Improving Neural Cross-Lingual Abstractive Summarization via Employing Optimal Transport Distance for Knowledge Distillation

Thong Thanh Nguyen[1], **Anh Tuan Luu**[2*]

[1] VinAI Research, Vietnam
[2] Nanyang Technological University, Singapore

Abstract

Current state-of-the-art cross-lingual summarization models employ multi-task learning paradigm, which works on a shared vocabulary module and relies on the self-attention mechanism to attend among tokens in two languages. However, correlation learned by self-attention is often loose and implicit, inefficient in capturing crucial cross-lingual representations between languages. The matter worsens when performing on languages with separate morphological or structural features, making the cross-lingual alignment more challenging, resulting in the performance drop. To overcome this problem, we propose a novel Knowledge-Distillation-based framework for Cross-Lingual Summarization, seeking to explicitly construct cross-lingual correlation by distilling the knowledge of the monolingual summarization teacher into the cross-lingual summarization student. Since the representations of the teacher and the student lie on two different vector spaces, we further propose a Knowledge Distillation loss using Sinkhorn Divergence, an Optimal-Transport distance, to estimate the discrepancy between those teacher and student representations. Due to the intuitively geometric nature of Sinkhorn Divergence, the student model can productively learn to align its produced cross-lingual hidden states with monolingual hidden states, hence leading to a strong correlation between distant languages. Experiments on cross-lingual summarization datasets in pairs of distant languages demonstrate that our method outperforms state-of-the-art models under both high and low-resourced settings.

Introduction

Cross-Lingual Summarization (CLS) is the task of condensing a document of one language into its shorter form in the target language. Most of contemporary works can be classified into two categories, i.e. low-resourced and high-resourced CLS approaches. In high-resourced scenarios, models are provided with an enormous number of document /summary pairs on which they can be trained (Zhu et al. 2019; Cao, Liu, and Wan 2020; Zhu et al. 2020). On the other hand, in low-resourced settings, those document/summary pairs are scarce, which restrains the amount of information that a model can learn. While high-resourced settings are preferred, in reality it is difficult to attain a sufficient amount of data, especially for less prevalent languages.

Most previous works resolving the issue of little training data concentrate on multi-task learning framework by utilizing the relationship of Cross-Lingual Summarization (CLS) with Monolingual Summarization (MLS) or Neural Machine Translation (NMT). Their approach can be further divided into two groups. The first group equips their module with two independent decoders, one of them targets the auxiliary task (MLS or NMT). Nevertheless, since two decoders do not share their parameters, this approach undermines the model's ability to align between two tasks (Bai, Gao, and Huang 2021), making the ancillary and the main task less relied upon each other. Hence, the trained model might produce output that does not match up the topic, or miss important spans of text.

The second group decides to employ a single decoder dealing with both CLS and MLS tasks. To this end, the method concatenates the monolingual to cross-lingual summary and designate the model to sequentially generate the monolingual summary, and then the cross-lingual one. Unfortunately, notwithstanding lessening the computational overhead during training by using solely one decoder, this method is not efficacious in capturing the connection between two languages in the output, consequently producing representations that do not take into account language relationships (Luo et al. 2021). In that case, the correlation of cross-lingual representations will be tremendously impacted by the structural and morphological similarity of those languages (Bjerva et al. 2019). As a result, in case of summarizing the document from one language to another that possesses distinct morphology and structure properties, such as from Chinese to English, the decoder might be prone to underperformance, due to the dearth of language correlation between two sets of hidden representations in the bilingual vector space (Luo et al. 2021).

To solve the aforementioned problem, we propose a novel Knowledge-Distillation framework for Cross-Lingual Summarization task. Particularly, our framework consists of a teacher model targetting Monolingual Summarization, and a student for Cross-Lingual Summarization. We initiate our procedure by finetuning the teacher model on monolingual document/summary pairs. Subsequently, we continue to distill summarization knowledge of the trained teacher

*Corresponding author

into the student model. Because the hidden vectors of the teacher and student lie upon two disparate monolingual and cross-lingual spaces, respectively, we propose a Sinkhorn-Divergence-based Knowledge Distillation loss, for the distillation process. Whereas multiple distances such as Cosine Distance or Euclidean Distance demand two sets share the sample size and are sensitive to outliers (Zimek, Schubert, and Kriegel 2012), Sinkhorn divergence does not enforce any requirement that relates to the number of samples and is also robust to noise (Séjourné et al. 2019). Furthermore, compared with other types of divergences such as KL divergence, the computation of Sinkhorn divergence does not require two distributions to lie on the same probability space. This is important because two languages might possess distinct features that cannot be projected one-to-one, such as the vocabulary set. Consequently, employing divergences different from Sinkhorn would need additional constraint to the distillation loss. Lastly, Sinkhorn divergence is able to capture geometric nature (Feydy et al. 2019) which has been shown to benefit myriad cross-lingual and multilingual representation learning settings (Huang et al. 2021). We will empirically prove the superiority of Sinkhorn divergence in the Experiment section.

