Exploring Cross-Lingual Guidance in Abstractive Summarization

Anonymous NAACL submission

Abstract

Cross-lingual guidance (CLG) as an augmentation method is often applied in cross-lingual summarization (CLS) to improve its performance. In this paper, we empirically explore how cross-lingual information of different quality benefits the encoding and decoding procedures for both cross-lingual and monolingual abstractive summarization (MLS). We specifically propose a summarization model DualSum which can utilize CLG in both encoding and decoding, and construct a dataset BiRead with high-quality parallel bilingual document-summary pairs. The empirical experiments will show how CLS and MLS are influenced by CLG. [1]

1 Introduction

With the development of machine translation (MT), the task of cross-lingual summarization (CLS) is proposed, aiming to generate a summary in another language different from the input document (Leuski et al., 2003). Initially, the techniques adopted are intuitive, first-translate-then-summarize or first-summarize-then-translate (Leuski et al., 2003; Orasan and Chiorean, 2008; Wan et al., 2010; Wan, 2011). Subsequently, the use of sequence-to-sequence (Seq2Seq) (Sutskever et al., 2014) techniques brought significant improvements over traditional pipeline methods with one language as input and another language as output (Duan et al., 2019). In this case, the translated documents or summaries often served as augmented data of a Seq2Seq model. This kind of cross-lingual guidance (CLG) has proven useful in CLS (Shen et al., 2018; Duan et al., 2019; Zhu et al., 2019; Cao et al., 2020; Bai et al., 2021).

In this paper, we would further explore the CLG problem: Can the translated text be flexibly used in both the encoder and decoder side of a Seq2Seq model to improve the performance of cross-lingual and monolingual abstractive summarization? In our opinion, for the CLS task, the parallel document-summary pairs can offer more information in guiding the generation of a precise summary. However, to the best of our knowledge, such cross-lingual guidance in CLS has only been discussed in the decoding process and still lacks a comprehensive review. Due to the maturity of MT techniques, the low-cost translation makes the exploration of both the encoding and decoding processes feasible. As a byproduct, provided with parallel bilingual documents and summaries, we can also explore: Is CLG able to boost the performance of monolingual summarization (MLS)? As we know, in commonly studied MLS, various external guidances such as keywords and fact triples were used to improve the performance of a summarization model and cross-lingual information is barely involved (Dou et al., 2021). In addition, while accompanying the exploration of cross-lingual guidance, MT techniques are not perfect and may bring some translation errors. So, we also put forward the question: how will the quality of the cross-lingual information influence the summarization performance?

Oriented with the questions above, in this paper, we empirically study how CLG benefits summarization learning under the Seq2Seq framework. For the convenience of this study, we propose a new summarization model and dataset, namely DualSum and BiRead, respectively. DualSum is a transformer based Seq2Seq model composed of a Dual Encoder and a Bilingual Decoder, which can utilize CLG in both encoding and decoding phases. BiRead is a high-quality summarization dataset which we collect for studying the effectiveness of guidance quality. Each sample in BiRead contains two human-written document-summary pairs in Chinese and English that are semantically parallel. Compared to machine translated CLG applied in previous works, guidance from BiRead

[1] We will release our data and code upon acceptance.

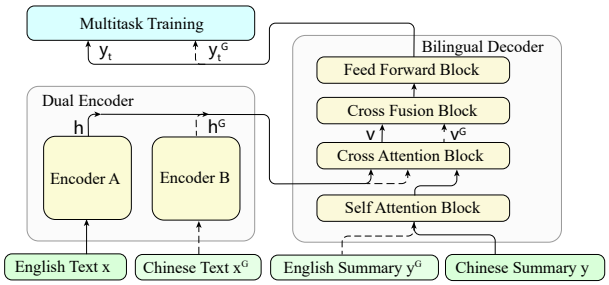

Figure 1: The Architecture of DualSum (we take English-to-Chinese Summarization as example).

English Text: However, a new study conducted by the University of Limerick in Ireland found that resistance exercise training, like weightlifting, may actually help soothe anxiety.
English Summary: Research finds that weightlifting has surprisingly calming effects.
Chinese Text: 然而，爱尔兰利默里克大学进行的一项新研究发现，像举重一样的阻力运动训练实际上可能有助于缓解焦虑。(However, a new study carried out by University of Limerick in Ireland foud that, resistance exercise like weight lifting maybe actually help relieve anxiety.)
Chinese Summary: 研究发现举重具有惊人的镇静效果。(Reaches find that weightlifting has surprisingly calming effect.)
Chinese Summary (Machine Translated): 研究发现，举重效果令人惊讶。 (Reaches find that the effect of weightlifting is surprising.)

