electronics

Article

Oversea Cross-Lingual Summarization Service in Multilanguage Pre-Trained Model through Knowledge Distillation

Xiwei Yang, Jing Yun *, Bofei Zheng, Limin Liu and Qi Ban

School of Data Science and Applications, Inner Mongol University of Technology, Hohhot 010080, China; 20211800677@imut.edu.com (X.Y.); ax979710378@126.com (B.Z.); liulimin@imut.edu.com (L.L.); 20221800740@imut.edu.com (Q.B.)
* Correspondence: yunjing@imut.edu.cn

Abstract: Cross-lingual text summarization is a highly desired service for overseas report editing tasks and is formulated in a distributed application to facilitate the cooperation of editors. The multilanguage pre-trained language model (MPLM) can generate high-quality cross-lingual text summaries with simple fine-tuning. However, the MPLM does not adapt to complex variations, like the word order and tense in different languages. When the model performs on these languages with separate syntactic structures and vocabulary morphologies, it will lead to the low-level quality of the cross-lingual summary. The matter worsens when the cross-lingual summarization datasets are low-resource. We use a knowledge distillation framework for the cross-lingual summarization task to address the above issues. By learning the monolingual teacher model, the cross-lingual student model can effectively capture the differences between languages. Since the teacher and student models generate summaries in two languages, their representations lie on different vector spaces. In order to construct representation relationships across languages, we further propose a similarity metric, which is based on bidirectional semantic alignment, to map different language representations to the same space. In order to improve the quality of cross-lingual summaries further, we use contrastive learning to make the student model focus on the differentials among languages. Contrastive learning can enhance the ability of the similarity metric for bidirectional semantic alignment. Our experiments show that our approach is competitive in low-resource scenarios on cross-language summarization datasets in pairs of distant languages.

Keywords: multilingual pre-trained language model; cross-lingual summary; knowledge distillation; similarity metric

Citation: Yang, X.; Yun, J.; Zheng, B.; Liu, L.; Ban, Q. Oversea Cross-Lingual Summarization Service in Multilanguage Pre-Trained Model through Knowledge Distillation. *Electronics* **2023**, *12*, 5001. https://doi.org/10.3390/electronics12245001

Academic Editor: Arkaitz Zubiaga

Received: 3 November 2023
Revised: 8 December 2023
Accepted: 12 December 2023
Published: 14 December 2023

1. Introduction

The cross-lingual summarization (CLS) task emerged to convert documents from one language to a summary in another. It is a highly desired service in foreign report editing and formulated in a distributed application to facilitate the cooperation of editors. This service plays a crucial role in report writing on a global scale, facilitating collaboration and communication among editorial teams through concise summaries and multilingual translations.

MPLMs, such as mBART [1] and mT5 [2], have led to significant breakthroughs in the cross-lingual summarization task. In particular, ref. [3] discovers that the mBART (mbart.cc25) model outperforms many multitask models on large-scale cross-lingual summary datasets through simple fine-tuning. However, ref. [4] proposes that the syntactic structure of the language will still influence the relevance of cross-linguistic representations of models trained on multilingual corpora. Therefore, for the mBART model, learning feature relationships between two languages with separate syntactic structures and vocabulary morphologies is challenging. Specifically, the mBART model will be pre-trained in multiple languages. When generating a cross-lingual summary of two languages, the model may mistake the lingual features in the pre-trained corpus for features of two languages.

Constructing cross-lingual relationships between two languages with separate syntactic structures and vocabulary morphologies is more challenging. Significantly, the matter worsens in low-resource scenarios where the cross-lingual summary dataset contains only a few languages and a limited number of samples for each language.

To address the above issues, we propose a knowledge distillation framework to mitigate cross-linguistic interference that can improve the quality of the generated target language summaries. Our model consists of a monolingual teacher model and a cross-lingual student model. By distilling the attention weight and word distribution of the fine-tuned teacher model into the student model, the student model can effectively construct cross-lingual relationships in low-resource scenarios. To improve the quality of the cross-lingual summary, we propose a text similarity metric to evaluate the similarity of the summaries in two languages. As the linguistic representations of the two summaries lie on different vector spaces, we adopt the UMH in [5] to map the linguistic representations into a unified vector space. Then, we view the bidirectional semantic alignment of summaries as an optimal transmission problem. Additionally, we use contrast learning to capture the semantic similarity between two summaries. It helps the student model focus on differentials among languages and extract critical information from the text.

