Mixed-Lingual Pre-training for Cross-lingual Summarization

Ruochen Xu*, **Chenguang Zhu***, **Yu Shi, Michael Zeng, Xuedong Huang**
Microsoft Cognitive Services Research Group
{ruox,chezhu,yushi,nzeng,xdh}@microsoft.com

Abstract

Cross-lingual Summarization (CLS) aims at producing a summary in the target language for an article in the source language. Traditional solutions employ a two-step approach, i.e. translate→summarize or summarize→translate. Recently, end-to-end models have achieved better results, but these approaches are mostly limited by their dependence on large-scale labeled data. We propose a solution based on mixed-lingual pre-training that leverages both cross-lingual tasks such as translation and monolingual tasks like masked language models. Thus, our model can leverage the massive monolingual data to enhance its modeling of language. Moreover, the architecture has no task-specific components, which saves memory and increases optimization efficiency. We show in experiments that this pre-training scheme can effectively boost the performance of cross-lingual summarization. In Neural Cross-Lingual Summarization (NCLS) (Zhu et al., 2019b) dataset, our model achieves an improvement of 2.82 (English to Chinese) and 1.15 (Chinese to English) ROUGE-1 scores over state-of-the-art results.

1 Introduction

Text summarization can facilitate the propagation of information by providing an abridged version for long articles and documents. Meanwhile, the globalization progress has prompted a high demand of information dissemination across language barriers. Thus, the cross-lingual summarization (CLS) task emerges to provide accurate gist of articles in a foreign language.

Traditionally, most CLS methods follow the two-step pipeline approach: either translate the article into the target language and then summarize it (Leuski et al., 2003), or summarize the article in the source language and then translate it (Wan et al., 2010). Although this method can leverage off-the-shelf summarization and MT models, it suffers from error accumulation from two independent subtasks. Therefore, several end-to-end approaches have been proposed recently (Zhu et al., 2019b; Ouyang et al., 2019; Duan et al., 2019), which conduct both translation and summarization simultaneously. Easy to optimize as these methods are, they typically require a large amount of cross-lingual summarization data, which may not be available especially for low-resource languages. For instance, NCLS (Zhu et al., 2019b) proposes to co-train on monolingual summarization (MS) and machine translation (MT) tasks, both of which require tremendous labeling efforts.

On the other hand, the pre-training strategy has proved to be very effective for language understanding (Devlin et al., 2018; Holtzman et al., 2019) and cross-lingual learning (Lample and Conneau, 2019; Chi et al., 2019). One of the advantages of pre-training is that many associated tasks are self-learning by nature, which means no labeled data is required. This greatly increases the amount of training data exposed to the model, thereby enhancing its performance on downstream tasks.

Therefore, we leverage large-scale pre-training to improve the quality of cross-lingual summarization. Built upon a transformer-based encoder-decoder architecture (Vaswani et al., 2017), our model is pre-trained on both monolingual tasks including masked language model (MLM), denoising autoencoder (DAE) and monolingual summarization (MS), and cross-lingual tasks such as cross-lingual masked language model (CMLM) and machine translation (MT). This mixed-lingual pre-training scheme can take advantage of massive unlabeled monolingual data to improve the model's language modeling capability, and leverage cross-lingual tasks to improve the model's cross-lingual representation. We then finetune the model on the

* Equal contribution

downstream cross-lingual summarization task.

Furthermore, based on a shared multi-lingual vocabulary, our model has a shared encoder-decoder architecture for all pre-training and finetuning tasks, whereas NCLS (Zhu et al., 2019b) sets aside task-specific decoders for machine translation, monolingual summarization, and cross-lingual summarization.

In the experiments, our model outperforms various baseline systems on the benchmark dataset NCLS (Zhu et al., 2019b). For example, our model achieves 3.27 higher ROUGE-1 score in Chinese to English summarization than the state-of-the-art result and 1.28 higher ROUGE-1 score in English to Chinese summarization. We further conduct an ablation study to show that each pretraining task contributes to the performance, especially our proposed unsupervised pretraining tasks.