Since the proposed module perpetuates the one-decoder employment, our framework is able to explicitly correlate representations from two languages, thus resolving the issue of two distant languages without demanding any additional computation overhead. To evaluate the efficacy of our framework, we proceed to conduct the experiments on myriad datasets containing document /summary pairs of couples of distant languages, for example, English-to-Chinese, English-to-Arabic, Japanese-to-English, etc. The empirical results demonstrate that our model outperforms previous state-of-the-art Cross-Lingual Summarization approaches. In sum, our contributions are three-fold:

- We propose a Knowledge Distillation framework for Cross-Lingual Summarization task, which seeks to enhance the summarization performance on distant languages by aligning the cross-lingual with monolingual summarization, through distilling the knowledge of monolingual teacher into cross-lingual student model.
- We propose a novel Knowledge Distillation loss using Optimal-Transport distance, i.e. Sinkhorn Diveregence, with a view to coping with the spatial discrepancy formed by the hidden representations produced by teacher and student model.
- We conducted extensive experiments in both high and low-resourced settings on multiple Cross-Lingual Summarization datasets that belong to pairs of morphologically and structurally distant languages, and found that our method significantly outperforms other baselines in both automatic metrics and by human evaluation.

Related Work

Neural Cross-Lingual Summarization

Due to the advent of Transformer architecture with its self-attention mechanism, Text Generation has received ample attention from researchers (Tuan, Shah, and Barzilay 2020; Lyu et al. 2021; Zhang et al. 2021), especially Document Summarization (Zhang et al. 2020; Nguyen et al. 2021). In addition to Monolingual Summarization, Neural Cross-Lingual Summarization has been receiving a tremendous amount of interest, likely due to the burgeoning need in cross-lingual information processing.

Conventional approaches designate a pipeline in two manners. The first one is translate-then-summarize, which copes with the task by initially translating the document into the target language and then performing the summarization (Wan, Li, and Xiao 2010; Ouyang, Song, and McKeown 2019; Wan 2011; Zhang, Zhou, and Zong 2016). The second approach is summarize-then-translate, which firstly summarizes the document and then creates its translated version in the target language (Lim, Kang, and Lee 2004; Orăsan and Chiorean 2008; Wan, Li, and Xiao 2010). Nonetheless, both of these approaches are vulnerable to error propagation caused by undertaking multiple steps (Zhu et al. 2019).

Recent works apply a general architecture combined with large-scale training to conduct Cross-Lingual Summarization. The main approach is to utilize the multi-task framework, in which CLS task benefits from the process of making use of other tasks such as Monolingual Summarization or Machine Translation (Zhu et al. 2019). Further approaches design ancillary mechanisms such as pointer-generator to exploit the translation scheme in the cross-lingual summary (Zhu et al. 2020). Other work uses a pair of encoders and decoders to co-operate the cross-lingual alignment with summarization (Cao, Liu, and Wan 2020).

Optimal Transport in Natural Language Processing

Introduced in 19th century as a method to find the optimal solution to transport a mass from one place to another destination, researchers have found its use in a wide variety of scientific fields, such as computational fluid mechanics (Benamou and Brenier 2000), economics (Carlier, Oberman, and Oudet 2015), physics (Cole et al. 2021), and notably machine learning (Peyré, Cuturi et al. 2019; Cuturi 2013; Courty et al. 2016; Danila et al. 2006).

Recently, beside Contrastive Learning framework (Nguyen and Luu 2021; Pan et al. 2021a,b), Optimal Transport has been omnivorously employed in Natural Language Processing field, as used through Optimal Transport distance, for instance Word Mover's Distance (Werner and Laber 2019), to estimate the necessary quantity of alignment. Its application includes text classification (Kusner et al. 2015), capturing spatial alignment in word embedding (Alvarez-Melis and Jaakkola 2018), machine translation (Chen et al. 2019), abstractive summarization (Chen et al. 2019), etc. Nevertheless, the adaptation of Optimal Transport distance, especially Sinkhorn divergence, for Neural Cross-Lingual Summarization task has been attracting limited amount of research effort.

Background

Neural Cross-Lingual Summarization

Given a document $X^{L_1} = \{x_1, x_2, \ldots, x_N\}$, a monolingual summarization model's task is to create a summary $Y^{L_1} = \{y_1^{L_1}, y_2^{L_1}, \ldots, y_{M_1}^{L_1}\}$, where both X^{L_1} and Y^{L_1} are in language L_1. On the contrary, a cross-lingual summarization model will produce a cross-lingual summary $Y^{L_2} = \{y_1^{L_2}, y_2^{L_2}, \ldots, y_{M_2}^{L_2}\}$ that is in language L_2. It is worth noting here that $M_1 < N$ and $M_2 < N$.

Analogous to monolingual summarization, current state-of-the-art cross-lingual summarization methods employ the Transformer-based architecture. Relying mainly on self-attention mechanism, Transformer-based architecture consists of an encoder and a decoder. The bidirectional self-attention in the encoder will extract contextualized representations of the input, which will be fed to the decoder to generate the output. Due to its generation nature, the decoder will use unidirectional self-attention to learn the context of previously generated tokens. During training procedure, the whole framework is updated based upon the cross-entropy loss as follows

$$\mathcal{L}_{\text{CLS}} = -\sum_{t=1}^{M_2} \log P(y_t^{L_2} | y_{<t}^{L_2}, X^{L_1}) \tag{1}$$

Knowledge Distillation (KD)

