Cross-Lingual Abstractive Summarization
with Limited Parallel Resources

Yu Bai, Yang Gao, Heyan Huang*

School of Computer Science and Technology,
Beijing Institute of Technology, Beijing, China
Southeast Academy of Information Technology, Fujian, China
Beijing Engineering Research Center of High Volume Language Information
Processing and Cloud Computing Applications, Beijing, China
`{yubai,gyang,hhy63}@bit.edu.cn`

Abstract

Parallel cross-lingual summarization data is scarce, requiring models to better use the limited available cross-lingual resources. Existing methods to do so often adopt sequence-to-sequence networks with multi-task frameworks. Such approaches apply multiple decoders, each of which is utilized for a specific task. However, these independent decoders share no parameters, hence fail to capture the relationships between the discrete phrases of summaries in different languages, breaking the connections in order to transfer the knowledge of the high-resource languages to low-resource languages. To bridge these connections, we propose a novel Multi-Task framework for Cross-Lingual Abstractive Summarization (MCLAS) in a low-resource setting. Employing one unified decoder to generate the sequential concatenation of monolingual and cross-lingual summaries, MCLAS makes the monolingual summarization task a prerequisite of the cross-lingual summarization (CLS) task. In this way, the shared decoder learns interactions involving alignments and summary patterns across languages, which encourages attaining knowledge transfer. Experiments on two CLS datasets demonstrate that our model significantly outperforms three baseline models in both low-resource and full-dataset scenarios. Moreover, in-depth analysis on the generated summaries and attention heads verifies that interactions are learned well using MCLAS, which benefits the CLS task under limited parallel resources.

1 Introduction

Cross-lingual summarization (CLS) helps people efficiently grasp salient information from articles in a foreign language. Neural approaches to CLS require large scale datasets containing millions of cross-lingual document-summary pairs (Zhu et al.,
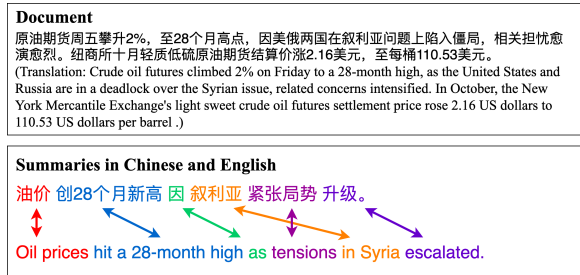
---

*Corresponding author.

Figure 1: An example of the alignments across summaries in different languages. Each color represents phrases with one specific meaning.

2019; Cao et al., 2020; Zhu et al., 2020). However, two challenges arise with these approaches: 1) most languages are low-resource, thereby lacking document-summary paired data; 2) large parallel datasets across different languages for neural-based CLS are rare and expensive, especially under the current trend of neural networks. Therefore, a low-resource setting is more realistic, and challenging, one for cross-lingual summarization. To our best knowledge, cross-lingual summarization under low-resource settings has not been well investigated and explored. Therefore, in this paper, we will develop a new model for cross-lingual abstractive summarization under limited supervision.

For low-resource settings, multi-task learning has been shown to be an effective method since it can borrow useful knowledge from other relevant tasks to use in the target task (Yan et al., 2015; Wang et al., 2020; Motiian et al., 2017). Cross-lingual summarization can be viewed as the combination of two tasks, i.e., monolingual summarization (MS) and cross-lingual translation (Zhu et al., 2019). A wealth of relationships exist across the target summaries of MS and CLS tasks, such as translation alignments and summarization patterns. Illustrated in Figure 1, "叙利亚" is mapped to "Syria", and similar maping is done with the other

6910

aligned phrases. Obviously, leveraging these relationships is crucial for the task of transferring summarization knowledge from high-resource languages to low-resource languages. Unfortunately, existing multi-task frameworks simply utilize independent decoders to conduct MS and CLS task separately (Zhu et al., 2019; Cao et al., 2020), which leads to failure in capturing these relationships.

To solve this problem, we establish reliant connections between MS and CLS tasks, making the monolingual task a prerequisite for the cross-lingual task. Specifically, one decoder is shared by both MS and CLS tasks; this is done by setting the generation target as a sequential concatenation of a monolingual summary and the corresponding cross-lingual summary. Sequentially generating monolingual and cross-lingual summaries, the decoder also conducts the translation task between them, which enhances the interactions between different languages. These **interactions** implicitly involve translation alignments, similarity in semantic units, and summary patterns across different lingual summaries. To demonstrate these decoder interactions, we further visualize them by probing Transformer attention heads in the model. Based on this process, the new structure with these advanced interactions enhances low-resource scenarios which require the model to be capable of transferring summary knowledge from high-resource languages to low-resource language. We name our model Multi-task Cross-Lingual Abstractive Summarization (MCLAS) under limited resources.

In terms of a training strategy under limited resources, we first pre-train MCLAS on large-scale monolingual document-summary parallel datasets to well-equip the decoder with general summary capability. Given a small amount of parallel cross-lingual summary samples, the model is then fine-tuned and is able to transfer the learned summary capability to the low-resource language, leveraging the interactions uncovered by the shared decoder.

Experiments on Zh2EnSum (Zhu et al., 2019) and a newly developed En2DeSum dataset demonstrate that MCLAS offers significant improvements when compared with state-of-the-art cross-lingual summarization models in both low-resource scenarios and full-dataset scenario. At the same time, we also achieved competitive performances in the En2ZhSum dataset (Zhu et al., 2019). Human evaluation results show that MCLAS produces more fluent, concise and informative summaries

than baselines models under limited parallel resources. In addition, we analyzed the length of generated summaries and the success of monolingual generation to verify advantages offered by identifying interactions between languages. We further investigate the explainability of the proposed multi-task structure by probing the attention heads in the unified decoder, proving that MCLAS learns the alignments and interactions between two languages, and this facilitates translation and summarization in the decoder stage. Our analysis provides a clear explanation of why MCLAS is capable of supporting CLS under limited resources. Our implementation and data are available at `https://github.com/WoodenWhite/MCLAS`.

2 Related Work

2.1 Cross-Lingual Summarization

Recently, cross-lingual summarization has received attention in research due to the increasing demand to produce cross-lingual information.

Traditional CLS systems are based on a pipeline paradigm (Wan et al., 2010; Wan, 2011; Zhang et al., 2016). These pipeline systems first translate the document and then summarize it or vice versa. Shen et al. (2018) propose the use of pseudo summaries to train the cross-lingual abstractive summarization model. In contrast, Duan et al. (2019a) and Ouyang et al. (2019) generate pseudo sources to construct the cross-lingual summarization dataset.