2 Related Work

2.1 Pre-training

Pre-training language models (Devlin et al., 2018; Dong et al., 2019) have been widely used in NLP applications such as question answering (Zhu et al., 2018), sentiment analysis (Peters et al., 2018), and summarization (Zhu et al., 2019a; Yang et al., 2020). In multi-lingual scenarios, recent works take input from multiple languages and shows great improvements on cross-lingual classification (Lample and Conneau, 2019; Pires et al., 2019; Huang et al., 2019) and unsupervised machine translation (Liu et al., 2020). Artetxe and Schwenk (2019) employs the sequence encoder from a machine translation model to produce cross-lingual sentence embeddings. Chi et al. (2019) uses multi-lingual pre-training to improve cross-lingual question generation and zero-shot cross-lingual summarization. Their model trained on articles and summaries in one language is directly used to produce summaries for articles in another language, which is different from our task of producing summaries of one language for an article from a foreign language.

2.2 Cross-lingual Summarization

Early literatures on cross-lingual summarization focus on the two-step approach involving machine translation and summarization (Leuski et al., 2003; Wan et al., 2010), which often suffer from error propagation issues due to the imperfect modular systems. Recent end-to-end deep learning models have greatly enhanced the performance. Shen et al. (2018) presents a solution to zero-shot cross-lingual headline generation by using machine translation and summarization datasets. Duan et al. (2019) leverages monolingual abstractive summarization to achieve zero-shot cross-lingual abstractive sentence summarization. NCLS (Zhu et al., 2019b) proposes a cross-lingual summarization system for large-scale datasets for the first time. It uses multi-task supervised learning and shares the encoder for monolingual summarization, cross-lingual summarization, and machine translation. However, each of these tasks requires a separate decoder. In comparison, our model shares the entire encoder-decoder architecture among all pre-training and finetuning tasks, and leverages unlabeled data for monolingual masked language model training. A concurrent work by Zhu et al. (2020) improves the performance by combining the neural model with an external probabilistic bilingual lexicon.

3 Method

3.1 Pre-training Objectives

We propose a set of multi-task pre-training objectives on both monolingual and cross-lingual corpus. For monolingual corpus, we use the masked language model (MLM) from Raffel et al. (2019). The input is the original sentence masked by sentinel tokens, and the target is the sequence consists of each sentinel token followed by the corresponding masked token. The other monolingual task is the denoising auto-encoder (DAE), where the corrupted input is constructed by randomly dropping, masking, and shuffling a sentence and the target is the original sentence. Since our final task is summarization, we also include monolingual summarization (MS) as a pre-training task.

To leverage cross-lingual parallel corpus, we introduce the cross-lingual masked language model (CMLM). CMLM is an extension of MLM on the parallel corpus. The input is the concatenation of a sentence in language A and its translation in language B. We then randomly select one sentence and mask some of its tokens by sentinels. The target is to predict the masked tokens in the same way as MLM. Different from MLM, the masked tokens in CMLM are predicted not only from the context within the same language but also from their translations in another language, which encourages the model to learn language-invariant representations. Note that CMLM is similar to the Translation Language Model (TLM) loss proposed in Lample

Objective	Supervised	Multi-lingual	Inputs	Targets
Masked Language Model			France <X> Morocco in <Y> exhibition match.	<X> beats <Y> an
Denoising Auto-Encoder			France beats <M> in <M> exhibition .	France beats Morocco in an exhibition match.
Monolingual Summarization	✓		World champion France overcame a stuttering start to beat Morocco 1-0 in a scrappy exhibition match on Wednesday night.	France beats Morocco in an exhibition match.
Cross-lingual MLM	✓	✓	France <X> Morocco in <Y> exhibition match. 法国队在一场表演赛中击败摩洛哥队。	<X> beats <Y> an
Cross-lingual MLM	✓	✓	France beats Morocco in an exhibition match. <X>队在一场表演赛中<Y>摩洛哥队。	<X>法国<Y>击败
Machine Translation	✓	✓	France beats Morocco in an exhibition match.	法国队在一场表演赛中击败摩洛哥队。

