aim to train one model to generate summaries in any target language for an input article from another language by providing them explicit cross-lingual supervision. Fine-tuning pretrained language models (Devlin et al., 2019; Xue et al., 2021) have shown state-of-the-art results on monolingual and multilingual abstractive text summarization (Rothe et al., 2020; Hasan et al., 2021). Many pretrained multilingual generation models are currently available, some of the prominent ones being mBART (Tang et al., 2020), CRISS (Tran et al., 2020), mT5 (Xue et al., 2021). Among them, we chose the mT5 model for fine-tuning because of its broad coverage of 101 languages with a support for 41 languages from CrossSum. Though CRISS is pretrained with a cross-lingual objective in contrast to the multilingual objective of mT5, we choose not to use CRISS due to its limited coverage of languages.

As the baseline for comparison, we use a summarize-then-translate pipeline. We use the mT5 checkpoint available on Hugging Face\(^3\) fine-tuned on the XL-Sum multilingual summarization dataset to first summarize the source article. Then we use the M2M-100 model (Fan et al., 2021) to translate the summary into the target language.

We fine-tune the mT5-base model with batch size 256, mini-batch size 32 (i.e. 8 mini-batches in one batch) on CrossSum (and also XL-Sum) with multistage sampling factors \(\alpha = 0.5\), \(\beta = 0.75\) for 25k steps on 8 Nvidia Tesla P100 GPUs for 3 days. We discard a language-pair from consideration if it has fewer than 32 training samples in order not to have duplicates in a mini-batch. We limit the input length to 512 tokens and output to 84 tokens. We use language-specific start tokens (Wu et al., 2016) for guiding the decoder to generate summaries in the intended target language. During inference, we use the language-specific start tokens for decoding and use a length penalty of 0.6. We use ROUGE and LaSE as the evaluation metrics.

4.1 Evaluation with Different Pivots

Cross-lingual text generation models are typically trained with one or more pivot languages which either occur in either the source or the target side of all training samples. This essentially makes the models perform better on those languages with others achieving sub-par performance (Johnson et al., 2017b). In this work, we make no such distinction on the choice of the pivots during fine-tuning and only rely on the multistage sampling technique for choosing the source and target languages. To assess how the model performs on non-English languages, in addition to English, two other languages: Hindi (hi) and Russian (ru) are benchmarked. For each of these languages, we pick five languages that have the highest number of test samples with them, and evaluate their performance on cross-lingual summarization when they are both on the source side.

\(^3\)https://huggingface.co/csebuetnlp/mT5_multilingual_XLSum
Figure 3: ROUGE-2 and LaSE scores for different pivot languages. Scores indicate that out many-to-many (m2m) significantly outperforms the baseline model on most languages. The remaining comparisons are shown in the appendix.
or the target side. We show the results in Figure 4.

5 Conclusion & Future Works

In this paper, we present CrossSum, a large-scale, non-English-centric cross-lingual abstractive summarization dataset, containing 1.65 million samples across 1500+ language pairs. For many of these pairs, CrossSum provides the first publicly available cross-lingual summarization dataset and benchmarks. We also make the alignment scripts available for the researchers, which will help to produce better alignments in future. We introduced a new multistage sampling algorithm that can be generalized to any cross-lingual generation task and a new language-agnostic metric for evaluating cross-lingual summaries when references in the target languages may not be available. Additionally, we demonstrate that training one multilingual model can help towards better cross-lingual summarization than baselines. Moreover, CrossSum is also useful for low-resource languages in zero- and few-shot cross-lingual transfer.

In future, we will investigate the use of our dataset for other summarization tasks, e.g., multi-document (Fabbri et al., 2019) and multi-modal summarization (Zhu et al., 2018).

Acknowledgements

This work was performed using the OzSTAR national facility at the Swinburne University of Technology. The OzSTAR program receives funding in part from the Astronomy National Collaborative Research Infrastructure Strategy (NCRIS) allocation provided by the Australian Government.

References

Mikel Artetxe and Holger Schwenk. 2018. Margin-based parallel corpus mining with multilingual sentence embeddings. arXiv preprint arXiv:1811.01136.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. Transactions of the Association for Computational Linguistics, 7:597–610.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8440–8451, Online. 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, USA. Association for Computational Linguistics.

Angelina 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. Journal of Machine Learning Research, 22(107):1–48.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. arXiv preprint arXiv:2007.01852.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XLsum: Large-scale multilingual abstractive summarization for 44 languages. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 4693–4703, Online. Association for Computational Linguistics.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017a. Billion-scale similarity search with gpus. arXiv preprint arXiv:1702.08734.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017b. Google’s multilingual neural machine translation system: Enabling zero-shot translation. Transactions of the Association for Computational Linguistics, 5:339–351.
Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017), page 6000–6010, Long Beach, California, USA.

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.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google’s neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 483–498, Online. Association for Computational Linguistics.

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

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. arXiv preprint arXiv:1909.02622.

Junnan Zhu, Haoran Li, Tianshang Liu, Yu Zhou, Ji-ajun Zhang, and Chengqing Zong. 2018. Msomo: Multimodal summarization with multimodal output. In Proceedings of the 2018 conference on empirical methods in natural language processing, pages 4154–4164.

Junnan Zhu, Qian Wang, Yining Wang, Yu Zhou, Ji-ajun Zhang, Shaoman 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.

Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. 2017. Overview of the second bncu shared task: Spotting parallel sentences in comparable corpora. In Proceedings of the 10th Workshop on Building and Using Comparable Corpora, pages 60–67.
Appendix
Figure 4: The remaining comparisons on ROUGE-2 and LaSE.
Table 1: An article-summary statistics of the CrossSum dataset containing a total of 1,645,226 cross-lingual samples. The rows indicate the articles’ language and columns that of their summaries’. For example, the cell on the second column of the fourth row indicates the number of samples where the article is in Bengali and the summary in Arabic.