The first large-scale cross-lingual summarization datasets are acquired by use of a round-trip translation strategy (Zhu et al., 2019). Additionally, Zhu et al. (2019) propose a multi-task framework to improve their cross-lingual summarization system. Following Zhu et al. (2019), more methods have been proposed to improve the CLS task. Zhu et al. (2020) use a pointer-generator network to exploit the translation patterns in cross-lingual summarization. Cao et al. (2020) utilize two encoders and two decoders to jointly learn to align and summarize. In contrast to previous methods, MCLAS generates the concatenation of monolingual and cross-lingual summaries, thereby modeling relationships between them.

2.2 Low-Resource Natural Language Generation

Natural language generation (NLG) for low-resource languages or domains has attracted lots of attention. Gu et al. (2018) leverage meta-learning

to improve low-resource neural machine translation. Meanwhile, many pretrained NLG models have been proposed and adapted to low-resource scenarios (Song et al., 2019; Chi et al., 2020; Radford et al., 2019; Zhang et al., 2019a). However, these models require large-scale pretraining. Our work does not require any large pretrained generation models or translation models, enabling a vital decreases in training cost.

3 Background

3.1 Neural Cross-lingual Summarization

Given a source document $D^A = \{x_1^A, x_2^A, \ldots, x_m^A\}$ in language A, a monolingual summarization system converts the source into a summary $S^A = \{y_1^A, y_2^A, \ldots, y_n^A\}$, where m and n are the lengths of D^A and S^A, respectively. A cross-lingual summarization system produces a summary $S^B = \{y_1^B, y_2^B, \ldots, y_{n'}^B\}$ consisting of tokens y^B in target language B, where n' is the length of S^B. Note that the mentioned x^A, y^A, and y^B are all tokens.

Zhu et al. (2019) propose using the Transformer (Vaswani et al., 2017) to conduct cross-lingual summarization tasks. The Transformer is composed of stacked encoder and decoder layers. The encoder layer is comprised of a self-attention layer and a feed-forward layer. The decoder layer shares the same architecture as the encoder except for an extra encoder-decoder attention layer, which performs multi-head attention over the output of stacked encoder layers. The whole Transformer model θ is trained to maximize the conditional probability of the target sequence S^B as follows:

$$L_{\text{NCLS}} = \sum_{t=1}^{N} \log P(y_t^B | y_{<t}^B, D^A) \tag{1}$$

3.2 Improving NCLS with Multi-Task Frameworks

Considering the relationship between CLS and MS, in which they share the same goal to summarize important information in a document, Zhu et al. (2019) proposed employing a one-to-many multi-task framework to enhance the basic Transformer model. In this framework, one encoder is employed to encode the source document D^A. Two separate decoders simultaneously generate a monolingual summary S^A and a cross-lingual summary S^B, leading to a loss as follows:

$$L_{\text{NCLS+MS}} = \sum_{t=1}^{n} \log P(y_t^A | y_{<t}^A, D^A) \\ + \sum_{t=1}^{n'} \log P(y_t^B | y_{<t}^B, D^A) \tag{2}$$
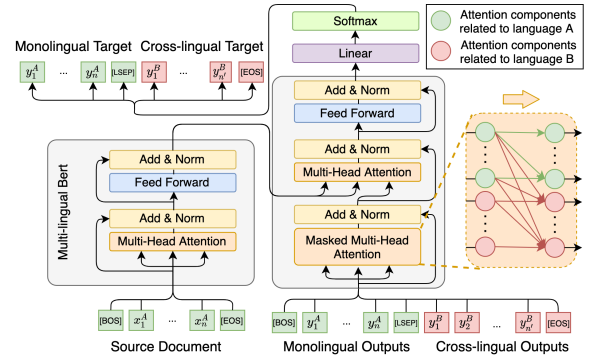

Figure 2: An overview of our proposed MCLAS. A unified decoder produces both monolingual (green) and cross-lingual (red) summaries. The green and red lines represent the monolingual and cross-lingual summaries' attention, respectively.

This multi-task framework shares encoder representation to enhance cross-lingual summarization. However, independent decoders in this model are incapable of establishing alignments and connections between cross-lingual summaries.

4 MCLAS with Limited Parallel Resources

To strengthen the connections mentioned, we propose making the monolingual task a prerequisite for the cross-lingual task through modeling interactions. According to previous work (Wan et al., 2010; Yao et al., 2015; Zhang et al., 2016), interactions between cross-lingual summaries (important phrase alignments, sentence lengths, and summary patterns, etc) are crucial for the final summary's quality. We leverage these interactions to further transfer the rich-resource language knowledge. Detailed descriptions of this step are presented in following sections.

4.1 Multi-Task Learning in MCLAS

To model interactions between languages, we need to share the decoder's parameters. Inspired by Dong et al. (2019), we propose sharing the whole decoder to carry out both the translation and the summarization tasks. Specifically, we substitute the generation target S^A with the sequential concatenation of S^A and S^B:

$$S^{AB} = \{[\text{BOS}], y_1^A, y_2^A, \ldots, y_n^A, \\ [\text{LSEP}], y_1^B, y_2^B, \ldots, y_{n'}^B, [\text{EOS}]\} \tag{3}$$

where [BOS] and [EOS] are the beginning and end token of the output summaries, respectively. And

[LSEP] is the special token used as the separator of S^A and S^B.

With the new generation target, the decoder learns to first generate S^A, and then generate S^B conditioned on S^A and D^A. The whole generation process is illustrated in Figure 2.

Formally, we maximize the joint probability for monolingual and cross-lingual summarization:

$$L_{\text{MCLAS}} = \begin{array}{l} \sum_{t=1}^{n} \log P(y_t^A | y_{<t}^A, D^A) \\ + \sum_{t=1}^{n'} \log P(y_t^B | y_{<t}^B, S^A, D^A) \end{array} \quad (4)$$

The loss function can be divided into two terms. When generating S^A, the decoder conducts the MS task based on D^A, corresponding to the first term in Equation (4). When generating S^B, the decoder already knows the information of corresponding monolingual summaries. In this way, it performs the translation task (for S^A) and the CLS task (for D^A), achieved by optimizing the second term in Equation (4). With the modification of the target, our model can easily capture interactions between cross-lingual summaries. The trained model shows effectiveness in aligning the summaries. Not only the output tokens, but also the attention distributions are aligned. The model we designed leverages this phenomenon to enable monolingual knowledge to be transferred under low-resource scenarios. Detailed investigation is presented in Section 6.