Our main contributions can be described as follows:

- We propose a knowledge distillation framework for the cross-lingual summarization task. The student model constructs strong relationships between two languages by learning the attention weight and word distribution of the teacher model.
- We propose a text similarity metric that maps texts from different languages into a unified similarity space for comparison. Furthermore, we introduce contrastive learning to push similar text vectors toward each other, combined with the text similarity metric to compute the bidirectional semantics of two summaries.
- We validate our model by training on datasets of language pairs with separate syntactic structures and vocabulary morphologies in a low-resource scenario. We find that our approach outperforms other baselines in most cases.

2. Related Work

Early cross-lingual summary generation mainly focused on pipe-lined approaches [6–8], which suffered from error propagation problems. Since [9] first proposed that end-to-end methods are significantly better than pipe-lined methods, such methods are gradually becoming the mainstream methods for generating cross-lingual text summaries [3,10].

Since cross-lingual summarization still lacks large-scale, high-quality supervised datasets [11], methods that can be pre-trained from large-scale unsupervised corpora have been extensively investigated. Applying a pre-trained model to the cross-lingual summarization task involves pre-training and fine-tuning. In the former, the model uses many unlabeled data to learn generic linguistic knowledge; the latter uses the labeled data to fine-tune the cross-lingual summarization task. However, early pre-training models were trained in only one language [12–15]. It is difficult for models to learn general knowledge of another language in a cross-lingual task.

Then, multilanguage pre-trained models started to be pre-trained in multiple languages [1,2,16], which allowed them to learn syntactic, lexical, and semantic similarities between languages. Overall, the application of pre-training methods in cross-lingual summarization has been widely explored and studied [17–19] and achieved good performances in different languages. For instance, ref. [20] proposes an mBERT-based knowledge distillation framework that improves the summarization performance for remote languages by aligning cross-language summaries with monolingual summaries. Ref. [21] propose a many-to-one summarization model using the mT5 model combined with a contrastive learning approach to unify cross-lingual and monolingual representations to enrich low-resource data.

In contrast to the former study, we introduce contrastive learning into the knowledge distillation framework to construct strong correlations between languages. Moreover, we introduce the attention weight and word distribution of the teacher model to form adequate alignment supervision information.

3. Methods

The MPLM will be pre-trained in multiple languages. When generating a cross-lingual summary of two languages, the model may mistake the lingual features in the pre-trained corpus for features of two languages, thus interfering with the quality of the summary. We propose a knowledge distillation framework and a similarity metric based on contrast learning to address the problem that the cross-lingual summarization model suffers from data interference in low-resource settings.

We construct cross-lingual supervision by linking a monolingual teacher model with a cross-lingual student model. Then, we use a bidirectional semantic alignment of the similarity metric to construct strong correlations among languages. Finally, we use contrast learning to enhance the ability of the similarity metric for bidirectional semantic alignment. The overall model framework is shown in Figure 1. In this way, we can improve the performance of the model on languages with separate syntactic structures and vocabulary morphologies.

