End-to-End Cross-Lingual Summarization with Pre-training

Tony Ding Yixuan Huang Cindy Lin
Massachusetts Institute of Technology
Cambridge, MA, USA
{tding,yixuanhu,cindylin}@mit.edu

Abstract

Cross-lingual summarization (CLS) generates summaries in a target language from texts in a source language, serving as a valuable tool in cross-cultural and global communications. Traditionally achieved through a pipeline of machine translation (MT) and monolingual summarization (MS), CLS tasks are increasingly being explored with various end-to-end mode frameworks without the intermediate outputs. Building on the latest work on end-to-end CLS, we adopt two different approaches to integrating multilingual pre-training into generating Chinese summaries from English source texts: 1) using pre-trained multilingual embeddings only from the translation model mBART, and 2) fine-tuning the entire mBART model on CLS data. The second approach significantly outperforms the baseline model and the first approach, achieving a ROUGE-1 score of 22.37 on the in-distribution evaluation data and 3.82 on an external dataset. These scores surpass the baseline transformer model by 932% and 5,357%, and the first model by 1,002% and 3,720%, respectively.

1 Introduction

Cross-lingual summarization (CLS) refers to the task of generating summaries in a target language from longer texts in a source language. The existence of language barriers present issues in many settings, including transfer of knowledge or communication in both academic or personal settings. Without adequate translation tools, individuals may struggle to access or understand information, particularly those restricted to niche language groups. CLS addresses many of these issues and can be extremely valuable in various contexts. For example, scholars can comprehend a broad set of foreign academic work more efficiently, professionals can glean insights from international reports, and individuals can quickly digest global information without language barriers. This can help information sharing and knowledge distillation.

Traditionally, the task of CLS can be approached through sequential steps involving translation and summarization. Namely, previous methods have explored two sequential frameworks: translate then summarize or summarize then translate. However, these methods can cause inefficiencies, accumulate errors, and create more barriers for those without the expertise to effectively execute both tasks. Therefore, end-to-end CLS models can potentially improve efficiency and accuracy in information synthesis across languages.

1.1 Pipeline Methods

With the proliferation of research and advancement in neural models as well as the increasing availability of large pre-trained models, cross-lingual summarization has attracted more attention in the past few years. Early works focused on breaking down the task into a pipeline of two sub-tasks: machine translation (MT) and monolingual summarization (MS). For example, under the “MS-MT” approach, model-generated summaries are translated by MT service tools (Orăsan and Chiorean, 2008; Wan, 2011). Conversely, the "MT-MS" pipeline, which has garnered more interest over time, involves translating documents from one language to another and then summarizing the translated content. Within this paradigm, Wan (2011) translates English documents into Chinese, leveraging bilingual information for Chinese summary generation. Similarly, the approach of Boudin et al. (2011) involved the translation of English documents into French for subsequent summarization tasks.

1.2 End-to-End Methods

Instead of a pipeline method with two separate subtasks, Zhu et al. (2019) demonstrated the possibility of using an end-to-end, transformer-based model for this task. While this work is limited to summarizing short texts, recent work has expanded to long document summarization (Zheng et al., 2023), extreme, TLDR-style summarization (Cachola et al., 2020), and summary of academic articles (Cachola et al., 2020).

1.3 Multilingual Pre-Training

Pre-training, the process of training a model on a large, general dataset before fine-tuning it on specific tasks, has shown to increase performance in a variety of language tasks, such as question answering (Zhu et al., 2019) and sentiment analysis (Peters et al., 2018). Multilingual pre-training models, in particular, are equipped with knowledge and understanding of many languages and can thus be used for CLS tasks. For example, Xu et al. (2020) pre-train a cross-lingual masked language model (CMLM) then fine-tune it on a CLS corpus. In what follows, we discuss mBART (Liu et al., 2020), a multilingual pre-training model for Seq2Seq task that has been explored recently for CLS tasks.

mBART is a Seq2Seq denoising auto-encoder pre-trained on massive monolingual corpora in multiple languages using the BART objective (Lewis et al., 2019; Liu et al., 2020). Noises are added to the input texts by masking phrases and permuting sentences, and a bidirectional encoder along with an autoregressive decoder are learned to recover the original texts. mBART significantly improved both supervised and unsupervised MT performance at both the sentence level and the document level (Liu et al., 2020).