We adopt Transformers as our base model. In addition, we use multilingual BERT (Devlin et al., 2019) to initialize the encoder, improving its ability to produce multilingual representations. Additionally, having tried many different position embedding and language segmentation embedding methods, we find that [LSEP] is enough for the model to distinguish whether it is generating S^B. Hence keeping the original position embedding (Vaswani et al., 2017) and employing no segmentation embedding are best for performance and efficiency.

4.2 Learning Schemes for MCLAS under Limited Resources

Since our proposed framework enforces interactions between cross multilingual summaries, it has further benefits to the low-resource scenario, as only a few training summary samples are available in a cross-language. Yet, simply training from scratch can not make the best of our proposed model in low-resource scenarios. Hence we use a pre-training and fine-tuning paradigm to transfer the rich-resource language knowledge.

Scenarios	Zh2EnSum	En2DeSum	En2ZhSum
Minimum	5,000 (0.3%)	2,619 (0.6%)	1,500 (0.4%)
Medium	25,000 (1.5%)	12,925 (3.0%)	7,500 (2.0%)
Maximum	50,000 (3.0%)	25,832 (6.0%)	15,000 (4.0%)
Full-dataset	1,693,713	429,393	364,687

Table 1: Sample sizes of different low-resource scenarios. Three low-resource scenarios with various sample sizes are created for each dataset. Minimum, Medium, and Maximum represent sample sizes in the minimum low-resource scenario, medium low-resource scenario, and maximum low-resource scenario, respectively.

First, we train the model in a monolingual summarization dataset. In this step, the model learns how to produce a monolingual summary for a given document. Then, we jointly learn MS and CLS with few training samples, optimizing Equation (4). We adopt similar initialization to existing CLS methods, which is introduced in Section 5.3.

5 Experiments

5.1 Datasets

we conduct experiments on the En2ZhSum, Zh2EnSum CLS datasets[1] (Zhu et al., 2019) and a newly constructed En2DeSum dataset. En2ZhSum is an English-to-Chinese dataset containing 364,687 training samples, 3,000 validation, and 3,000 testing samples. The dataset is converted from the union set of CNN/DM (Hermann et al., 2015) and MSMO (Zhu et al., 2018) using a round-trip translation strategy. Converted from the LCSTS dataset, Zh2EnSum contains 1,693,713 Chinese-to-English training samples, 3,000 validation, and 3,000 testing samples. To better verify the CLS ability of MCLAS, we construct a new English-to-German dataset (En2DeSum), using the same methods proposed by Zhu et al. (2019). We use WMT'19 English-German winner[2] as our translation model to process the English Gigaword dataset.[3] We set the threshold $T_1 = 0.6$ and $T_2 = 0.2$. The final En2DeSum contains 429,393 training samples, 4,305 validation samples, and 4,099 testing samples.

All the training samples contain a source document, a monolingual summary, and a cross-lingual summary. For the full-dataset scenario, we train the model with the whole dataset. For low-resource scenarios, we randomly select 3 different amounts

---

[1] www.nlpr.ia.ac.cn/cip/dataset.htm
[2] https://github.com/pytorch/fairseq/tree/master/examples/translation
[3] LDC2011T07

(minimum, medium, and maximum) of training samples for all datasets to evaluate our model's performance under low-resource scenarios. Detailed numbers are presented in Table 1.

5.2 Training and Inference

We use multilingual BERT (mBERT) (Devlin et al., 2019) to initialize our Transformer encoder. The decoder is a Transformer decoder with 6 layers. Each attention module has 8 different attention heads. The hidden size of the decoder's self-attention is 768 and that of the feed-forward network is 2048. The final model contains 296,046,231 parameters. Because the encoder is pretrained when the decoder is randomly initialized, we use two separate optimizers for the encoder and the decoder (Liu and Lapata, 2019). The encoder's learning rate η_e is set as 0.005, while the decoder's learning rate η_d is 0.2. Warmup-steps for the encoder are 10,000 and 5,000 for the decoder. We train the model on two TITAN RTX GPUs for one day with gradient accumulation every 5 steps. Dropout with a probability 0.1 is applied before all the linear layers. We find that the target vocabulary type doesn't have much influence on the final result. Therefore, we directly use mBERT's subwords vocabulary as our target vocabulary. Nevertheless, in case tokens would be produced in the wrong language, we constructe a target token vocabulary for each target language. In the inference period, we only generate tokens from the corresponding vocabulary. During the decoding stage, we use beam search (size 5) and trigram block to avoid repetition. Length penalty is set between 0.6 and 1. All the hyperparameters are manually tuned using PPL and accuracy metric on the validation set.

5.3 Baselines

We compare MCLAS in low-resource scenarios with the following baselines:

NCLS CLS model proposed by Zhu et al. (2019). In low-resource scenarios, we initialize our model with the pretrained MS model and then use a few samples to optimize Equation (1).

NCLS+MS Multi-task framework proposed by Zhu et al. (2019). We find that NCLS+MS fails to converge when it is partly initialized by the pretrained MS model (the CLS decoder is randomly initialized). Hence, we fully initialize the multi-task model using the pretrained MS model. Specifically, the two separate decoders are both initialized

by the pretrained monolingual decoder. Then the model is optimized with Equation (2).

TLTran Transformer-based Late Translation is a pipeline method. First, a monolingual summarization model summarizes the source document. A translation model is then applied to translate the summary. The summarization model is trained with monolingual document-summary pairs in three datasets. Specifically, we continue using WMT'19 English-German winner as the translation model for En2DeSum.

Some recent proposed models improve the performance of CLS task. Methods **NCLS+MT**, **TETran** (Zhu et al., 2019), and the system proposed by Ouyang et al. (2019) require external long document machine translation (MT) corpora. The method proposed by Cao et al. (2020) requires not only parallel summaries but also document pairs translated by MT systems. Another method proposed by Zhu et al. (2020) requires bilingual lexicons extracted from large parallel MT datasets (2.08M sentence pairs from eight LDC corpora). We choose not to use these models as baselines since comparing MCLAS with them is unfair.

5.4 Automatic Evaluation Results

The overall results under low-resource scenarios and full-dataset scenario are shown in Table 2. We reimplement a variety of models and evaluate them using F1 scores of the standard ROUGE metric (Lin, 2004) (ROUGE-1, ROUGE-2, and ROUGE-L) and BERTScore[4] (Zhang et al., 2019b). The following analysis is from our observations.