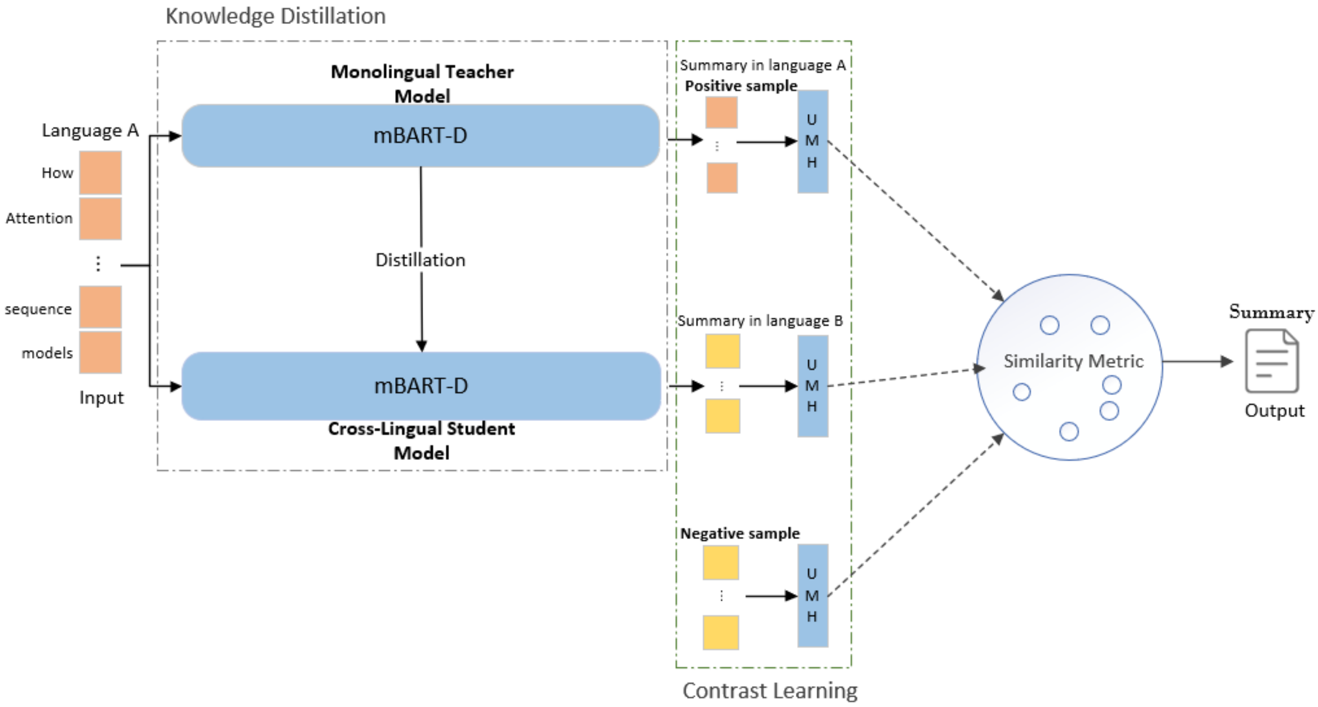

Figure 1. Diagram of knowledge distillation framework for cross-lingual summary.

3.1. Knowledge Distillation

The knowledge distillation framework transfers knowledge from a teacher model trained on a specific corpus to a student model. While preserving the effective encoding of the source language by the multilingual pre-trained model, this framework alleviates interference among different languages in the pre-training corpus during the generation of target language summaries by the decoder. In particular, we use a new model named the mBART-D model to construct teacher and student models. The architecture diagram of the mBART-D model is shown in Figure 2. In addition, the cross-lingual summary model requires understanding both the source and target languages and constructing the relationship between the two languages. To make the student model have the same ability to generate summaries as the teacher model, we transfer the attention weights and word distributions from the teacher model to the student model. The specific distillation is shown in Figure 3.

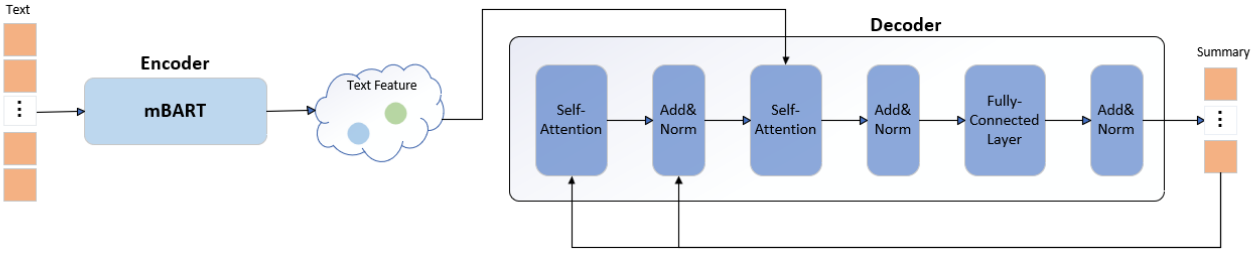

Figure 2. Overview of the proposed mBART-D model. The encoder of the mBART model extracts the text features of the source text and then inputs the text features into an additionally constructed decoder to generate the final summary.