1.4 Our Contributions

Following the end-to-end framework in Zhu et al. (2019), our work aims to develop unified models to accomplish the CLS task, with the aid of multilingual pre-training. Specifically, we propose and discuss two approaches: a representation model using transformer backbone with mBART embeddings and a fine-tuned mBART.

Although recent work on incorporating mBART into CLS has shown great promise in improving CLS performance (Tang et al., 2021), there is a noticeable gap in the literature regarding the comparative analysis of different pre-training utilization strategies in CLS. Our research hopes to contribute to bridging this gap by exploring and comparing such strategies. We hypothesize that leveraging mBART's pre-trained multilingual representations will enhance the model's ability to understand and generate summaries across different languages. Additionally, we aim to assess if a fine-tuning approach on the pre-trained model could provide benefits.

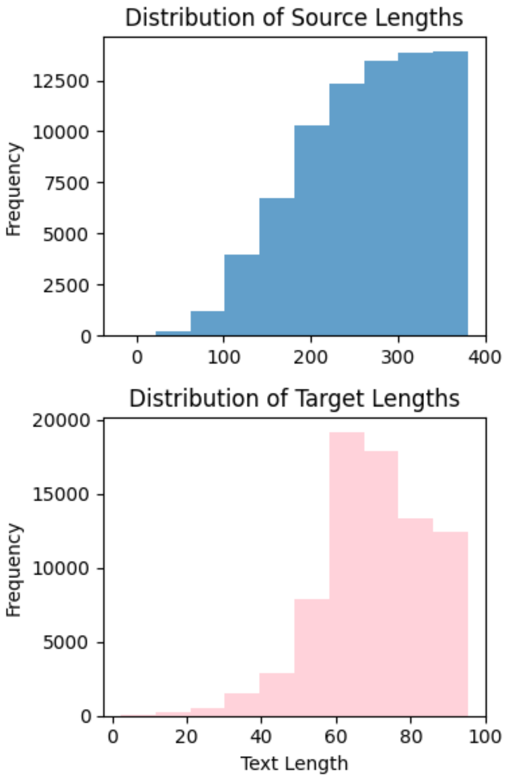

Figure 1: Text lengths of EN2ZHSUM subset

2 Methodology and Experimental Setup

2.1 Data and Preprocessing

Data: We train all our models on the *EN2ZHSUM* dataset from Zhu et al. (2019) with paired English texts and Chinese summaries. The dataset was constructed by first obtaining news articles and their corresponding highlights as sources and targets for monolingual summarization, then using a "round-trip" strategy to translate the summaries to the target language then back to the source language. The translation results were then evaluated using ROUGE to only include the results with high quality in the final dataset. The corpus of EN2ZHSUM contains 370,759 CLS pairs.

For model evaluation, in addition to the test set of EN2ZHSUM, we also use *ClidSum* (Wang et al., 2022) to assess the models' suitability to be generalized to out-of-distribution dataset. Vastly different from EN2ZHSUM, ClidSum consists of more than 67,000 dialogues and 112,000 annotated summaries. As a result, ClidSum presents an alternate text medium to assess the performance of our models on general CLS tasks. Due to the issue of data availability and the constraint of computing resources, for evaluation we only use 1,000 samples from *XMediaSum40k*, a subset of ClidSum that contains media interviews and their summaries.

Preprocessing: To account for computational constraints and ensure consistency across all models, we select the subset of EN2ZHSUM where the article text lengths are under 400 English words and summary text lengths 100 Chinese characters for training, validation, and test sets. The distribution of text lengths after subsetting is shown in Figure 1.

Additionally, the remaining source sequences in the data are then truncated to a length of at most 200 words, following (Zhu et al., 2019)'s approach to keep training more tractable. The sequences are then tokenized with an mBART Tokenizer, which is capable of tokenizing both English source texts and Chinese target texts.

Our final dataset contains 75,886 training samples, 644 evaluation samples, and 652 test samples.

2.2 Objectives

For a corpus of CLS data $C = (X, Y)$, where X is the set of texts in the source language and Y is the set of corresponding summaries in the target language, we aim to maximize the likelihood:

$$L = \prod_t p(y_t | y_{<t}, x, \theta).$$

2.3 Baseline Transformer Model

Our baseline model adopts the Transformer (Vaswani et al., 2017) framework, originally designed for sequence-to-sequence translation tasks with strong performance across various language modeling tasks. The baseline transformer consists of an encoder and a decoder each with 6 attention blocks. The embedding dimension is 250,054 and the model dimension is 1,024. The encoder of the model processes English articles first through embedding and positional encoding, and then self-attention and feed-forward layers. The decoder generates summaries in Chinese using cross-attention mechanisms.