In the Zh2EnSum and En2DeSum datasets, MCLAS achieves significant improvements over baselines in all the low-resource scenarios. It is worth noting that combining NCLS+MS in our experiments does not bring much improvement to the NCLS model. We consider that this is because mBERT has already provided multilingual encoding for our models.

However, we find that in the En2ZhSum dataset, MCLAS did not perform as well as that in the other two datasets. We speculate that is due to the imbalance of English reference and Chinese reference. The average length of S^A and S^B in En2ZhSum is 55.21 and 95.96, respectively (Zhu et al., 2019). This condition largely breaks the alignment between languages, leading to MCLAS

---

[4]https://github.com/Tiiiger/bert_score

Models		Zh2EnSum				En2DeSum				En2ZhSum			
		R-1	R-2	R-L	BERTScore	R-1	R-2	R-L	BERTScore	R-1	R-2	R-L	BERTScore
Minimum Low-resource Scenario	NCLS	20.93	5.88	17.58	**0.5041**	17.59	5.01	16.58	0.7202	**34.14**	12.45	**21.20**	0.7096
	NCLS+MS	20.50	5.45	17.25	0.5025	17.52	5.27	16.57	0.7198	33.96	12.38	21.07	**0.7102**
	MCLAS	**21.03**	**6.03**	**18.16**	0.5023	**19.19**	**5.91**	**18.43**	**0.7282**	32.03	**13.17**	21.17	0.6529
Medium Low-resource Scenario	NCLS	26.42	8.90	22.05	0.5373	23.55	8.09	22.13	0.7400	35.98	15.88	23.79	**0.7298**
	NCLS+MS	26.86	9.06	22.47	0.5377	23.60	8.35	22.14	0.7431	38.95	18.09	**25.39**	0.7172
	MCLAS	**27.84**	**10.41**	**24.12**	**0.5464**	**27.22**	**10.09**	**26.00**	**0.7575**	37.28	**18.10**	25.26	0.6839
Maximum Low-Resource Scenario	NCLS	29.05	10.88	24.32	0.5492	25.84	9.78	24.25	0.7483	**40.18**	19.86	26.52	0.7435
	NCLS+MS	28.63	10.63	24.00	0.5485	25.59	9.58	23.96	0.7484	39.86	**19.87**	**26.64**	**0.7445**
	MCLAS	**30.73**	**12.26**	**26.51**	**0.5633**	**30.31**	**12.32**	**28.88**	**0.7682**	38.35	19.75	26.41	0.6921
Full dataset Scenario	TLTran	33.64	15.58	29.74	-	28.57	13.31	26.34	-	30.20	12.20	27.02	-
	NCLS	35.60	16.78	30.27	**0.5835**	31.61	14.24	29.63	0.7680	**44.16**	24.28	**30.23**	**0.7407**
	NCLS+MS	34.84	16.05	29.47	0.5807	31.33	13.86	29.31	0.7675	42.68	23.51	29.24	0.7361
	MCLAS	**35.65**	**16.97**	**31.14**	0.5770	**36.48**	**17.21**	**34.86**	**0.7897**	42.27	**24.60**	30.09	0.7069

Table 2: F1 scores of ROUGE and BERTScore in Zh2EnSum, En2DeSum and En2ZhSum dataset. R-1, R-2, and R-L represents ROUGE-1, ROUGE-2, and ROUGE-L, respectively.

Models	Minimum			Medium			Maximum		
	IF	CC	FL	IF	CC	FL	IF	CC	FL
MCLAS	**-0.264**	**0.164**	**-0.021**	0.000	**0.236**	**0.164**	**0.057**	**0.464**	**0.214**
NCLS	-0.243	-0.386	-0.364	**0.036**	-0.221	-0.257	-0.129	-0.329	-0.186
NCLS+MS	-0.371	-0.407	-0.286	-0.343	-0.536	-0.407	-0.179	-0.364	-0.214
GOLD	0.879	0.629	0.671	0.300	0.529	0.500	0.257	0.221	0.179

Table 3: Human evaluation results in Zh2EnSum dataset. The best results are in bold.

Scenarios	Fleiss' Kappa	Overall Agreement
Minimum	0.37	60.48%
Medium	0.22	51.35%
Maxmium	0.20	50.16%

Table 4: Fleiss' Kappa and overall agreement percent of our human evaluation results. A higher value indicates higher agreements among participants.

Scenarios	Models	En2DeSum	Zh2EnSum
Minimum Low-resource Scenario	NCLS	13.48 (+4.69)	18.49 (+3.51)
	NCLS+MS	12.83 (+4.04)	18.68 (+3.70)
	MCLAS	7.80 (−**0.90**)	13.16 (−**1.82**)
Medium Low-resource Scenario	NCLS	13.13 (+4.34)	18.60 (+3.62)
	NCLS+MS	12.90 (+4.11)	18.57 (+3.59)
	MCLAS	8.65 (−**0.14**)	13.10 (−**1.88**)
Maximum Low-resource Scenario	NCLS	13.37 (+4.58)	18.44 (+3.46)
	NCLS+MS	13.37 (+4.58)	18.75 (+3.77)
	MCLAS	8.46 (−**0.33**)	12.83 (−**2.15**)
Gold		8.79	14.98

Table 5: Target summary length generated by various models. The best results are in bold.

Models	En2DeSum			Zh2EnSum		
	R-1	R-2	R-L	R-1	R-2	R-L
MS-Pretrain	39.16	19.21	35.42	41.71	**27.76**	**37.97**
NCLS+MS	36.06	16.84	32.72	39.98	26.03	36.13
MCLAS	**45.59**	**23.77**	**42.51**	41.72	27.69	37.92

Table 6: Monolingual summary results in Zh2EnSum and En2ZhSum datasets. MS-Pretrain refers to the pre-trained model for monolingual summarization.

the performing slightly weaker. Despite this, results in En2DeSum and Zh2EnSum demonstrate that our proposed MCLAS model is effective for CLS under limited resources.

Finally, our proposed model also has superior performance compared to baseline models given the full training dataset, achieving the best ROUGE score in En2DeSum and Zh2EnSum datasets.