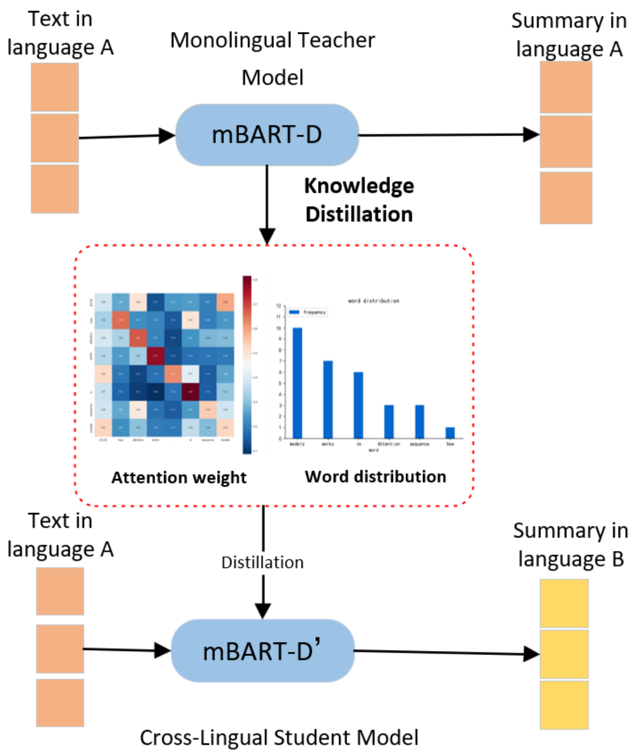

Figure 3. Framework for the refinement process. Both the monolingual teacher and the cross-language student models use mBART-D (for better differentiation, we set the student model to mBART-D'). The aim is to refine the attentional weights and word distributions from the monolingual teacher model into the cross-language student model. The teacher and student models generate monolingual and cross-lingual summaries, respectively.

3.1.1. mBART-D Model

Since the mBART model is a multilingual pre-trained model, many datasets are needed to fine-tune it for cross-language text summary generation tasks. However, when the mBART model is used in low-resource scenarios, the model may mistake linguistic features in the pre-trained corpus as features in both languages during text summary generation. It will affect the quality of the final generated text summaries.

To address the above issues, we construct an mBART-D model consisting of the mBART encoder and the decoder initialized by Xavier [22]. Specifically, the mBART-D model performs text feature extraction using the mBART encoder to obtain contextual and semantic information then employs the Xavier initialization method to initialize the weight values of the decoder, which inputs the text features into the decoder. Compared to generating summaries using the mBART model in low-resource scenarios, the mBART-D model inherits the encoding capabilities of the mBART model. At the same time, the initialized decoder is not affected by other linguistic features. We use the mBART-D model to mitigate the problem that the model may misidentify linguistic features in the pre-trained corpus as two linguistic features when outputting summaries.

3.1.2. Teacher Model

We utilize the mBART-D model as the teacher model to generate the monolingual summary. In detail, given the text input $T^A = \{X_1^A, X_2^A, X_3^A, \ldots, X_n^A\}$ in language A, the teacher model uses the maximum likelihood function with cross-entropy loss to generate the monolingual summary $S^A = \{Y_1^A, Y_2^A, Y_3^A, \ldots, Y_m^A\}$, where the m and n are the lengths of the input text and the monolingual summary and $m < n$. The formula is as follows:

$$S_{mls} = -\sum_{j=1}^{m} logP(Y_j^A | Y_{<j}^A, X) \tag{1}$$

We use a monolingual summary dataset to train the mBART-D model as a teacher model. Adequate training samples allow the mBART-D model to capture source language semantic information and generate accurate summaries efficiently.

3.1.3. Student Model

We train the student model using the mBART-D model and a cross-lingual summary corpus. In detail, given the text input $T^A = \{X_1^A, X_2^A, X_3^A, \ldots, X_n^A\}$, also using the maximum likelihood function with cross-entropy loss, it finally generates cross-linguistic text summaries $S^B = \{Y_1^B, Y_2^B, Y_3^B, \ldots, Y_l^B\}$, where the m and l are the lengths of the input text and the cross-lingual summary and $l < n$. The task form is as follows:

$$S_{cls} = -\sum_{j=1}^{l} logP(Y_j^B | Y_{<j}^B, X) \tag{2}$$

After fine-tuning the teacher model, we trained the student model on a cross-lingual summary dataset. Such training enabled the student model to inherit the same encoding capabilities as the teacher model and the ability to decode in the target language.