Table 1: Examples of inputs and targets used by different objectives for the sentence "France beats Morocco in an exhibition match" with its Chinese translation. We use <X> and <Y> to denote sentinel tokens and <M> to denote shared mask tokens.

and Conneau (2019). The key differences are: 1) TLM randomly masks tokens in sentences from both languages, while CMLM only masks tokens from one language; 2) TLM is applied on encoder-only networks while we employ CMLM on the encoder-decoder network. In addition to CMLM, we also include standard machine translation (MT) objective, in which the input and output are the unchanged source and target sentences, respectively.

The examples of inputs and targets used by our pre-training objectives are shown in Table 1.

3.2 Unified Model for Pre-training and Finetuning

While NCLS (Zhu et al., 2019b) uses different decoders for various pre-training objectives, we employ a unified Transformer (Vaswani et al., 2017) encoder-decoder model for all pre-training and finetuning tasks. This makes our model learn a cross-lingual representation efficiently. A shared dictionary across all languages is used. To accommodate multi-task and multilingual objectives, we introduce language id symbols to indicate the target language, and task symbols to indicate the target task. For instance, for the CMLM objective where the target language is Chinese, the decoder takes <cmlm> and <zh> as the first two input tokens. We empirically find that our model does not suffer from the phenomenon of forgetting target language controllability as in Chi et al. (2019), which requires manual freezing of encoder or decoder during finetuning. After pretraining, we conduct finetuning on cross-lingual summarization data.

4 Experiments

4.1 Dataset

We conduct our experiment on NCLS dataset (Zhu et al., 2019b), which contains paired data of English articles with Chinese summaries, and Chinese articles with English summaries. The cross-lingual training data is automatically generated by a machine translation model. For finetuning and testing, we followed the same train/valid/test split of the original dataset. We refer readers to Table 1 in Zhu et al. (2019b) for detailed statistics of the dataset.

For pre-training, we obtain monolingual data for English and Chinese from the corresponding Wikipedia dump. There are 83 million sentences for English monolingual corpus and 20 million sentences for Chinese corpus. For parallel data between English and Chinese, we use the parallel corpus from Lample and Conneau (2019), which contains 9.6 million paired sentences. For monolingual summarization objective, we use CNN/DailyMail dataset (Nallapati et al., 2016) for English summarization and LCSTS dataset (Hu et al., 2015) for Chinese summarization.

4.2 Implementation Details

Our transformer model has 6 layers and 8 heads in attention. The input and output dimensions d_{model} for all transformer blocks are 512 and the inner dimension d_{ff} is 2048.

We use a dropout probability of 0.1 on all layers. We build a shared SentencePiece (Kudo and Richardson, 2018) vocabulary of size $33,000$ from a balanced mix of the monolingual Wikipedia corpus. The model has approximately 61M parameters.

For MLM we use a mask probability of 0.15. For DAE, we set both the mask and drop out rate

	English→Chinese			Chinese→English		
	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
TETran	26.15	10.60	23.24	23.09	7.33	18.74
GETran	28.19	11.40	25.77	24.34	9.14	20.13
TLTran	30.22	12.20	27.04	33.92	15.81	29.86
GLTran	32.17	13.85	29.43	35.45	16.86	31.28
NCLS	36.82	18.72	33.20	38.85	21.93	35.05
NCLS-MS	38.25	20.20	34.76	40.34	22.65	36.39
NCLS-MT	40.23	22.32	36.59	40.25	22.58	36.21
XNLG	39.85	24.47	28.28	38.34	19.65	33.66
ATS	40.68	24.12	**36.97**	40.47	22.21	36.89
Ours	**43.50**	**25.41**	29.66	**41.62**	**23.35**	**37.26**

Table 2: ROUGE-1, ROUGE-2, ROUGE-L for English to Chinese and Chinese to English summarization on NCLS dataset.

to 0.1. For all pre-training and finetuning we use RAdam optimizer (Liu et al., 2019) with $\beta_1 = 0.9$, $\beta_2 = 0.999$. The initial learning rate is set to 10^{-9} for pre-training and 10^{-4} for finetuning. The learning rate is linearly increased to 0.001 with $16,000$ warmup steps followed by an exponential decay. For decoding, we use a beam size of 6 and a maximum generation length of 200 tokens for all experiments.