5.5 Human Evaluation

In addition to automatic evaluation, we conduct a human evaluation to verify our model's performance. We randomly chose 60 examples (20 for each low-resource scenario) from the Zh2EnSum test dataset. Seven graduate students with high levels of fluency in English and Chinese are asked to assess the generated summaries and gold summaries from independent perspectives: **informativeness**, **fluency**, and **conciseness**. We follow the Best-Worst Scaling method (Kiritchenko and Mohammad, 2017). Participants are asked to indicate

the best and worst items from each perspective. The result scores are calculated based on the percentage of times each system is selected as best minus the times it is selected as worst. Hence, final scores range from -1 (worst) to 1 (best). Results are shown in Table 3.

As the data size increases, all the models achieve better results. Our proposed MCLAS outperformed NCLS and NCLS+MS in all the metrics. We notice that MCLAS is especially strong in conciseness. This phenomenon will be analyzed in Section 5.7

We show Fleiss' Kappa scores of our conducted human evaluation in Table 4, which demonstrates a good inter-agreement among the participants.

Metrics	MS-Pretrain	NCLS+MS	MCLAS	Ground Truth
R-1 Recall	**58.37**	52.62	46.88	-
R-1 Precision	30.45	28.41	**46.31**	-
R-2 Recall	**30.21**	25.91	24.54	-
R-2 Precision	14.65	12.99	**24.11**	-
R-L Recall	**52.97**	47.89	43.74	-
R-L Precision	27.51	25.71	**43.15**	-
Avg. Length	18.64 (+9.53)	17.98 (+8.87)	9.15 (**+0.04**)	9.11

Table 7: Analysis on monolingual summary generation ability of MCLAS trained with En2DeSum dataset.
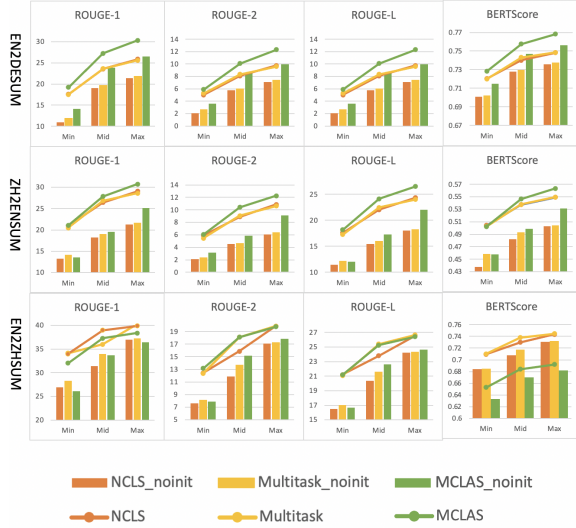
Figure 3: Line and column chart of En2DeSum and Zh2EnSum results. Lines represent models initialized with pretrained monolingual summarization model. Columns represent models trained from scratch.

5.6 Analysis on Initialization Methods

We use a monolingual summarization model to initialize our model. However, whether this initialization method works is still in question. Therefore we compare our models with non-initialized models, shown in Figure 3. Among the three datasets, the initialization methods bring a huge improvement to all of the models.

5.7 Analysis on Summary Length

One of the goals of automatic summarization is to produce brief text. Yet many neural auto-regressive models tend to produce a longer summary to improve the recall metric. Results in Table 5 show that interactions enable MCLAS to generate shorter summaries than other models, which more closely resembles human summaries. We can safely conclude that MCLAS can keep the summary in a fairly appropriate length, leading to concise generated summaries. We speculate that this is due to its ability to capture interactions between languages,

Figure 4: An example of generated cross-lingual summaries. Important phrases are bold while incorrect information generated by each model is italicized. Non-fluent parts in sentences are bold and italicized.

conditioning cross-lingual summaries on relatively precise monolingual summaries.

5.8 Analysis on Monolingual Summarization

Modeling interactions between languages brings many advantages. Specifically, we find that MCLAS can preserve more monolingual summarization knowledge than the NCLS+MS model during low-resource fine-tuning, or even promote its performance. We generate monolingual summaries with models trained in the maximum low-resource scenario. In Table 6, we can clearly see that MCLAS retains more monolingual summarization knowledge in the Zh2EnSum dataset. In the En2DeSum dataset, monolingual summarization performance is even significantly improved. We speculate that this is due to MCLAS's ability to provide the interactions between languages.

We focus specifically on digging into results in En2DeSum, evaluating its detailed ROUGE and average summary length, presented in Table 7. We find that ROUGE improvement mainly resulted from precision while recall barely decrease the performances. This and the Avg. length metric shows that MCLAS produces more precise summaries while retaining most of the important information, leading to the metric increase in ROUGE.
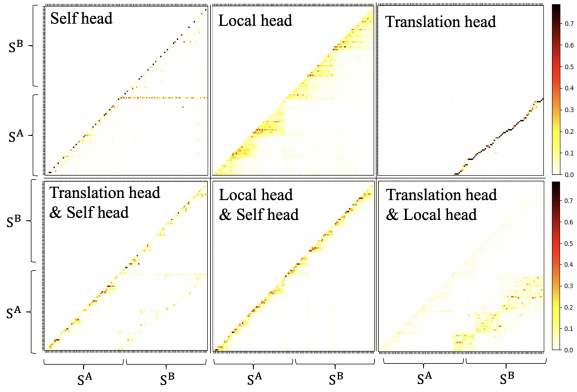
Figure 5: Different types of self-attention heads in MCLAS's decoder. The x-axis and y-axis are both concatenated source-language summary S^A and target-language summary S^B tokens. Darker color shows the more highly related associations between tokens. The horizontal line in self head represents the [LSEP] token. Some attention heads attend to [LSEP] to confirm whether it is generating a cross-lingual summary.

Figure 6: Different types of encoder-decoder attention heads in MCLAS's decoder. The x-axis represents concatenated source-language summary S^A and target-language summary S^B tokens while the y-axis is the document D^A tokens. In news texts, important information often gathers in the front part of the document. We only retain the informative part of the y-axis, omitting the blank part that the model do not attend to.

5.9 Case Study

In Figure 4, on the Zh2EnSum dataset, there is a list comparing the reference summary and outputs of models trained in the maximum low-resource scenario. Clearly, the NCLS model loses the information "two cars" and generates the wrong information "No.2 factory". The NCLS+MS model is not accurate when describing the number of injured people, dropping important information "more than". Additionally, the NCLS+MS model also has fluency and repetition issues: "in zhengzhou" appears twice in its generated summary. In contrast, MCLAS captures all of this information mentioned in both its Chinese and English output, and the English summary is well aligned with the Chinese summary. Finally, all of the models ignore the information "foxconn printed on the body of the car". See Appendix A for more examples.