3.1.4. Attention Weight

The summary model generates summaries by automatically assigning weights to each word in each input sentence. These weights are the importance of each word in the summary generation process. In the teacher model, the encoder processes the source language input X^A, while the decoder receives the source language summary Y^A, thereby facilitating the provision of attention alignment information from X^A to Y^A. In the student model, the encoder also takes the source language X^A as input, allowing the teacher model to furnish direct and effective supervisory information. Ref. [10] defined the attention weight distribution matrix in a cross-lingual sentence summarization model as A_j. During training, it is desired that the attention weight B_j of the student model approximates the attention weight A_j of the teacher model. Therefore, we use the attention relay approach and Euclidean distance to encourage the consistency of the attention weights of the student model and the teacher models' attention weights. The formula is as follows:

$$L_{att} = -\sqrt{\sum_j (A_j - B_j)^2} \tag{3}$$

where j represents the location of the attention weight. Through Equation (3), the student model can learn the attention weights from the teacher model to better understand the semantics of the input sentences.

Note that the teacher and student models use a Transformer architecture that contains multiple attention heads but omits self-attention in the Transformer. Therefore, we use the average attention method, which averages all attention heads' weights in the same layer.

3.1.5. Word Distribution

The word distribution reflects the ability of the teacher model to understand and summarize the input text. To better understand the associations and similarities between different words, the student model learns the word distribution of the teacher model. This can improve the accuracy and coherence of the cross-lingual summaries. Therefore, we use the cross-entropy loss to encourage the similarity of the probability distributions of the summary words generated by the two models. The formula is as follows:

$$L_{word} = -P(Y_i^A|Y_{<i-1}^A, X) log P(Y_i^B|Y_{<i-1}^B, X) \tag{4}$$

where $P(Y_i^A|Y_{<i-1}^A, X)$ denotes the summary word distribution of the teacher model and $P(Y_i^B|Y_{<i-1}^B, X)$ denotes the summary word distribution of the student model.

3.2. Similarity Metric

To understand the relationships between texts that are context-rich and cross-linguistic, we propose a similarity metric based on bidirectional semantic alignment inspired by [23]. The approach uses bidirectional semantic alignment, allowing the model to capture more contextual information at different locations in the text and making the metric more focused on the overall context of the text. The process is shown in Figure 4.