2.4 Representation Model

Building on the baseline, we propose our first modeling approach for better cross-lingual understanding and summarization: a Transformer with mBART's word representations (Representation Model). Specifically, we replace both the input and output embedding layers and positional encoding layers of our baseline model with those of mBART, while other parts of the architecture remain the same. We then train this model using the same set of preprocessed CLS data, while keeping the mBART layers frozen. With this design, we hope to investigate if multilingual word representations from pre-training alone improves model performance. Compared to the baseline model where the embedding layers are initialized randomly and trained from scratch, the Representation Model's usage of mBART's pre-trained embedding weights may exhibit better language comprehension abilities from learning from large amounts of natural language.

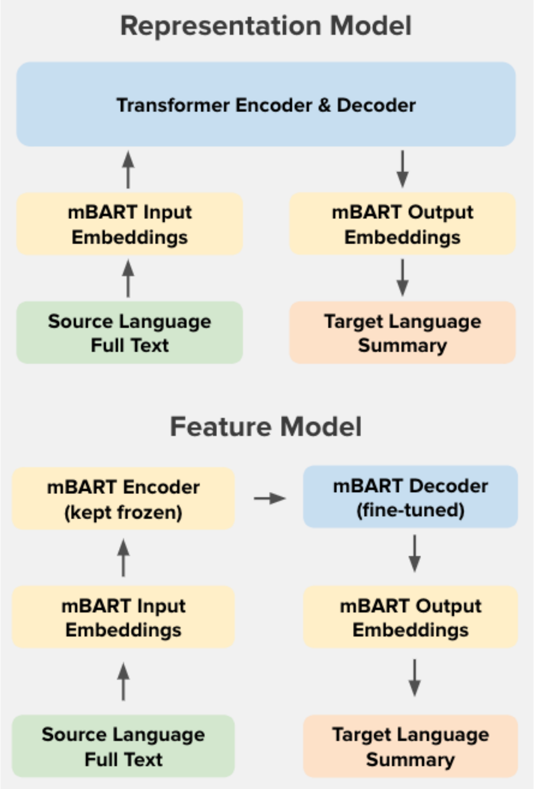

Figure 2: Framework designs for our two models

2.5 Feature Model

The Feature Model further integrates multilingual pre-training into CLS. In this architectural design, we fine-tune mBART for the CLS task on our training data. Specifically, we fine-tune the linear layer of mBART to adapt it to the requirements of summarization so that it can generate concise and accurate summaries. Additionally, we also unfreeze two out of ten cross-attention blocks of the mBART decoder. This allows the decoder to recalibrate and learn attention weights with regards to the source

Model	EN2ZHSUM				ClidSum			
	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum
Baseline	2.20	0.76	2.24	2.18	0.07	0.00	0.07	0.07
Representative	2.06	0.23	2.03	2.03	0.10	0.01	0.10	0.10
Feature	22.70	6.11	22.33	22.38	3.82	0.50	3.69	3.74

Table 1: Model performances on test datasets. We use 652 test samples from EN2ZHSUM and 1,000 test samples from ClidSum for evaluation.

texts that are more suited for summarization tasks, which may differ from those required in translation tasks. By doing so, the model is expected to gain a more nuanced understanding of the text and facilitate better content selection and synthesis.

2.6 Evaluation Metrics

Given that we want to evaluate the text summarizations generated by our models, we thus employ ROUGE, which is a set of metrics commonly used for the automatic evaluation of machine-generated text, particularly in the context of text summarization (Lin, 2004). ROUGE focuses on measuring content similarity and is relatively easy to interpret, as a higher ROUGE score generally indicates better content overlap between the generated text and the reference text. For our work. we employ the following ROUGE metrics for performance evaluation (Lin, 2004):

- ROUGE-1: measures the overlap of unigrams between the generated summary and the reference summary.

- ROUGE-2: measures the overlap of bigrams between the generated summary and the reference summary.

- ROUGE-L: measures the longest common subsequence (LCS) of words between the generated summary and the reference summary. ROUGE-L takes into account the order of words and is more lenient in terms of word overlap. ROUGE-L also rewards longer shared sequences between the generated and reference summaries.