6 Probing into Attention Heads

We have observed a successful alignment between S^A and S^B produced by our model in Section 5.9. In this section, we dig into this and analyze how the model learns the relationships. For a CLS task from document D^A to S^B, our hypotheses are: (1) the unified decoder is implicitly undertaking translation from S^A to S^B; (2) the unified decoder also conducts both monolingual and cross-lingual summarization. To verify these hypotheses, we visualize attention distributions of the Transformer decoders trained on En2ZhSum. Neural models
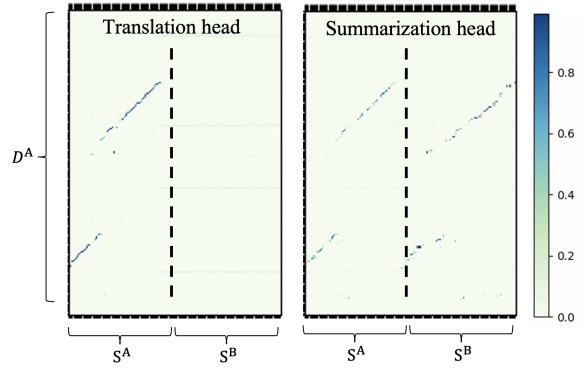
can be explicitly explained using probing into the attention heads (Michel et al., 2019; Voita et al., 2019). We follow the previous work and visualize the function of all attention heads in the decoder to verify the relationships of the concatenated cross-lingual summaries (i.e., translation) and cross-lingual document-summary pairs (i.e., summarization).

6.1 Analysis on Translation

We assume that the decoder translates only if the source summary S^A and the target summary S^B align well. This means that MCLAS is transferring knowledge from S^A to S^B. We visualize and probe all 48 **self-attention heads** in the unified decoder. We find 23 (47.9%) **translation heads**, defined as the heads attending from y_j^B to the corresponding words in language A. These heads undertake a translation function. 19 (39.6%) heads are **local heads**, attending to a few words before them and modeling context information. 12 (25%) heads are **self heads**, which only attend to themselves to retain the primary information. Some of the heads can be categorized into two types. Note that all of the heads behave similarly across different samples. We find that most of the heads are translation heads, indicating that our unified decoder is translating S^A into S^B. We sample some representative heads in Figure 5 to show their functionalities.

6.2 Analysis on Summarization

To analyze whether the decoder for S^B is simply translating from S^A or that it also summarizes the source document, we visualize the distribution of 48 **encoder-decoder attention** heads. We find 28 (58.3%) **summarization heads** that attend to the document's important parts when generating both the monolingual summary and the cross-lingual summary. We also find 20 (41.7%) **translation heads**, which focus on the source document when generating S^A, while focusing on nothing when generating S^B. We speculate that summarization heads are responsible for the summarization function and that translation heads cut down the relation between S^B and source document D^A, leaving space for translation. Again, all the heads behave similarly across different samples. We select two representative samples in Figure 6.

The existence of both summarization and translation heads in encoder-decoder attention components supports our views: the unified decoder simultaneously conducts translation and summarization. Therefore, our model enhances the interactions between different languages, being able to facilitate cross-lingual summarization under low-resource scenarios. See Appendix B for detailed visualization results.

7 Discussions

An ideal low-resource experiment should be conducted with real low-resource languages. Although possible, it takes much effort to acquire such datasets. Hence, it is the second-best choice that we simulate our low-resource scenarios by artificially limiting the amount of the available data. Some may question it about the feasibility of our method in real low-resource languages since machine translation systems, which is used to generate document-summary pairs, would be of lower quality for truly low-resource languages. For this concern, we consider it still possible to acquire thousands of high-quality human translated parallel summaries, as Duan et al. (2019b) adopt on their test set, to apply our method.

8 Conclusion

In this paper, we propose a novel multi-task learning framework MCLAS to achieve cross-lingual abstractive summarization with limited parallel resources. Our model shares a unified decoder that sequentially generates both monolingual and cross-lingual summaries. Experiments on two cross-lingual summarization datasets demonstrate that our framework outperforms all the baseline models in low-resource and full-dataset scenarios.

References

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*, pages 6220–6231.

Zewen Chi, Li Dong, Furu Wei, Wenhui Wang, Xian-Ling Mao, and Heyan Huang. 2020. Cross-lingual natural language generation via pre-training. In *AAAI*, pages 7570–7577.

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.

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*, volume 32. Curran Associates, Inc.

Xiangyu Duan, Mingming Yin, Min Zhang, Boxing Chen, and Weihua Luo. 2019a. 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.

Xiangyu Duan, Mingming Yin, Min Zhang, Boxing Chen, and Weihua Luo. 2019b. Zero-shot cross-lingual abstractive sentence summarization through teaching generation and attention. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3162–3172, Florence, Italy. Association for Computational Linguistics.

Jiatao Gu, Yong Wang, Yun Chen, Victor OK Li, and Kyunghyun Cho. 2018. Meta-learning for low-resource neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3622–3631.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in neural information processing systems*, pages 1693–1701.

Svetlana Kiritchenko and Saif Mohammad. 2017. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 465–470.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. 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 3721–3731.

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems*, pages 14014–14024.

Saeid Motiian, Quinn Jones, Seyed Mehdi Iranmanesh, and Gianfranco Doretto. 2017. Few-shot adversarial domain adaptation. In *NIPS*.

Jessica Ouyang, Boya Song, and Kathy 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, Minneapolis, Minnesota. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

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.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. Mass: Masked sequence to sequence pre-training for language generation. In *International Conference on Machine Learning*, pages 5926–5936.

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.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808.

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.

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.

Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. 2020. Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys (CSUR)*, 53(3):1–34.

Wang Yan, Jordan Yap, and Greg Mori. 2015. Multi-task transfer methods to improve one-shot learning for multimedia event detection. In *BMVC*, pages 37–1.

Jin-ge Yao, Xiaojun Wan, and Jianguo Xiao. 2015. Phrase-based compressive cross-language summarization. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 118–127.

Jiajun Zhang, Yu Zhou, and Chengqing Zong. 2016. Abstractive cross-language summarization via translation model enhanced predicate argument structure fusing. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(10):1842–1853.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J Liu. 2019a. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. *arXiv preprint arXiv:1912.08777*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019b. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Junnan Zhu, Haoran Li, Tianshang Liu, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2018. Msmo: 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, 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)*, pages 3045–3055.

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.

A Samples

We list some samples from outputs of various models. Samples from En2DeSum dataset are shown in Figure 7. Samples from Zh2EnSum dataset are shown in Figure 8. We randomly selected one sample from each low-resource scenario.