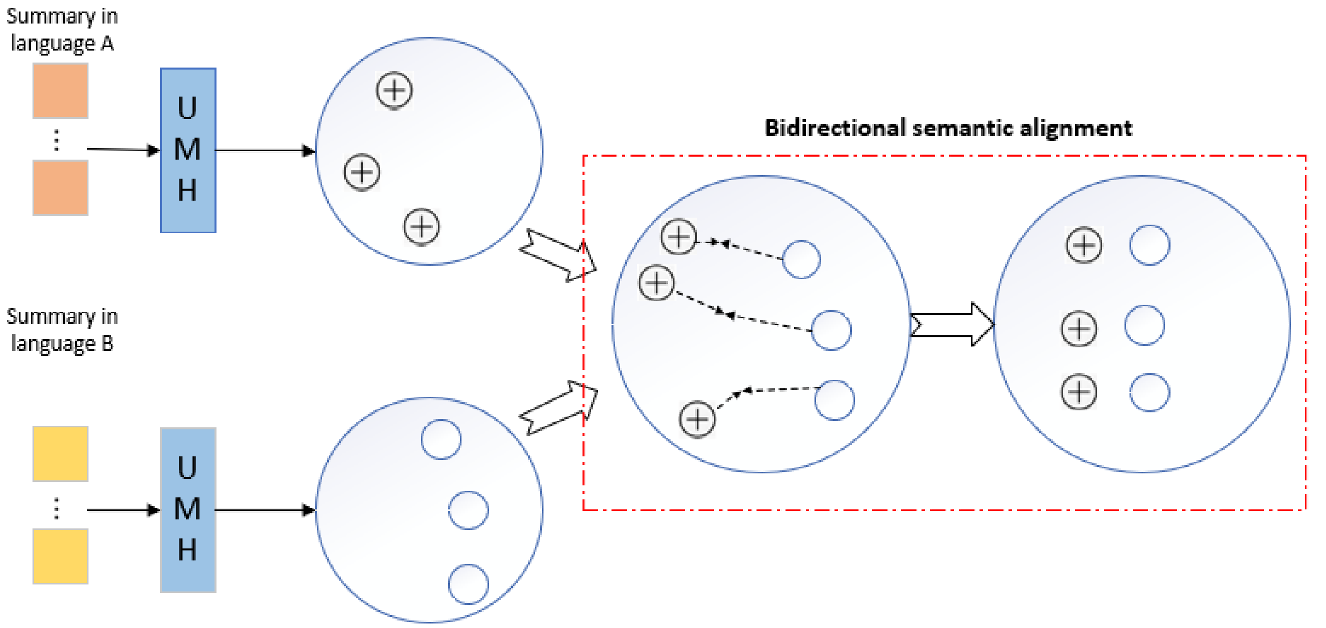

Figure 4. The similarity metric is based on bidirectional semantic alignment. The summaries in language A and language B are mapped to the same vector space by the UMH. Then, the distance between the vectors is measured using the similarity metric through bidirectional semantic alignment.

The teacher and student models generate summaries in two languages. Their linguistic representations lie on different vector spaces. So, the similarity of the two summaries cannot be calculated directly. We use UMH, an unsupervised hyper-alignment for multilingual word embeddings, to map the word vectors of summaries in different languages to a unified vector space. The formula is as follows:

$$\min_{Q_i, P_{i,j}} \sum_{i,j} \alpha_{i,j} l(X_i Q_i, P_{ij} X_j Q_j) \tag{5}$$

where α is the weighting coefficient and X_i, X_j represent different languages. P and Q denote the allocation and mapping matrices, respectively.

Then, we apply optimal transport in a contextual embedding space. It uses the optimal solution of a relaxed transport problem as a distance metric and converts the distance metric to the corresponding similarity measure with the following equation:

$$Sim_{s1}(s^1, s^2) = \frac{1}{L_1}\sum_{i=1}^{L_1} \max_j cos(x_i^1, x_j^2) \tag{6}$$

$$Sim_{s2}(s^2, s^1) = \frac{1}{L_2}\sum_{j=1}^{L_2} \max_i cos(x_i^2, x_j^1) \tag{7}$$

where *s1* and *s2* represent the summaries in two different languages and L_1 and L_2 represent the lengths of the summaries in the two languages. x_i^1 represents the sentence token of the summary in language 1, and x_j^2 represents the sentence token of the summary in language 2, where i and j represent the positions of the tokens. x_j^1 and x_i^2 mean the same as x_i^1 and x_j^2.

By comparing the word vectors in the two sentences one by one and choosing the maximum cosine similarity value, we can obtain two scores that measure the similarity of the sentences. These scores were averaged to obtain a combined bidirectional semantic alignment metric between the two summary sentences. The formula is as follows:

$$Sim(s^1, s^2) = \frac{1}{2}(Sim_{s1}(s^1, s^2) + Sim_{s2}(s^2, s^1)) \tag{8}$$

3.3. Contrastive Learning

We use contrastive learning to enhance the ability of the similarity metric for bidirectional semantic alignment. We consider the monolingual summaries generated by the teacher model as positive samples. Moreover, following the idea of [24], we use the other sentences in the batch as negative pairs. We then calculate the similarity metrics of the cross-lingual summaries to the positive and negative samples separately so that the cross-lingual summaries are close to the predicted values for the positive samples and far from a certain threshold for the negative samples. The process is shown in Figure 5. This enables the final generated cross-language text summaries to express the same content as the positive samples. So, the contrast learning for sentence i in the batch is defined as follows.

For the above distillation process, we define our distillation loss as follows:

$$L_{KD} = \lambda_1 L_{att} + \lambda_2 L_{word} + \lambda_3 L_{sim} \tag{9}$$

where L_{att} comes from Equation (3), L_{word} comes from Equation (4), and L_{sim} comes from Equation (8). λ_1, λ_2, and λ_3 are the weighted hyper-parameters.

$$L_{sim} = -log\frac{exp(\frac{Sim(s^i, S_+^i)}{\tau})}{\sum_{j=1}^{B}(exp(\frac{Sim(s^i, S_+^j)}{\tau}) + exp(\frac{Sim(s^i, S_-^j)}{\tau}))} \tag{10}$$

where τ is the temperature parameter, B is the batch size, and s_+ and s_- represent positive and negative samples, respectively.