	English→Chinese		
	ROUGE-1	ROUGE-2	ROUGE-L
Ours	43.50	25.41	29.66
- MS	42.48	24.45	28.49
- MT	42.12	23.97	28.74
- MLM, DAE	41.82	23.85	28.40
- All Pretraining	41.12	23.67	28.53

Table 3: Finetuning performance on English→Chinese summarization starting with various ablated pre-trained models.

4.3 Baselines

We first include a set of pipeline methods from Zhu et al. (2019b) which combines monolingual summarization and machine translation. **TETran** first translates the source document and then uses LexRank (Erkan and Radev, 2004) to summarize the translated document. **TLTran** first summarizes the source document and then translates the summary. **GETran** and **GLTran** replace the translation model in TETran and TLTran with Google Translator[1] respectively.

We also include three strong baselines from Zhu et al. (2019b): **NCLS**, **NCLS-MS** and **NCLS-MT**. NCLS trains a standard Transformer model on the cross-lingual summarization dataset. NCLS-MS and NCLS-MT both use one encoder and multiple decoders for multi-task scenarios. NCLS-MS combines the cross-lingual summarization task with monolingual summarization while NCLS-MT combines it with machine translation.

We finetune **XNLG** model from Chi et al. (2019) on the same cross-lingual summarization data. We finetune all layers of XNLG in the same way as our pretrained model.

Finally, we include the result of **ATS** from the concurrent work of Zhu et al. (2020).

4.4 Results

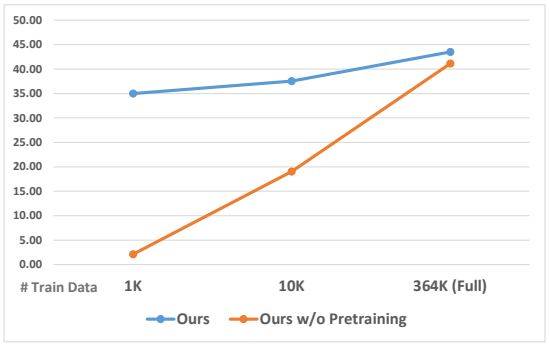
Table 2 shows the ROUGE scores of generated summaries in English-to-Chinese and Chinese-to-English summarization. As shown, pipeline models, although incorporating state-of-the-art machine translation systems, achieve sub-optimal performance in both directions, proving the advantages of end-to-end models.

Our model outperforms all baseline models in all metrics except for ROUGE-L in English-to-Chinese. For instance, our model achieves 2.82 higher ROUGE-1 score in Chinese to English summarization than the previously best result and 1.15 higher ROUGE-1 score in English to Chinese summarization, which shows the effectiveness of utilizing multilingual and multi-task data to improve cross-lingual summarization.

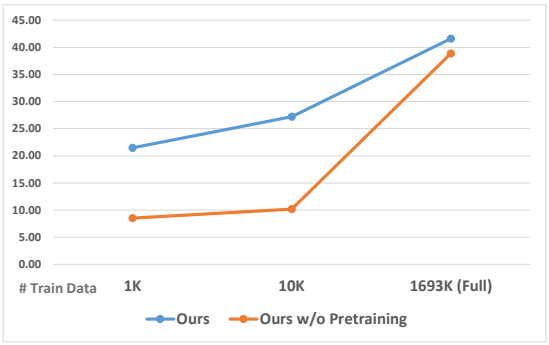
4.5 Ablation Study

Table 3 shows the ablation study of our model on English to Chinese summarization. We remove

[1] https://translate.google.com/



(a) English→Chinese ROUGE-1



(b) Chinese→English ROUGE-1

Figure 1: ROUGE-1 performance on NCLS dataset when the cross-lingual summarization training data is subsampled to size of 1k and 10k. The result on the full dataset is also shown.

from the pre-training objectives i) all monolingual unsupervised tasks (MLM, DAE), ii) machine translation (MT), iii) monolingual summarization (MS), and iv) all the objectives. Note that "- All Pretraining" and NCLS both only train on the cross-lingual summarization data. The performance difference between the two is most likely due to the difference in model size, vocabulary, and other hyperparameters.