- ROUGE-Lsum: applies the same calculation method as that of ROUGE-L, but at the sentence level. ROUGE-Lsum is more suitable for evaluating tasks where sentence level extraction is valuable, such as extractive summarization tasks.

2.7 Implementation Details

We train all models mainly using NVIDIA A100 GPU with 40G VRAM for 200,000 steps with a batch size of 32. We use the AdamW optimizer ($\beta_1 = .9, \beta_2 = .998$).

3 Results and Analysis

Evaluation results on the test set of EH2ZHSUM and the external dataset ClidSum are presented in Table 1 above. We also include sample summarization output in Chinese from all models in Table 2, along with English translations provided for reference. Overall, the Feature Model outperforms the other two significantly, based on both the quantitative and qualitative results.

3.1 Evaluation Results

EN2ZHSUM: On the EN2ZHSUM test set, the Representative Model performs slightly better than Baseline on three of the four metrics, while the Feature Model performs better than both the baseline and the Representation Model by large margins on all metrics. The Baseline Model lacks the extensive pre-training on diverse datasets that mBART offers. Without pre-training, the Baseline Model starts with no prior knowledge and must learn language representations from the ground up. This process is inherently less efficient and can result in a model that is less adept at handling the nuances of language, particularly in a cross-lingual context.

In the Representation Model, replacing the embedding layers with those from mBART is shown to be useful. The pre-trained mBART is equipped with a much more refined and nuanced understanding of syntax and semantics across different languages, including the ones in our source and target texts. This knowledge then helps better capture the essence of texts for summarization tasks. However, since only the embedding and positional encoding layers are replaced, the rest of the model still needs to adapt and learn how to effectively use these embeddings for the task of summarization.

	EN2ZHSUM	ClidSum
Ground Truth in Chinese (Translation for Reference)	一家汽车公司发起了对假冒气囊零件的搜索。丰田“非常担心假气囊的安全”。该公司今年早些时候召回了5种不同型号的汽车。 (A car company has launched a search for counterfeit air bag parts. Toyota is "very concerned about the safety of fake air bags". The company called back 5 different models of cars earlier this year.)	科林·凯珀尼克与NFL达成了协议。体育社会学家哈里·爱德华兹接受NPR记者麦克·马丁,讨论这对NFL抗议者的未来意味着什么。 (Colin Kaepernick reached an agreement with NFL. Sports sociologist Harry Edwards talked to NPR reporter Mike Martin about what this meant for the future of NFL protesters.)
Baseline Model Output (Translation)	周六,一名妇女在车祸中丧生。该地区在车祸 (On Saturday, a woman died in a car accident. The area is car accident)	这位前锋在周六晚上的家中被发现。他被指控在推特 (The striker was discovered at his home on Saturday night. He was accused on Twitter)
Representation Model Output (Translation)	一辆汽车在一辆汽车撞上一辆汽车后被发现。汽车在一辆汽车 (A car after a car crashed into another car is discovered. Car at a car)	这位前前前前前足球运动员在推特上发布了一段视频。前足球 (The former former former former former soccer player posted a video on Twitter. Former soccer)
Feature Model Output (Translation)	汽车公司发起了一项搜索,以找出假冒的假气囊零件。据信这些零件是由丰田经销商出售的,并可能对司机造成严重风险。 (The car company launched a search to find counterfeit fake air bag parts. The parts are believed to be sold by Toyota dealers and could pose serious risks to drivers.)	科林·卡佩尼克与美国国家橄榄球联盟签署了保密协议。这位前49人四分卫在2016年国歌上膝盖受伤引发了抗议。 (Colin Kaepernick signed a non-disclosure agreement with the National Football League. The former 49ers quarterback sparked protests when he injured his knee during the national anthem in 2016.)

Table 2: Sample ground truth and model generated outputs for the two evaluation datasets EN2ZHSUM and CLIMDSUM. The literal English translations of the Chinese summaries are provided for reference. Grammatical and punctuation errors in the English translations reflect those in the generated Chinese summaries.

The Feature Model takes the integration of mBART a step further by utilizing not just the embeddings but also the pre-trained weights from most of its encoder and decoder layers. The much-improved performance compared to the Representation Model demonstrates the benefits of additionally incorporating mBART's learned-ability for translation through loading the pre-trained model weights. While embeddings provide the foundational representation of words, subwords, and symbols and multilingual embeddings further possess the capacity of encapsulating both languages into the same space, they are only a small portion of the architecture that contributes to mBART's overall success in translation. As a result, using more parameters from mBART provides better translation ability which improves the quality of cross-lingual summarization.