Proposed by (Hinton, Vinyals, and Dean 2015), knowledge distillation is a method to train a model, called the student, by leveraging valuable information provided by soft targets output by another model, called the teacher. In particular, the framework initially trains a model on one designated task to extract useful features. Subsequently, given a dataset $D = \{(X_1, Y_1), (X_2, Y_2), \ldots (X_{|D|}, Y_{|D|})\}$, where $|D|$ is the size of the dataset, the teacher model will generate the output $H_i^T = \{\mathbf{h}_1^T, \mathbf{h}_2^T, \ldots, \mathbf{h}_{L_T}^T\}$ for each input X_i. Dependent on the researchers' decision, the output might be hidden representations or final logits. As a consequence, in order to train the student model, the framework will use a KD loss that discriminates the output of the student model $H_i^S = \{\mathbf{h}_1^S, \mathbf{h}_2^S, \ldots, \mathbf{h}_{L_S}^S\}$ given input X_i from the teacher output H_i^T. Eventually, the KD loss for input X_i will possess the form as follows

$$\mathcal{L}_{\text{KD}} = dist(H_i^T, H_i^S) \tag{2}$$

where $dist$ is a distance function to estimate the discrepancy of teacher and student outputs.

The explicated Knowledge Distillation framework has shown its efficiency in a tremendous number of tasks, such as Neural Machine Translation (Tan et al. 2019; Wang et al. 2021; Li and Li 2021; Sun et al. 2020), Question Answering (Hu et al. 2018; Arora, Khapra, and Ramaswamy 2019; Yang et al. 2020b), Image Classification (Yang et al. 2020a; Chen, Chang, and Lee 2018; Fu et al. 2020), etc. Nonetheless, its application for Neural Cross-Lingual Summarization has received little interest.

Methodology

To resolve the issue of distant languages, the output representations from two vector spaces denoting two languages should be indistinguishable, or easily transported from one space to another. In order to accomplish that goal, we seek to relate the cross-lingual output of the student model to the monolingual output of the teacher model, via utilizing Knowledge Distillation framework and Sinkhorn Divergence calculation. The complete framework is illustrated in Figure 1.

Knowledge Distillation Framework for Cross-Lingual Summarization

We inherit the architecture of Transformer model for our module. In particular, both the teacher and student model uses the encoder-architecture paradigm combined with two fundamental mechanisms. Firstly, the self-attention mechanism will attempt to learn the context of the tokens by attending tokens among each other in the input and output document. Secondly, there is a cross-attention mechanism to correlate the contextualized representations of the output tokens to ones of the input tokens.

In our KD framework, we initiate the process by training the teacher model on monolingual summarization task. In detail, given an input $X^{L_1} = \{x_1, x_2, \ldots, x_N\}$, the teacher model will aim to generate its monolingual summary $Y^{L_1} = \{y_1^{L_1}, y_2^{L_1}, \ldots, y_{M_1}^{L_1}\}$. Similar to previous monolingual summarization schemes, our model is trained by maximizing the likelihood of the groundtruth tokens, which takes the cross-entropy form as follows

$$\mathcal{L}_{\text{MLS}} = -\sum_{t=1}^{M_1} \log P(y_t^{L_1} | y_{<t}^{L_1}, X^{L_1}) \tag{3}$$

After finetuning the teacher model, we progress to train the student model, which also employs the Transformer architecture. Contrary to the teacher, the student model's task is to generate the cross-lingual output $Y^{L_2} = \{y_1^{L_2}, y_2^{L_2}, \ldots, y_{M_2}^{L_2}\}$ in language L_2, given the input document X^{L_1} in language L_1. We update the parameters of the student model by minimizing the objective function that is formulated as follows

$$\mathcal{L}_{\text{CLS}} = -\sum_{t=1}^{M_2} \log P(y_t^{L_2} | y_{<t}^{L_2}, X^{L_1}) \tag{4}$$

With a view to pulling the cross-lingual and monolingual representations nearer, we implement a KD loss to penalize the large distance of two vector spaces. Particularly, let $H^T = \{\mathbf{h}_1^T, \mathbf{h}_2^T, \ldots, \mathbf{h}_{L_T}^T\}$ denote the contextualized representations produced by the decoder of the teacher model, and $H^S = \{\mathbf{h}_1^S, \mathbf{h}_2^S, \ldots, \mathbf{h}_{L_S}^S\}$ denote the representations from the decoder of the student model, we define our KD loss as follows

$$L_{\text{KD}} = dist(H^T, H^S) \tag{5}$$

where $dist$ is the Optimal-Transport distance to evaluate the difference of two representations, which we will delineate in the following section.