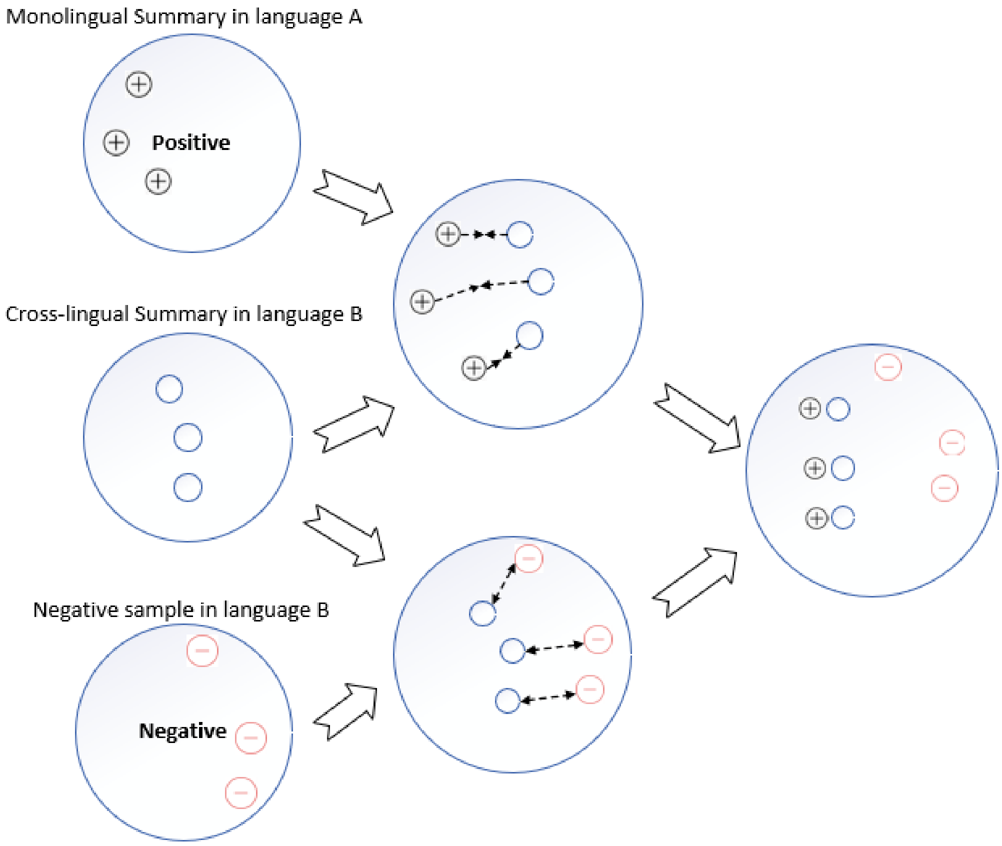

Figure 5. Contrast learning. We used the monolingual summary in language A as a positive sample. Following the idea of [24], we used other sentences in language B in the batch as negative sentence pairs. Then, the positive and negative samples and the cross-lingual summary are aligned using the similarity measure.

3.4. Training Objective

We propose a cross-lingual summary model based on knowledge distillation. Its overall model framework is shown in Figure 4. Our total training loss is computed as follows for each input:

$$L = S_{cls} + \delta L_{KD} \tag{11}$$

where S_{cls} comes from Equation (2) and L_{KD} comes from Equation (10). δ is a weighted hyper-parameter that balances the weights between S_{cls} and L_{KD}.

4. Experiments

4.1. Datasets and Evaluation

To validate the effectiveness of our method on the languages with separate syntactic structures and vocabulary morphologies in low-resource settings, we processed the Wikilingua [25] dataset. Inspired by [20], we use a back-translation pre-processing strategy to convert each sample into a document consisting of three parts: a document, a monolingual summary, and a cross-language summary.

For the evaluation, we specifically chose four linguistic variants that exhibited variations in either structure or morphology. These variants include En2ArSum (summary from English to Arabic), En2ViSum (summary from English to Vietnamese), and En2JaSum (summary from English to Japanese). Moreover, to validate the performance of the model in cross-language summarization tasks where the source language is not English, we also introduce Ja2EnSum (summary from Japanese to English) with the same content as En2JaSum. The selected dataset encompasses diverse syntactic structures and character sets across several languages. The En2ViSum dataset presents a low-resource scenario with a mere 7500 samples. The average lengths of the source texts, monolingual summaries, and cross-lingual summaries for each dataset and the size of each dataset are shown in Table 1.