We note that ROUGE-2 scores are consistently lower for all models, across both test datasets. This difference in performance is likely due to the fact that ROUGE-2 measures the overlap of bigrams between the generated summary and the ground truth summary. Although the generated translations from the feature model are semantically accurate, the words generated do not match the ground truth exactly. Alternately, ROUGE-1 considers unigrams and thus suffers less since one-word matches are generally more likely. ROUGE-L and ROUGE-Lsum are more lenient in terms of word overlap and hence yields higher values than ROUGE-2.

ClidSum: Evaluating our model's performance on an external dataset, ClidSum, we observe that the representation model marginally outperforms the baseline model across all four metrics. Additionally, though the feature model achieves much higher performance on all metrics for the ClidSum dataset, its performance remains poor in contrast to model evaluation on EN2ZHSUM. The performance of all three models on ClidSum is noticeably worse than its EN2ZHSUM counterparts. We note that this difference is likely a result of the nature of the text input, as ClidSum involves dialogues whereas EN2ZHSUM primarily includes news articles.

3.2 Sample Outputs

Qualitative assessment of the model outputs from the test sets further demonstrates Feature Model's enhanced performance compared to the other models. As shown in Table 2, the summaries generated by the Feature Model for both the EN2ZHSUM and ClidSum samples are more accurate and coherent. In contrast, the Baseline and Representation Models either fail to produce full, sensible sentences, or are unable to capture the essence of the original texts.

A common error by these two models is the unnecessary repetitions of words and phrases, as exemplified by the sample outputs for ClidSum.

Additionally, as expected, the Baseline and Representation Model's translation and language understanding skills are worse than that of the Feature Model and are more prone to errors. In the sample ClidSum outputs generated by the Baseline and the Representation Model, an American football player is incorrectly referred to as a soccer player. This reflects not only the models' inferior translation abilities but also their lack of context understanding, as the source text contains discussion of a former quarterback of an American football team. The Feature Model, however, successfully captures this context.

3.3 Ablation Study

We conduct an ablation study on the Feature Model based on the training sample size. The evaluation results are presented in Figure 3 below. Generally, model performance improves with increasing training sample sizes, although even a small sample of 1K training examples can achieve relatively decent performance. This ablation study is not conducted on the ClidSum dataset. Given that the model, when trained with the full dataset of more than 75K examples, demonstrates sub-optimal performance on ClidSum, it's deemed less meaningful to assess its performance with smaller training sizes.

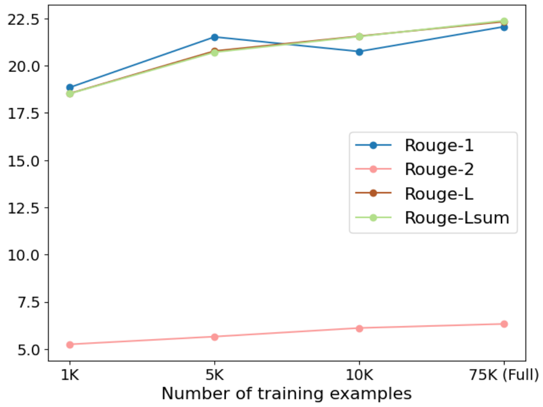

Figure 3: Evaluation results on EN2ZHSUM using Feature Model with various training sample sizes.

4 Discussion

Our findings demonstrate that the efficacy of models in cross-lingual summarization is significantly enhanced by multilingual pre-training. In particular, employing a greater number of pre-trained weights from many layers of a multilingual model results in a much higher degree of success compared to utilizing only the embeddings. This highlights the importance of knowledge transfer in multilingual pre-training for improving the performance of cross-lingual summarization. By leveraging the comprehensive linguistic knowledge acquired during pre-training, these models can achieve a higher level of success in generating accurate, relevant, and contextually appropriate summaries across different languages.

4.1 Limitations

There are several limitations to our work. Due to computational resource constraints, we conducted training on only a small subset of EN2ZHSUM and truncated each article to 200 words. This hinders our models' ability to learn from a larger amount of data and likely has a negative impact on model performance. In addition, not using the full set of EN2ZHSUM prevents us from performing comprehensive comparisons between our results and other relevant work that also uses the same data. Although we explore the advantages of using word representations and fine-tuning on a multilingual pre-trained model in the CLS task, we are unable to draw conclusions on the best approach applied to broader, larger, and more diverse datasets.