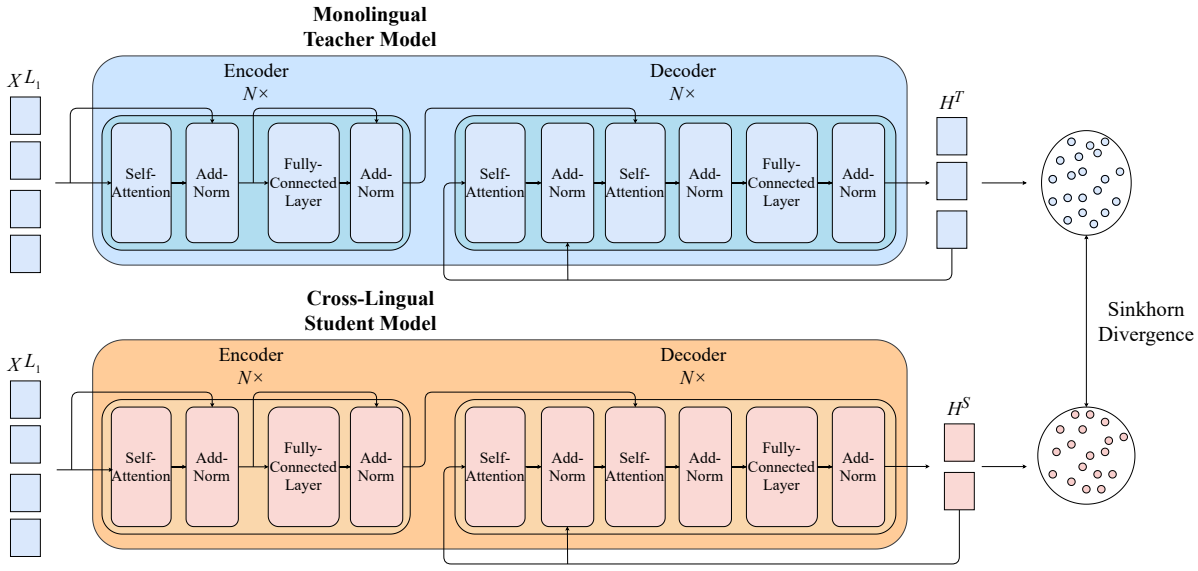

Figure 1: Diagram of Knowledge Distillation Framework for Cross-Lingual Summarization

Sinkhorn Divergence for Knowledge Distillation Loss

Due to the dilemma that the hidden representations of the teacher and student model stay upon two disparate vector spaces (as they represent two different languages), we will consider the distance of the two spaces as the distance of two probability measures. To elaborate, we choose to adapt Sinkhorn divergence, a variant of Optimal Transport distance, to calculate the aforementioned spatial discrepancy. Let H^T, H^S denote the representations of the teacher decoder and the student decoder, we encode the sample measures of them

$$\alpha = \sum_{i=1}^{L_T} \alpha_i \delta_{\mathbf{h}_i^T}, \quad \beta = \sum_{j=1}^{L_S} \beta_j \delta_{\mathbf{h}_j^S} \tag{6}$$

where α and β are probability distributions that satisfy $\sum_{i=1}^{L_T} \alpha_i = 1$ and $\sum_{j=1}^{L_S} \beta_j = 1$.

Inspired by (Feydy et al. 2019), we estimate the difference of the representations through determining the Sinkhorn divergence between them

$$dist(H^T, H^S) = \text{OT}(\alpha, \beta) - \frac{1}{2}\text{OT}(\alpha, \alpha) - \frac{1}{2}\text{OT}(\beta, \beta) \tag{7}$$

where

$$\text{OT}(\alpha, \beta) = \sum_{i=1}^{N} \alpha_i f_i + \sum_{j=1}^{M} \beta_j g_j \tag{8}$$

in which f_i, g_j are estimated by Sinkhorn loop. We thoroughly delineate the loop in Algorithm 1.

Training Objective

We amalgamate the Cross-Lingual Summarization and Knowledge Distillation objective to obtain the ultimate objective function. Mathematically, for each input, our training loss is computed as follows

$$\mathcal{L} = \mathcal{L}_{\text{CLS}} + \lambda \cdot \mathcal{L}_{\text{KD}} \tag{9}$$

where λ is the hyperparameter that controls the influence of the cross-lingual alignment of two vector spaces.

Algorithm 1: Sinkhorn loop

Input: Probability distributions α, β, regularization hyperparameter ε, number of iterations N_I, log-sum-entropy function $\text{LSE}_{k=1}^{N}(z_k) = \log \sum_{k=1}^{N} \exp(z_k)$, distance function $C(\mathbf{x}, \mathbf{y}) = ||\mathbf{x} - \mathbf{y}||^2$

1: **for** $i = 1$ to N_I **do**
2: Compute $f_i = \varepsilon \cdot \text{LSE}_{k=1}^{L_S}[\log(\beta_k) + \frac{1}{\varepsilon} g_k - \frac{1}{\varepsilon} C(\mathbf{h}_i^T, \mathbf{h}_k^S)]$
3: Compute $g_j = \varepsilon \cdot \text{LSE}_{k=1}^{L_T}[\log(\alpha_k) + \frac{1}{\varepsilon} f_k - \frac{1}{\varepsilon} C(\mathbf{h}_k^T, \mathbf{h}_j^S)]$
4: **end for**

Experiments

Datasets

We evaluate the effectiveness of our methods on En2Zh and Zh2En datasets processed by (Bai, Gao, and Huang 2021). We also inherit their minimum, medium, and maximum settings in order to verify the effectiveness of our method under limited-resourced settings. The sample size under each setting is depicted in Table 2. Furthermore, to further evaluate the performance of our method in various languages, we also preprocess datasets of Wikilingua (Ladhak et al. 2020) in the manner that every sample is converted to a triple of document, MLS summary, and CLS summary. We choose 4 variants of Wikilingua to proceed our evaluation, i.e. English to Arabic (En2Ar), English to Japanese (En2Ja), Japanese to English (Ja2En), and English to Vietnamese (En2Vi). It

should be noted here that (En, Ja), (En, Ar), (En, Zh), and (En, Vi) are all couples of languages that are distant in terms of structure or morphology. The statistics of the datasets is demonstrated in Table 1.

Dataset	l_{Input}	l_{CLS}	l_{MLS}
Zh2En	105	19	19
En2Zh	912	97	69
En2Ar	1589	227	133
En2Ja	1463	212	133
Ja2En	2103	133	212
En2Vi	1657	175	135

Table 1: Statistics of Cross-Lingual Summarization datasets.