Figure 2: Example of BiRead dataset. The first four blocks show a sample of BiRead. The fifth block shows a Chinese summary translated with machine.

is considered to be of higher quality. Finally, by experimenting with different settings of DualSum, we discover that CLG with high quality is not only beneficial to CLS but also to MLS in both encoding and decoding.

2 Model

Here we briefly introduce our model DualSum for empirical studies. Figure 1 illustrates the architecture of DualSum, which is composed of two modules: the **Dual Encoder** and the **Bilingual Decoder**. The dual encoder includes two separate text encoders respectively encoding the input document in source language and its translated text and produce contextual representations for both languages. The bilingual decoder utilizes the representations in both languages for decoding. To leverage CLG in the decoding stage, we apply multi-task learning on DualSum to jointly learn generating two summaries in both source and target languages.

Formally, we denote the input document in source language as x and its translated text as x^G; the reference summary in target language as y and the corresponding translated summary as y^G. We aim to study how x^G and y^G separately benefits this task, and how CLG of different quality benefits CLS and MLS.

Dual Encoder Two text encoders separately encodes x and x^G into their contextual representation h and h^G. Each encoder is composed of a stack of transformer layers (Vaswani et al., 2017). Following Cao et al. (2020), We apply a unified BPE (Sennrich et al., 2016) dictionary for the model and share two encoders' parameters to enhance the isomorphism of both contextual representations.

Bilingual Decoder During training, the decoder takes y and y^G as target of generation. Similar to the encoder, the decoder also adopts transformer layers. Take y as an example, in each transformer layer, we first apply a multi-head self-attention layer to get y's contextual representation o, then attend o to the encoded bilingual representations h and h^G from the encoder to produce v and v^G, respectively.

$$v, v^G = \text{MultiHead}(h, h, o), \text{MultiHead}(h^G, h^G, o)$$

where $\text{MultiHead}(x, y, z)$ represents applying x, y and z as key, value and query for multi-head attention, respectively. Later, we fuse v and v^G by:

$$u = \text{LayerNorm}((W([v; v^G]) + b)) \quad (1)$$

where W and b are trainable parameters, u is then passed to the FFN block. Similarly, we can get u^G.

Multi-Task Learning We apply multi-task learning to take advantage of CLG. We first teach the model to generate the reference summary y with Negative Log-Likelihood(NLL) loss:

$$\mathcal{L}_T = -\sum_t \log P(y_t | y_{<t}, x, x^G) \quad (2)$$

where t is the generation step of decoding. To optimize reference summary generation, we also generate the translated summary as an auxiliary task, and its NLL loss $\mathcal{L}^G$ is computed analogously. The overall training loss is defined as: $\mathcal{L} = \mathcal{L}_T + \lambda \mathcal{L}_G$, where $\lambda \in (0, 1]$ controls the weights of the auxiliary task.

3 BiRead Dataset

To evaluate the influence of CLG quality over model performance, we newly construct a dataset named BiRead which contains 247,157 bilingual news and headlines written in both Chinese and

Model	En2En			Zh2Zh			Zh2En			En2Zh		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
Trans	41.46	20.92	36.15	42.35	27.28	35.57	31.12	11.38	25.85	39.44	23.94	32.82
EncGuid (low)	41.13	20.90	36.04	42.97	28.91	36.59	32.99	12.71	28.07	40.77	25.70	33.79
EncGuid (high)	41.64	21.16	36.39	43.99	29.41	37.29	41.64	**24.16**	36.39	43.99	29.40	37.30
DecGuid (low)	42.64	21.71	37.10	44.21	29.49	37.42	35.20	14.29	29.79	40.33	25.05	33.06
DecGuid (high)	42.64	21.81	**37.22**	44.32	29.75	37.71	36.05	15.01	30.82	42.97	27.75	35.72
DualSum	**42.74**	**21.91**	37.15	**45.28**	**30.56**	**38.50**	**42.74**	21.91	**37.15**	**45.28**	**30.56**	**38.50**

Table 1: Experiment results on BiRead. **EncGuid** and **DecGuid** respectively stands for CLG in encoder and decoder side. **low** and **high** indicate CLG of low or high quality. Underline means the highest score in the fourth block. Highest score of a column is denoted in **bold**.

Model	R-1	R-2	R-L
TETran	26.15	10.60	23.24
TLTran	30.22	12.20	27.04
TNCLS	36.82	18.72	33.20
CLS+MT	40.23	22.32	36.59
CLS+MS	38.25	20.20	34.76
ATS-A	40.47	22.21	36.89
(Cao et al., 2020)	38.12	16.76	33.86
DualSum	**41.32**	**22.57**	**37.52**

Table 2: Comparison with baselines and previous works on En2ZhSum dataset. Our DualSum outcompetes all the baselines and previous models.