This limitation is also reflected in our model's subpar performance on the ClidSum dataset, with all ROUGE scores associated with ClidSum being drastically lower than those obtained when the model is evaluated on EN2ZHSUM.

Furthermore, the choice of our evaluation metrics poses yet another limitation, as ROUGE scores focus exclusively on lexical overlap and are unable to effectively capture semantic nuances. To elaborate, ROUGE score penalizes outputs that use synonyms or paraphrasing, as it relies heavily on exact word matches or overlaps. In addition, ROUGE is sensitive to the length of generated summary, as longer summaries may achieve higher ROUGE scores even if they contain more redundant or less relevant information, thus leading to potential bias in evaluation.

4.2 Future Directions

For future work, we can employ more well-rounded metrics like BERTScore, which measures the similarity between generated and reference sentences using contextual embeddings from pre-trained BERT models (Zhang et al., 2019). As a result, BERTscore, compared to ROUGE scores, will be much better at capturing the semantic similarity as well as less sensitive to word order variations, thus making it a good evaluation metric to employ. BARTScore is another metric that can capture the fluency and coherence of the generated text (Yuan et al., 2021), which also makes it a good candidate as an evaluation metric.

Furthermore, future work can focus on datasets with other language pairs, especially for low-resource languages or for pairs that are not within the same language family such as Korean and German, Swahili and Spanish, and Turkish and French.

Another potential future direction could be to train our model on multiple target languages and summarization lengths so that the model can produce summaries in different languages or of different lengths as desired.

References

Florian Boudin, Stéphane Huet, and Juan-Manuel Torres-Moreno. 2011. A graph-based approach to cross-language multi-document summarization. *Polibits*, (43):113–118.

Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel S. Weld. 2020. Tldr: Extreme summarization of scientific documents.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

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.

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.

Constantin Orăsan and Oana Andreea Chiorean. 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. European Language Resources Association (ELRA).

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Xiaojun Wan. 2011. Using bilingual information for cross-language document summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1546–1555, Portland, Oregon, USA. Association for Computational Linguistics.

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

Ruochen Xu, Chenguang Zhu, Yu Shi, Michael Zeng, and Xuedong Huang. 2020. Mixed-lingual pre-training for cross-lingual summarization. *arXiv preprint arXiv:2010.08892*.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277.

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

Shaohui Zheng, Zhixu Li, Jiaan Wang, Jianfeng Qu, An Liu, Lei Zhao, and Zhigang Chen. 2023. Long-document cross-lingual summarization. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, WSDM '23, page 1084–1092, New York, NY, USA. Association for Computing Machinery.

Junnan Zhu, Qian Wang, Yining Wang, Yu Zhou, Jiajun Zhang, Shaonan Wang, and Chengqing Zong. 2019. Ncls: Neural cross-lingual summarization. *arXiv preprint arXiv:1909.00156*.

A Impact Statement

Accurate and efficient cross-lingual summarization (CLS) techniques can improve accessibility to information and knowledge for a wide range of audiences. It aids in understanding news, books, reports, academic papers, and other content in unfamiliar languages by providing translated summaries and thus help reducing language barriers in knowledge distillation.

CLS can be especially valuable in fields like medicine, where effective information exchange, retrieval, and synthesis across different populations and languages is crucial. For example, some rare diseases may have only been studied and documented in small sub-populations with low-resource languages. By enabling efficient translations and summarizations of medical findings, CLS can facilitate collaboration and advancements which can be otherwise expensive and sparse.

However, potential harm that comes with efficient CLS should not be ignored. Like most language modeling tasks, CLS faces the common risk of learning and perpetuating harmful content present in natural language datasets. This content can become embedded in the model's knowledge base and reasoning mechanisms, leading to the generation of biased or inappropriate outputs. Moreover, an additional challenge unique to CLS is the loss of contexts and nuance from compressing rich, detailed information into short and sometimes oversimplified summaries. This could lead to summaries that might misrepresent the source material or fail to convey its true intent and tone, or provide incentives for readers to shy away from deeper understanding of subjects, therefore contradicting the objective of enhancing knowledge sharing.