Scenarios	Minimum	Medium	Maximum	Full-dataset
Zh2En	5,000	25,000	50,000	1,693,713
En2Zh	1,500	7,500	15,000	364,687

Table 2: Dataset sizes of multiple low-resource scenarios for CLS datasets.

Implementation Details

We initialize the encoder with multilingual BERT (Devlin et al. 2018), whereas the decoder with Xavier intialization (Glorot and Bengio 2010). The dimensions of our encoder and decoder hidden states are both 768. We use two seperate Adam optimizers for encoder and decoder, and the learning rate for encoder and decoder is 0.002 and 0.2, respectively. The model is trained with the warmup phase of 25000 steps. We train the model on one Nvidia GeForce A100 GPU that accumulates gradient every 5 steps. Moreover, we apply Dropout probability of 0.1 to all fully-connected layers in the model. The teacher and student model shares the architecture and scale of parameters in our Knowledge Distillation framework. To estimate the Sinkhorn divergence, we employ the entropic regularization rate ε of 0.0025 and the iteration length N_I of 14. The weight λ of KD Loss in Equation 9 is set to 1.

Baselines

We compare our proposed architecture against the following baselines:

- **NCLS** (Zhu et al. 2019): a Transformer-based model to conduct CLS.
- **NCLS + MS** (Zhu et al. 2019): a multi-task framework that leverages an auxiliary MS decoder to enhance cross-lingual summarization performance.
- **TLTran** (Bai, Gao, and Huang 2021): a CLS pipeline that firstly performs MLS and then utilizes a finetuned NMT model to translate the monolingual summary into the target language.
- **MCLAS** (Bai, Gao, and Huang 2021): a multi-task framework that sequentially performs MLS, and CLS which is based upon the MLS result.

Automatic Evaluation

Full-dataset Scenario The experimental results under the full-dataset scenario are given in Table 3, 4, 5, 6, 7, and 8.

For Zh2En dataset, our method outperforms MCLAS model by ROUGE-1 of 1.3 points, ROUGE-2 of 4.0 points, ROUGE-3 of 0.4 point, and ROUGE-L of 1.2 points. Our model also improves the performance of NCLS model for dataset En2Zh, with 0.6 point in ROUGE-1, 1.5 points in ROUGE-2, 0.1 point in ROUGE-3, and 0.8 point in ROUGE-L. For Arabic language, our model achieves the enhancement compared against NCLS model by 0.1 in ROUGE-1 score, 2.9 in ROUGE-2 score, 1.6 in ROUGE-3 score, and 5.1 in ROUGE-L score. In En2Ja dataset, we outperformed previous best method MCLAS by 0.6 point in ROUGE-1, 0.2 point in ROUGE-2, 0.2 point in ROUGE-3, and 0.5 point in ROUGE-L.

Additionally, for the reverse dataset Ja2En, our method significantly achieves higher performance with the improvement of 1.0 point of ROUGE-1, 0.5 point of ROUGE-2, 0.4 point of ROUGE-3, and 0.4 point of ROUGE-L, compared with MCLAS model. Those results substantiate our hypothesis that our framework is able to enhance the capability of apprehending and summarizing a document into a summary of another distant language, since English alphabet does not have any character in common with Japanese, Arabic, and Chinese counterparts.

For En2Vi dataset, our method also obtains notable improvement over other state-of-the-art methods. As shown in Table 8, our model outperforms MCLAS model by 0.1 in ROUGE-1, 2.9 in ROUGE-2, 1.6 in ROUGE-3, and 5.1 in ROUGE-L. This demonstrates that our method is also capable of buttressing the model capacity in situations where two languages are slightly morphologically or structurally similar, since Vietnamese and English do share a number of characters in their alphabets.

Model	R1	R2	R3	RL
TLTran	33.64	15.58	12.02	29.74
NCLS	35.60	16.78	12.57	30.27
NCLS+MS	34.84	16.05	12.28	29.47
MCLAS	35.65	16.97	12.78	31.14
Our Model	**36.93**	**20.99**	**13.20**	**32.33**

Table 3: Full-dataset Cross-Lingual Summarization results in Zh2En dataset

Model	R1	R2	R3	RL
TLTran	30.20	12.20	11.79	27.02
NCLS	44.16	24.28	17.13	30.23
NCLS+MS	42.68	23.51	15.62	29.24
MCLAS	42.27	24.60	16.07	30.09
Our Model	**44.75**	**25.76**	**17.20**	**31.05**

Table 4: Full-dataset Cross-Lingual Summarization results in En2Zh dataset

Model	R1	R2	R3	RL
NCLS	36.80	17.36	10.79	27.25
NCLS+MS	35.53	17.01	10.33	26.36
MCLAS	36.28	17.27	10.81	27.56
Our Model	**36.89**	**20.28**	**12.40**	**32.38**

Table 5: Full-dataset Cross-Lingual Summarization results in En2Ar dataset

Model	R1	R2	R3	RL
NCLS	29.55	15.99	10.25	23.03
NCLS+MS	29.42	15.83	10.12	23.00
MCLAS	29.60	16.08	10.14	33.20
Our Model	**30.21**	**16.27**	**10.46**	**23.90**

Table 6: Full-dataset Cross-Lingual Summarization results in En2Ja dataset

Model	R1	R2	R3	RL
NCLS	32.78	12.66	6.33	26.43
NCLS+MS	32.50	12.02	6.15	26.41
MCLAS	33.20	12.57	6.33	27.27
Our Model	**34.21**	**13.08**	**6.70**	**27.63**

Table 7: Full-dataset Cross-Lingual Summarization results in Ja2En dataset

Model	R1	R2	R3	RL
NCLS	36.75	16.37	8.04	28.69
NCLS+MS	36.28	16.14	8.03	28.61
MCLAS	36.31	15.91	7.75	28.62
Our Model	**37.38**	**16.20**	**8.09**	**28.97**

Table 8: Full-dataset Cross-Lingual Summarization results in En2Vi dataset

Low-resource Scenario We denote results of the experiments conducted under minimum, medium, and maximum scenarios in Table 9, 10, and 11.