English in parallel. These news and headlines are collected from several bilingual news websites and most of them are written by human.[2] Therefore, different from previous machine translated CLS datasets like En2ZhSum (Zhu et al., 2019), BiRead contains high-quality semantically parallel bilingual text and summary pairs, which facilitates our empirical study. After collecting raw data from the news websites, we conduct dataset preprocessing, such as data cleaning and tokenizing. Due to space limit, the details are not presented here. We randomly split BiRead into 241,157/ 3,000/ 3,000 for training, validation and testing, respectively. On average, each English document contains 593.83 words (17.6 sentences) and each Chinese document 1207.8 Chinese characters (29.15 sentences). Each English summary has an average size of 16.4 words (1.1 sentences) and each Chinese summary 29.2 chars (1.5 sentences).

Figure 2 is an example of BiRead. The first four blocks are a sample from BiRead, including the English and Chinese texts and summaries, which are semantically similar. The fifth block shows a Chinese summary translated from the parallel English Summary using google-translation API. Obviously, the Chinese summary in our sample is of higher quality than the machine translated one.

[2]Websites like CNN, VOA, DailyMail, contain bilingual news to benefit language learners.

4 Experiments

4.1 Experiment Setup

In the experiments, we evaluate the quality of generated summaries using ROUGE F_1 (Lin, 2004). To be specific, we use ROUGE-1, 2 and L. Besides BiRead, we also conduct experiments on two CLS datasets: En2ZhSum and Zh2EnSum, proposed by Zhu et al. (2019). En2ZhSum is an English-to-Chinese summarization dataset constructed using round-trip translation with 364,687 training pairs, 3,000 validation pairs, and 3,000 test pairs. Zh2EnSum is a Chinese-to-English summarization dataset using the same strategy as En2ZhSum with 1,699,713 training pairs, 3,000 validation pairs, and 3,000 test pairs. The implementation details are listed in Appendix B.

4.2 Cross-Lingual Guidance Analysis

Effectiveness on CLS and MLS We first evaluate the summarization performance on both CLS and MLS with and without CLG on our BiRead dataset and the experiment results are illustrated in Table 1. ***Trans*** denotes the DualSum model without CLG (i.e., a vanilla transformer model). We present the performance of DualSum with and without CLG respectively in the second and fourth blocks of Table 1. For MLS, results are listed in columns ***En2En*** and ***Zh2Zh***. Compared to Trans, DualSum can steadily promote its performance for about 1 point on English summarization, and 3 points on Chinese summarization. For CLS, DualSum achieves significant improvements of about 11 points on average in Chinese-to-English CLS (the

Zh2En column), and about 6 points in English-to-Chinese CLS (the ***En2Zh*** column). These results indicates the effectiveness of CLG on MLS and CLS, respectively. We also conduct extensive experiments on the En2ZhSum and Zh2EnSum datasets to further evaluate the effect of CLG on CLS. DualSum with MT-based CLG steadily outperforms Trans on En2ZhSum Zh2EnSum (check Appendix A for details), further validating the effect of CLG in promoting CLS performance. MT-based CLG is not so effective as human translated CLG (i.e., CLG in BiRead), meaning CLG of higher quality may bring more improvements.

Guidance from Encoder and Decoder Side To evaluate the effect of CLG from encoder or decoder side, we conduct ablation study on DualSum. The third row block of Table 1 shows the results. ***EncGuid*** means DualSum with guidance only from encoder side, where we keep the dual encoder and leave alone the guidance of translated summaries from decoder side. ***DecGuid*** denotes DualSum with guidance only from decoder side, where we keep one encoder and use bilingual decoder to train through multi-task learning. We can see that in most cases, model with EncGuid or DecGuid achieves better results than Trans, proving the effectiveness of CLG in both sides. Further, on average, EncGuid improves performance by 1.0 point on MLS and 8.2 points on CLS, and DecGuid improves performance by 1.6 points on MLS and 4.0 points on CLS, indicating that ***guidance from encoder is more helpful to CLS, guidance from decoder is more helpful to MLS***. Concerning that most of the previous works of CLS only adopt CLG in decoder side, we suggest more attention be paid to leveraging CLG in encoder side.

Effectiveness of CLG Quality We apply CLG of different quality to encoder and decoder side and present the experiment results in the third block of Table 1. CLG of ***low*** quality is translated by machine, which is obtained through google-translation API in practice, and CLG of ***high*** quality is translated by human, which is directly available in our dataset. CLG of high quality in encoder side outperforms the low-quality version by an average of 0.5 and 6.5 ROUGE scores in MLS and CLS, respectively. And for decoder side it is respectively 0.1 and 1.8 on average. We can conclude that CLG of higher quality can boost model performance in all cases. Additionally, we find that improving CLG quality is more beneficial for encoder module and for CLS task.