B Attention Distributions

In Section "Probing into Attention Heads", we selected some representative attention heads. We list all of our trained attention heads among 6 Transformer decoder layers in Figure 9 and Figure 10 for reference.

Sample 1 (Max Low-Resource Scenario)
Source: the federal agency that provides the public with bus safety data has some incomplete information and ac k now led ges it focuses on larger carriers , leaving smaller companies with little over sig ht .
MCLAS: agentur stellt daten von bussexporten in frage
NCLS: agentur sagt , dass sie uberspannte bussicherheit nur ein einfacher gewesen sein
NCLS+MS: agentur sagt , dass bussicherheit nicht mehr uberwache , wenn sie nicht uberwindet
Reference: informationen uber viele busunternehmen fehlen
Sample 2 (Med Low-Resource Scenario)
Source: bayern munich stol e a point from spartak moscow in group b of the champions ' cup on wednesday , when their pre - season signing from hamburg , marcus bab bel , scored an injury - time equal iser for a 1 - 1 draw .
MCLAS: bayern und moskau bindet moskauer 1 : 1 - unentschieden
NCLS: bayern schlagt moskauer 1 : 1 im champions - cup eingeladen mit 1 - unentschieden.
NCLS+MS: bayern frustrieren moskau , um sich im champions ' cup zu 1 : 1 - qualifikationsste
Reference: bayern klauen einen punkt
Sample 3 (Min Low-Resource Scenario)
Source: valencia ' s uruguay an striker diego alonso got the admission on tuesday from the spanish football federation to play as an eu player after receiving an italian passport , the club said .
MCLAS: spanischer alonso wird in spanien interessiert
NCLS: spaniens uruguay ' s alonso wird as eu - player , allegation zauberer zu erklaren
NCLS+MS: deportivo schickt alonso aus eu - fußballer zau - verbot " zauberer "
Reference: alonso erhalt italienischen pass

Figure 7: Examples of models trained in En2DeSum dataset.

Sample 1 （Max Low-Resource Scneario）
Source: 韩国检方 6 日指出，船体改装、超载及舵手操作不熟练等多个因素导致了岁月号事故，海警应对不力，他们甚至制作虚假文件隐瞒失误。载有 476 人的岁月号 4 月 16 日沉没，近 300 人遇难，10 人失踪。事故发生后，399 人被调查，其中 154 人被捕。(Translation: South Korean prosecutors pointed out on the 6th that multiple factors such as hull modification, overloading and unskilled operation of the helmsman led to the Suiyue accident. The coastal police did not respond well, and they even produced false documents to conceal mistakes. The Year, which contained 476 people, sank on April 16. Nearly 300 people were killed and 10 people were missing. After the accident, 399 people were investigated and 154 of them were arrested.)
MCLAS: korean marine police made false documents to hide errors
NCLS: south korea ' s response to the wreck accident of " year ' s " : more than 400 people were arrested
NCLS+MS: nearly 300 people were killed and 10 missing in the sunken ship accident in south korea for the time being
Reference: five months later , the cause of the sinking of the ship nian came to the surface .
Sample 2 （Mid Low-Resource Scneario）
Source: 《路透社》旗下 ifr 报道，深圳 a 股和 b 股均已上市的万科，正研究转板香港上市的可能性。市场预测万科很可能於明日 (4 日)内地股市复市后宣布香港上市计划。万科 a 股和 b 股已於 12 月 26 日起停牌。截至 12 月 25 日，万科市值 164 亿港元。(Translation: Ifr, a subsidiary of Reuters, reported that China Vanke, which has both listed a-shares and b-shares in Shenzhen, is studying the possibility of a Hong Kong listing. Market forecasts that Vanke is likely to announce its Hong Kong listing plan tomorrow (4th) after the mainland stock market resumes. Trading in Vanke's a-share and b-shares has been suspended since December 26. As of December 25, Vanke had a market value of HK$16.4 billion.)
MCLAS: foreign media : vanke research on hong kong ' s listing
NCLS: foreign media : vanke ' s research on hong kong listing next month , and h shares are suspended .
NCLS+MS: vanke intends to reconstrue hong kong ' s listing in hong kong next month
Reference: it is said that wanke is investigating b - to - h shares and announcing hong kong listing plan tomorrow
Sample 3 （Min Low-Resource Scneario）
Source: 在余姚一段未被水淹的区域里，饭店的老板俞熠华，从 9 日开始，从宁海采购了大量食材，为因城市内涝严重被困的余姚市民免费提供了上万份热饭热菜。像这样的故事，像俞熠华这样的人，在余姚、在整个浙江还有很多。(Translation: In a section of Yuyao that is not flooded, the owner of the restaurant, Yu Yihua, has purchased a large amount of ingredients from Ninghai starting on the 9th and provided tens of thousands of hot meals and dishes to Yuyao citizens who were severely trapped by the city's waterlogging. There are many stories like this, people like Yu Yihua, in Yuyao and the whole of Zhejiang..)
MCLAS: zhejiang citizens donated 10 , 000 kilograms of thousands of meals
NCLS: china ' s most beautiful " with a large number of thousands of people to migrant workers in zhejiang province
NCLS+MS: china ' s most beautiful " invites more than 20 , 000 yuan of public funds to hangzhou city
Reference: typhoon day " micro-love " condenses da ai folk quan yong the most beautiful zhejiang people