For the minimum setting, our model achieves the improvement over previous methods. In particular, we outperformed MCLAS model by 1.3 points of ROUGE-1, 0.5 point of ROUGE-2, 0.2 point of ROUGE-3, and 0.3 point of ROUGE-L in Zh2En dataset. For En2Zh dataset, we obtain an increase of 3.6 points in ROUGE-1, 0.6 point in ROUGE-2, 0.3 point in ROUGE-3, and 1.4 points in ROUGE-L.

Under the medium setting, the performance of our method is also higher than MCLAS model with 0.1 point in ROUGE-1, 1.1 points in ROUGE-2, 0.5 point in ROUGE-3, and 3.0 points in ROUGE-L. The improvement is more critical for dataset En2Zh with an increase of 3.0 in ROUGE-1, 1.9 in ROUGE-2, 0.6 in ROUGE-3, and 0.5 in ROUGE-L.

Last but not least, in maximum scenario, for dataset Zh2En, our gains compared against MCLAS model are 0.4 point in ROUGE-1, 0.4 point in ROUGE-2, 0.5 point in ROUGE-3, and 0.7 point in ROUGE-L. In dataset En2Zh, our improvements are 2.9 points in ROUGE-1, 0.3 point in ROUGE-2, 0.4 point in ROUGE-3, and 0.7 point in ROUGE-L.

Those aforementioned results have shown that our method is also capable of elevating the Cross-Lingual Summarization performance when the available training dataset is scarce.

Models	Zh2En	En2Zh
NCLS	20.93/5.88/2.47/17.58	34.14/12.45/4.38/21.20
NCLS+MS	20.50/5.45/2.22/17.25	33.96/12.38/4.36/21.07
MCLAS	21.03/6.03/2.68/18.16	32.03/13.17/4.28/21.17
Our Model	**22.37/6.50/2.91/18.47**	**35.59/13.77/4.57/22.56**

Table 9: Minimum Cross-Lingual Summarization Results

Models	Zh2En	En2Zh
NCLS	26.42/8.90/4.49/22.05	35.98/15.88/8.97/23.79
NCLS+MS	26.86/9.06/4.58/22.47	38.95/18.09/9.73/25.39
MCLAS	27.84/10.41/4.91/24.12	37.28/18.10/9.48/25.26
Our Model	**27.97/11.51/5.37/27.16**	**40.30/20.01/10.05/25.79**

Table 10: Medium Cross-Lingual Summarization Results

Models	En2Zh	Zh2En
NCLS	29.05/10.88/6.56/24.32	40.18/19.86/10.33/26.52
NCLS+MS	28.63/10.63/6.24/24.00	39.86/19.87/10.23/26.64
MCLAS	30.73/12.26/6.98/26.51	38.35/19.75/10.64/26.41
Our Model	**31.08/12.70/7.45/27.16**	**41.24/20.01/11.00/27.06**

Table 11: Maximum Cross-Lingual Summarization Results

Human Evaluation

Because automatic metrics do not completely betray the quality of the methods, we conduct further human evaluation for more precise assessment. To fulfil our objective, we design two tests in order to elicit human judgements in two manners.

In the first experiment, we present summaries generated by NCLS, MCLAS, our model, and the gold summary, then asked seven professional English speakers to indicate the best and worst summaries in terms of informativeness, faithfulness, topic coherence, and fluency. We randomly sampled 50 summaries from En2Vi dataset and 50 others from Ja2En dataset. The score of a model will be estimated as the percentage of times it was denoted as the best minus the percentage of times it was denoted as the worst.

For the second experiment, we decide to adapt Question Answering (QA) paradigm to our framework. For each sample, we create two independent questions that underscore the key information from the input document. Participants would read and answer each question as best as they could. The score of a system will be equal to the proportion of questions that the participants answer correctly.

Fleiss' Karpa scores of our experiments are shown in Table 12. It is obvious that the scores prove a strong inter-agreement among the participants.

The experimental results in Table 13 indicate that our model generates summaries that are conducive to human judgements, and have more likelihood to preserve important content in the original documents than summaries of other systems.

Test	Fleiss' Kappa	Overall Agreement
Preference	0.57	64.95%
QA	0.64	82.15%

Table 12: Fleiss' Kappa and Overall Agreement percentage of each human evaluation test. Higher score indicates better agreement.