4.3 Comparison with SOTA Models

We also compare our DualSum model with some baselines and previous work on En2ZhSum dataset. Results are presented in Table 2. As listed in the second block, TETran, TLTran and TNCLS are baseline models proposed by Zhu et al. (2019). The first two are pipeline models respectively adopting the translate-then-summarize strategy and summarize-then-translate strategy, and we use google-translation API to perform machine translation. The third is a transformer model performing end-to-end CLS. In the third block, we list the results of some previous works. CLS+MT and CLS+MS are two multi-task strategies proposed by Zhu et al. (2019) which simultaneously trains CLS with MT or MLS. ATS-A (Zhu et al., 2020) is a pointer-generator network utilizing translation patterns in CLS. Cao et al. (2020) is a multi-task framework which jointly learns cross-lingual alignment and summarization.

We can see that our model outcompetes TETran, TLTran and TNCLS by a large margin. Further, DualSum also produces better results than multitask training framework CLS+MT and CLS+MT repectively by more than 1 or 2 ROUGE points on average. We suppose it is because either CLS+MT or CLS+MS utilizes CLG only in decoder side, while DualSum adopts CLG on both encoder side and decoder side. It also outperforms ATS-A and Cao et al. (2020), which are competitive CLS models. With a simple architecture, our model DualSum can surpass some strong CLS models, and we can conclude that CLG still has an improvement space for improving CLS performance.

5 Conclusions

In this paper, we study how CLG benefits abstractive summmarization from a comprehensive view. For an empirical study, we propose the DualSum model which can utilize cross-lingual guidance in both encoding and decoding side, and construct a large-scale bilingual summarization dataset BiRead. Our work further verifies the effectiveness of CLG on both MLS and CLS tasks from encoder and decoder side. We also have some inspiring conclusions that guidance from encoder is more helpful to CLS and guidance from decoder is more helpful to MLS, and CLG of higher quality can produce better results.

References

Yu Bai, Yang Gao, and Heyan Huang. 2021. Cross-lingual abstractive summarization with limited parallel resources. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6910–6924, Online. Association for Computational Linguistics.

Yue Cao, Hui Liu, and Xiaojun Wan. 2020. Jointly learning to align and summarize for neural cross-lingual summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 6220–6231. Association for Computational Linguistics.

Zi-Yi Dou, Pengfei Liu, Hiroaki Hayashi, Zhengbao Jiang, and Graham Neubig. 2021. Gsum: A general framework for guided neural abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 4830–4842. Association for Computational Linguistics.

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 Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3162–3172. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

Anton Leuski, Chin-Yew Lin, Liang Zhou, Ulrich Germann, Franz Josef Och, and Eduard H. Hovy. 2003. Cross-lingual c*st*rd: English access to hindi information. *ACM Trans. Asian Lang. Inf. Process.*, 2(3):245–269.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Constantin Orasan and Oana Andreea Chiorean. 2008. Evaluation of a cross-lingual romanian-english multi-document summariser. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco*. European Language Resources Association.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

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

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Xiaojun Wan. 2011. Using bilingual information for cross-language document summarization. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA*, pages 1546–1555. The Association for Computer Linguistics.

Xiaojun Wan, Huiying Li, and Jianguo Xiao. 2010. Cross-language document summarization based on machine translation quality prediction. In *ACL 2010, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, July 11-16, 2010, Uppsala, Sweden*, pages 917–926. The Association for Computer Linguistics.

Junnan Zhu, Qian Wang, Yining Wang, Yu Zhou, Jiajun Zhang, Shaonan 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 2019, Hong Kong, China, November 3-7, 2019*, pages 3052–3062. Association for Computational Linguistics.

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, ACL 2020, Online, July 5-10, 2020*, pages 1309–1321. Association for Computational Linguistics.

A Other Results

Model	En2ZhSum			Zh2EnSum		
	R-1	R-2	R-L	R-1	R-2	R-L
Trans	38.56	21.13	35.12	39.75	**21.95**	35.66
DualSum	**41.33**	**22.57**	**37.52**	**40.57**	21.69	**35.76**

Table 3: Experiment results on En2ZhSum and Zh2EnSum. Meaning of **Trans** and **DualSum** is consistent with Table 1.

B Implementation Details

We use a unified BPE vocabulary with a size of approximately 15,000. For DualSum, the number of both encoder and decoder layers is 6, the hidden size is 512, and the number of heads in multi-head attention is 8. Fairseq (Ott et al., 2019) framework is used to implement all the models above. We apply Adam optimizer (Kingma and Ba, 2017) with $\beta_1 = 0.9$, $\beta_2 = 0.998$, and $\epsilon = 10^{-9}$. We set dropout rate to 0.1 and warmup steps to 8000.