As shown, the pre-training can improve ROUGE-1, ROUGE-2, and ROUGE-L by 2.38, 1.74, and 1.13 points respectively on Chinese-to-English summarization. Moreover, all pre-training objectives have various degrees of contribution to the results, and the monolingual unsupervised objectives (MLM and DAE) are relatively the most important. This verifies the effectiveness of leveraging unsupervised data in the pre-training.

Low-resource scenario. We sample subsets of size 1K and 10K from the training data of cross-lingual summarization and finetune our pre-trained model on those subsets. Figure 1 shows the the performance of the pre-trained model and the model trained from scratch on the same subsets. As shown, the gain from pre-training is larger when the size of training data is relatively small. This proves the effectiveness of our approach to deal with low-resource language in cross-lingual summarization.

5 Conclusion

We present a mix-lingual pre-training model for cross-lingual summarization. We optimize a shared encoder-decoder architecture for multi-lingual and multi-task objectives. Experiments on a benchmark dataset show that our model outperforms pipeline-based and other end-to-end baselines. Through an ablation study, we show that all pretraining objectives contribute to the model's performance.

References

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.

Zewen Chi, Li Dong, Furu Wei, Wenhui Wang, Xian-Ling Mao, and Heyan Huang. 2019. Cross-lingual natural language generation via pre-training. *arXiv preprint arXiv:1909.10481*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems*, pages 13063–13075.

Xiangyu Duan, Mingming Yin, Min Zhang, Boxing Chen, and Weihua Luo. 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.

Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22:457–479.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Baotian Hu, Qingcai Chen, and Fangze Zhu. 2015. Lcsts: A large scale chinese short text summarization dataset. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1967–1972.

Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. 2019. Unicoder: A universal language encoder by pretraining with multiple cross-lingual tasks. 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 2485–2494.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Anton Leuski, Chin-Yew Lin, Liang Zhou, Ulrich Germann, Franz Josef Och, and Eduard Hovy. 2003. Cross-lingual c* st* rd: English access to hindi information. *ACM Transactions on Asian Language Information Processing (TALIP)*, 2(3):245–269.

Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. 2019. On the variance of the adaptive learning rate and beyond. In *International Conference on Learning Representations*.

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. *arXiv preprint arXiv:2001.08210*.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Guclcehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290.

Jessica Ouyang, Boya Song, and Kathleen McKeown. 2019. A robust abstractive system for cross-lingual summarization. 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.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001.

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 preprint arXiv:1910.10683*.

Shi-qi Shen, Yun Chen, Cheng Yang, Zhi-yuan Liu, Mao-song Sun, et al. 2018. Zero-shot cross-lingual neural headline generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(12):2319–2327.

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 *Advances in neural information processing systems*, pages 5998–6008.

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. Association for Computational Linguistics.

Ziyi Yang, Chenguang Zhu, Robert Gmyr, Michael Zeng, Xuedong Huang, and Eric Darve. 2020. Ted: A pretrained unsupervised summarization model with theme modeling and denoising. *arXiv preprint arXiv:2001.00725*.

Chenguang Zhu, Ziyi Yang, Robert Gmyr, Michael Zeng, and Xuedong Huang. 2019a. Make lead bias in your favor: A simple and effective method for news summarization. *arXiv preprint arXiv:1912.11602*.

Chenguang Zhu, Michael Zeng, and Xuedong Huang. 2018. Sdnet: Contextualized attention-based deep network for conversational question answering. *arXiv preprint arXiv:1812.03593*.

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

Junnan Zhu, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2020. Attend, translate and summarize: An efficient method for neural cross-lingual summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1309–1321.