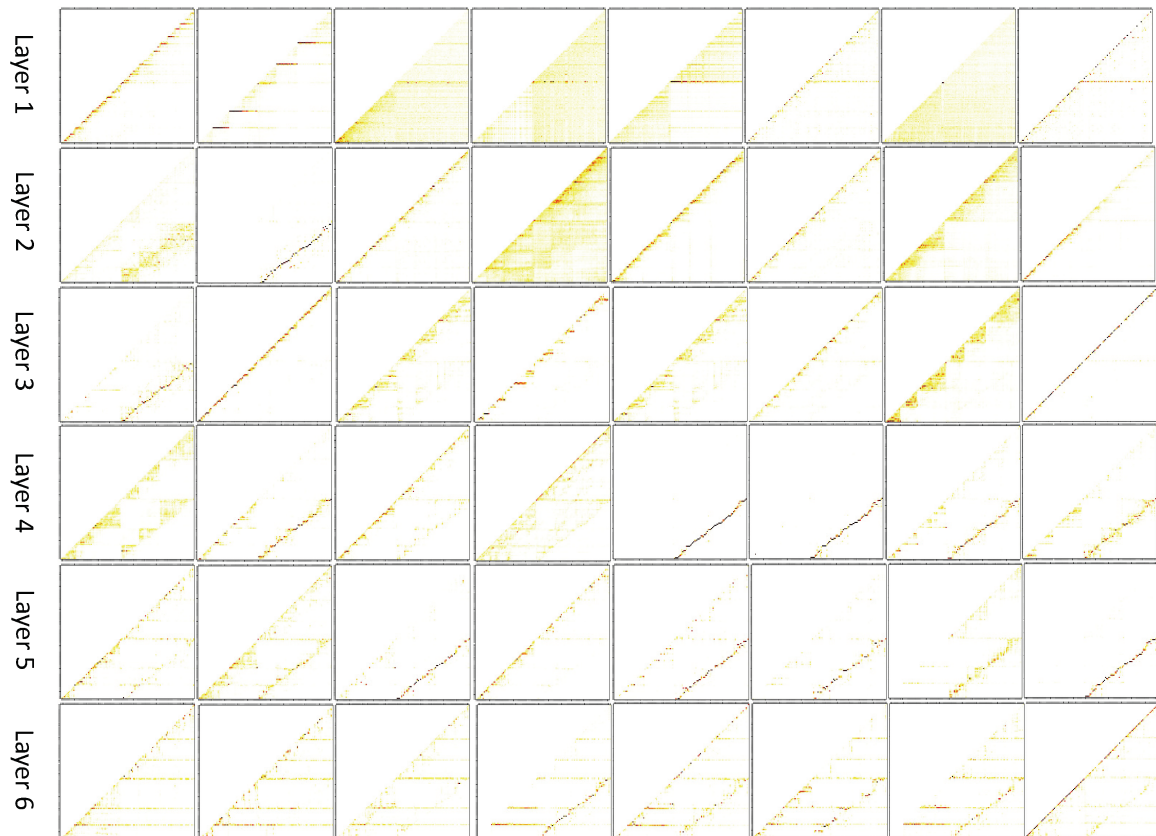Figure 8: Examples of models trained in Zh2EnSum dataset

Figure 9: Visualization of all the 48 self attention heads. The x-axis and y-axis are both concatenated source-language summary S^A and target-language summary S^B tokens. Each row contains all of the attention heads of corresponding layer from bottom to the top. The darker color shows the more highly related associations between tokens.
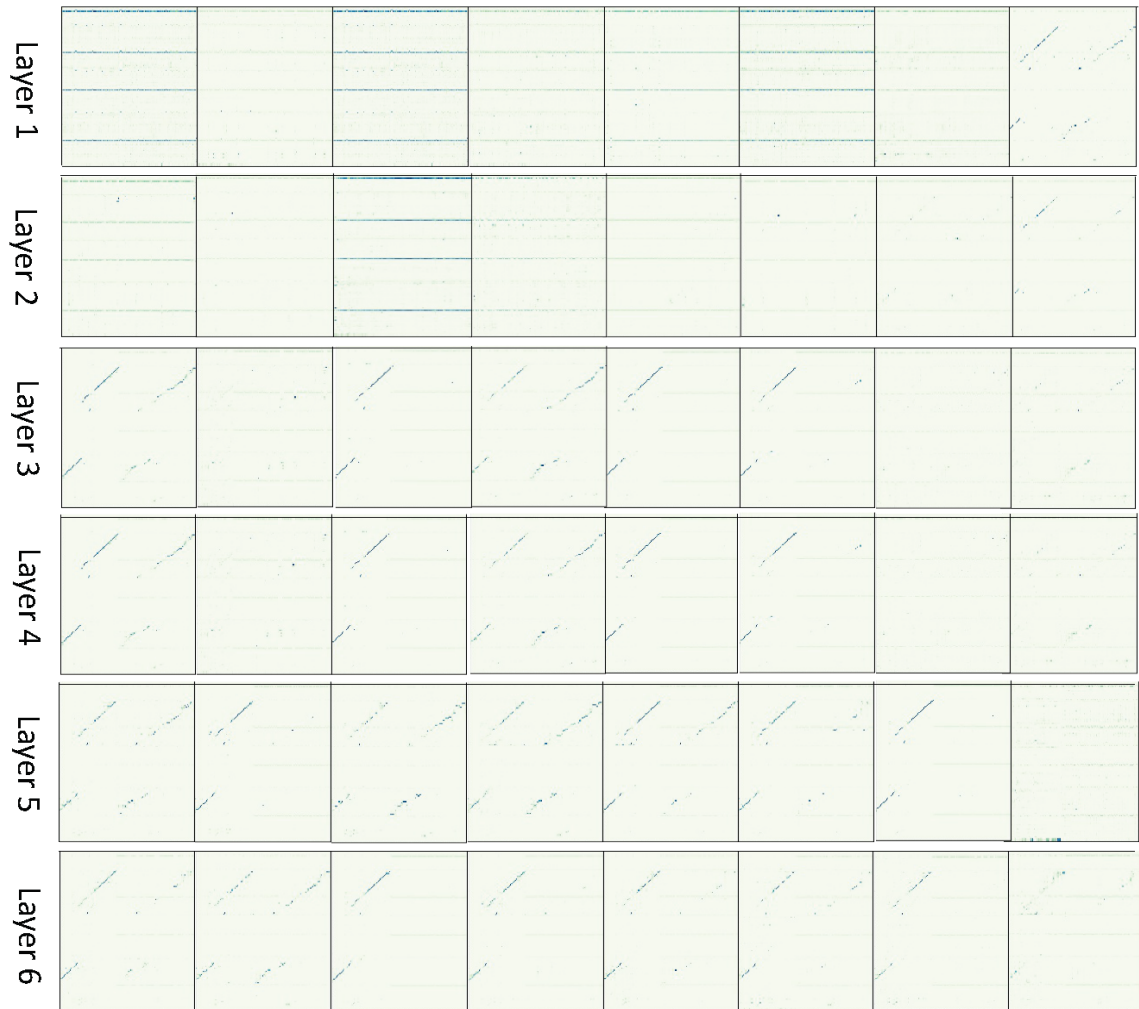
Figure 10: Visualization of all the 48 encoder-decoder attention heads. The x-axis is concatenated source-language summary S^A and target-language summary S^B tokens while the y-axis is document D^A tokens. Each row contains all of the attention heads of corresponding layer from bottom to the top. The darker color shows the more highly related associations between tokens.