Models	Preference Score	QA score
NCLS	-0.123	51.11
MCLAS	0.169	59.26
Our Model	0.498	71.85
Gold Summary	0.642	95.52

Table 13: Human evaluation

Analysis on Distance Methods

We compare our implemented Sinkhorn Divergence with other distance methods. Particularly, we perform the mean or max-pooling of the teacher and student hidden representations. Subsequently, we evaluate the teacher and student discrepancy via Cosine Similarity (CS) or Mean Squared Error (MSE) of two pooled vectors. We show the numerical results in Table 14. The results demonstrate the superiority of Sinkhorn Divergence over other approaches. We hypothesize that those approaches do not efficaciously capture the geometry nature of cross-lingual output representations.

Distance Methods	R-1	R-2	R-3	R-L
Mean-CS	44.20	24.54	16.96	30.27
Mean-MSE	44.14	24.27	16.22	30.19
Max-CS	44.29	25.65	17.07	30.82
Max-MSE	44.23	24.61	16.44	30.21
Our Method	**44.75**	**25.76**	**17.20**	**31.05**

Table 14: Results when applying different distance methods in En2Zh dataset under full-dataset setting.

Impact of Sinkhorn Divergence on Geometric Distance of Cross-Lingual Representations

We propose to adapt Sinkhorn Divergence to align the cross-lingual decoder hidden states of the student model with monolingual decoder hidden states of the teacher model. Nevertheless, whether this geometrically brings two sets of representations nearer remains a quandary. To further verify the benefit of leveraging Sinkhorn Divergence, we estimate the distances of those hidden vectors by using other metrics, i.e. Cosine Similarity and Mean Squared Error. Particularly, for each input, after getting the decoder to generate the hidden vectors of the output tokens, we take the average of those vectors and measure the distance between the mean of the vectors generated by the CLS model (NCLS, MCLAS, and Our Model) with the mean of the vectors created by the MLS model. We denote the expected value and standard deviation of each method in Table 15. As it can be obviously seen, employing Sinkhorn Divergence actually pulls the vectors in the cross-lingual spaces towards one another.

Models	Cosine Similarity	Mean Squared Error
NCLS	0.165 ± 0.038	19.434 ± 7.252
MCLAS	0.064 ± 0.057	17.207 ± 4.028
Our Model	$\mathbf{0.034 \pm 0.054}$	$\mathbf{13.517 \pm 4.013}$

Table 15: Results when applying different distance methods in Zh2En dataset under full-dataset setting.

Conclusion

In this paper, we propose a novel Knowledge Distillation framework to tackle Neural Cross-Lingual Summarization for morphologically or structurally distant languages. Our framework trains a monolingual teacher model, and then finetunes the cross-lingual student model which is distilled knowledge from the aforementioned teacher. Since the hidden representations of the teacher and student model lie upon two different lingual spaces, we continually proposed to adapt Sinkhorn Divergence to efficiently estimate the cross-lingual discrepancy. Extensive experiments show that our method significantly outperforms other approaches under both low-resourced and full-dataset settings.

References

Alvarez-Melis, D.; and Jaakkola, T. S. 2018. Gromov-Wasserstein alignment of word embedding spaces. *arXiv preprint arXiv:1809.00013*.

Arora, S.; Khapra, M. M.; and Ramaswamy, H. G. 2019. On knowledge distillation from complex networks for response prediction. 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)*, 3813–3822.

Bai, Y.; Gao, Y.; and Huang, H. 2021. Cross-Lingual Abstractive Summarization with Limited Parallel Resources. *arXiv preprint arXiv:2105.13648*.

Benamou, J.-D.; and Brenier, Y. 2000. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3): 375–393.

Bjerva, J.; Östling, R.; Veiga, M. H.; Tiedemann, J.; and Augenstein, I. 2019. What do language representations really represent? *Computational Linguistics*, 45(2): 381–389.

Cao, Y.; Liu, H.; and Wan, X. 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*, 6220–6231.

Carlier, G.; Oberman, A.; and Oudet, E. 2015. Numerical methods for matching for teams and Wasserstein barycenters. *ESAIM: Mathematical Modelling and Numerical Analysis*, 49(6): 1621–1642.

Chen, L.; Zhang, Y.; Zhang, R.; Tao, C.; Gan, Z.; Zhang, H.; Li, B.; Shen, D.; Chen, C.; and Carin, L. 2019. Improving sequence-to-sequence learning via optimal transport. *arXiv preprint arXiv:1901.06283*.

Chen, W.-C.; Chang, C.-C.; and Lee, C.-R. 2018. Knowledge distillation with feature maps for image classification. In *Asian Conference on Computer Vision*, 200–215. Springer.

Cole, S.; Eckstein, M.; Friedland, S.; and Życzkowski, K. 2021. Quantum Optimal Transport. *arXiv preprint arXiv:2105.06922*.

Courty, N.; Flamary, R.; Tuia, D.; and Rakotomamonjy, A. 2016. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9): 1853–1865.

Cuturi, M. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26: 2292–2300.

Danila, B.; Yu, Y.; Marsh, J. A.; and Bassler, K. E. 2006. Optimal transport on complex networks. *Physical Review E*, 74(4): 046106.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Feydy, J.; Séjourné, T.; Vialard, F.-X.; Amari, S.-i.; Trouvé, A.; and Peyré, G. 2019. Interpolating between optimal transport and MMD using Sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, 2681–2690. PMLR.

Fu, S.; Li, Z.; Xu, J.; Cheng, M.-M.; Liu, Z.; and Yang, X. 2020. Interactive knowledge distillation. *arXiv preprint arXiv:2007.01476*.

Glorot, X.; and Bengio, Y. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 249–256. JMLR Workshop and Conference Proceedings.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Hu, M.; Peng, Y.; Wei, F.; Huang, Z.; Li, D.; Yang, N.; and Zhou, M. 2018. Attention-guided answer distillation for machine reading comprehension. *arXiv preprint arXiv:1808.07644*.

Huang, K.-H.; Ahmad, W. U.; Peng, N.; and Chang, K.-W. 2021. Improving Zero-Shot Cross-Lingual Transfer Learning via Robust Training. *arXiv preprint arXiv:2104.08645*.

Kusner, M.; Sun, Y.; Kolkin, N.; and Weinberger, K. 2015. From word embeddings to document distances. In *International conference on machine learning*, 957–966. PMLR.

Ladhak, F.; Durmus, E.; Cardie, C.; and McKeown, K. 2020. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. *arXiv preprint arXiv:2010.03093*.

Li, Y.; and Li, W. 2021. Data Distillation for Text Classification. *arXiv preprint arXiv:2104.08448*.

Lim, J.-M.; Kang, I.-S.; and Lee, J.-H. 2004. Multi-Document Summarization Using Cross-Language Texts. In *NTCIR*.

Luo, F.; Wang, W.; Liu, J.; Liu, Y.; Bi, B.; Huang, S.; Huang, F.; and Si, L. 2021. VECO: Variable and Flexible Cross-lingual Pre-training for Language Understanding and Generation. 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)*, 3980–3994.

Lyu, C.; Shang, L.; Graham, Y.; Foster, J.; Jiang, X.; and Liu, Q. 2021. Improving Unsupervised Question Answering via Summarization-Informed Question Generation. *arXiv preprint arXiv:2109.07954*.

Nguyen, T.; and Luu, A. T. 2021. Contrastive Learning for Neural Topic Model. *Advances in Neural Information Processing Systems*, 34.

Nguyen, T.; Luu, A. T.; Lu, T.; and Quan, T. 2021. Enriching and controlling global semantics for text summarization. *arXiv preprint arXiv:2109.10616*.

Orăsan, C.; and Chiorean, O. A. 2008. Evaluation of a Cross-lingual Romanian-English Multi-document Summariser. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.

Ouyang, J.; Song, B.; and McKeown, K. 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)*, 2025–2031.

Pan, L.; Hang, C.-W.; Sil, A.; Potdar, S.; and Yu, M. 2021a. Improved Text Classification via Contrastive Adversarial Training. *arXiv preprint arXiv:2107.10137*.

Pan, X.; Wang, M.; Wu, L.; and Li, L. 2021b. Contrastive learning for many-to-many multilingual neural machine translation. *arXiv preprint arXiv:2105.09501*.

Peyré, G.; Cuturi, M.; et al. 2019. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6): 355–607.

Séjourné, T.; Feydy, J.; Vialard, F.-X.; Trouvé, A.; and Peyré, G. 2019. Sinkhorn divergences for unbalanced optimal transport. *arXiv preprint arXiv:1910.12958*.

Sun, H.; Wang, R.; Chen, K.; Utiyama, M.; Sumita, E.; and Zhao, T. 2020. Knowledge distillation for multilingual unsupervised neural machine translation. *arXiv preprint arXiv:2004.10171*.

Tan, X.; Ren, Y.; He, D.; Qin, T.; Zhao, Z.; and Liu, T.-Y. 2019. Multilingual neural machine translation with knowledge distillation. *arXiv preprint arXiv:1902.10461*.

Tuan, L. A.; Shah, D.; and Barzilay, R. 2020. Capturing greater context for question generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 9065–9072.

Wan, X. 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*, 1546–1555.

Wan, X.; Li, H.; and Xiao, J. 2010. Cross-language document summarization based on machine translation quality prediction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 917–926.

Wang, F.; Yan, J.; Meng, F.; and Zhou, J. 2021. Selective Knowledge Distillation for Neural Machine Translation. *arXiv preprint arXiv:2105.12967*.

Werner, M.; and Laber, E. 2019. Speeding up Word Mover's Distance and its variants via properties of distances between embeddings. *arXiv preprint arXiv:1912.00509*.

Yang, J.; Martinez, B.; Bulat, A.; and Tzimiropoulos, G. 2020a. Knowledge distillation via adaptive instance normalization. *arXiv preprint arXiv:2003.04289*.

Yang, Z.; Shou, L.; Gong, M.; Lin, W.; and Jiang, D. 2020b. Model compression with two-stage multi-teacher knowledge distillation for web question answering system. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, 690–698.

Zhang, A.; Wu, K.; Wang, L.; Li, Z.; Xiao, X.; Wu, H.; Zhang, M.; and Wang, H. 2021. Data Augmentation with Hierarchical SQL-to-Question Generation for Cross-domain Text-to-SQL Parsing. *arXiv preprint arXiv:2103.02227*.

Zhang, J.; Zhao, Y.; Saleh, M.; and Liu, P. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, 11328–11339. PMLR.

Zhang, J.; Zhou, Y.; and Zong, C. 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.

Zhu, J.; Wang, Q.; Wang, Y.; Zhou, Y.; Zhang, J.; Wang, S.; and Zong, C. 2019. NCLS: Neural cross-lingual summarization. *arXiv preprint arXiv:1909.00156*.

Zhu, J.; Zhou, Y.; Zhang, J.; and Zong, C. 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*, 1309–1321.

Zimek, A.; Schubert, E.; and Kriegel, H.-P. 2012. A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 5(5): 363–387.