Table 1. Statistics for the Wikilingua variant datasets.

Dataset	$\mathbf{Len}_{Source}$	$\mathbf{Len}_{MLS}$	$\mathbf{Len}_{CLS}$	Size
En2ArSum	1589	227	133	12,500
En2JaSum	1463	212	133	10,000
Ja2EnSum	2103	133	212	10,000
En2ViSum	1657	175	135	7500

After [26], most studies have used the standard ROUGE method to assess the performance of models. Therefore, we used the ROUGE method to evaluate the ROUGE-1, ROUGE-2, and ROUGE-L scores of our model on the Wikilingua variant dataset.

4.2. Implementation Details

We use mBART model to initialize our encoder and Xavier to initialize the decoder. The encoder and decoder have dimensions of 1024, and the model contains 1,277,437,076 parameters. The encoder and decoder use separate Adam optimizers, with learning rates of 0.002 and 0.2, respectively. The model is trained on a Tesla V100S-PCIE-32 GB with a training phase of 15,000 steps and gradient accumulation every five steps. To estimate the text similarity, we set the temperature parameter τ to 0.05 and the batch size to 128. The weighted hyper-parameters in the loss function are set to 1. In addition, the teacher model in the knowledge distillation framework has the same structure and parameters as the student model. In the training phase, the monolingual summary teacher model is first trained. Then, the cross-lingual summary student model learns the knowledge of the teacher model. In the inference phase, the student model is fed with textual content and undergoes three hours of inference to generate a summary.

4.3. Baseline

For performance comparison, we select several baselines that have better performances in the current study, comparing our proposed model architecture with the following baselines:

- mBART(mbart.cc25): Here, we fine-tune the mBART [1] model for a cross-lingual specific text summarization task, and we apply grid search to update hyper-parameters, such as the learning rate and batch size in the model, as a way to obtain the best performance of mBART in the downstream task.
- mBART + $MADP_D$: Inspired by [27], we use parallel data and connect each parallel instance of the two languages. Then, we combine these data with masking and noise reduction targets to train mBART further.
- MCLAS: The model in [28] is a strong baseline model that uses a unified decoder to generate sequential links between monolingual and cross-lingual summaries, making the monolingual summary task a prerequisite for the cross-lingual summarization task.
- Ref. [20] proposed a knowledge distillation framework based on the mBERT model. Furthermore, Sinkhorn scattering is utilized to bring the representation distance of summaries in two languages closer to generate high-quality cross-language text summaries.

4.4. Contrast Experiment

The experiments are validated on the En2ArSum, En2ViSum, En2JaSum, and Ja2EnSum datasets in the low-resource scenario, and the experimental results correspond to Tables 2, 3, 4 and 5, respectively.

Table 2. The ROUGE score on the En2ArSum dataset for cross-lingual summarization.

Model	ROUGE-1	ROUGE-2	ROUGE-L
mBART	30.13	13.58	22.47
mBART + $MADP_D$	30.58	13.97	22.63
MCLAS	36.28	17.27	27.56
[20]	36.89	20.28	32.38
Our model	35.89	16.20	25.21

Table 3. The ROUGE score on the En2ViSum dataset for cross-lingual summarization.

Model	ROUGE-1	ROUGE-2	ROUGE-L
mBART	31.46	9.69	18.32
mBART + $MADP_D$	31.81	10.05	19.09
MCLAS	36.31	15.91	28.62
[20]	37.38	16.20	28.97
Our model	37.79	15.27	25.83

Table 4. The ROUGE score on the En2JaSum dataset for cross-lingual summarization.

Model	ROUGE-1	ROUGE-2	ROUGE-3
mBART	26.17	10.01	17.94
mBART + $MADP_D$	25.41	10.27	18.26
MCLAS	29.60	16.08	23.20
[20]	30.21	16.27	23.90
Our model	30.12	14.23	21.52

Table 5. The ROUGE score on the Ja2EnSum dataset for cross-lingual summarization.

Model	ROUGE-1	ROUGE-2	ROUGE-3
mBART	27.93	8.19	18.09
mBART + $MADP_D$	28.22	9.57	19.04
MCLAS	33.20	12.57	27.27
[20]	34.21	13.08	27.63
Our model	31.74	14.01	24.47

In Table 2 on the En2ArSum dataset, our model outperforms the mBART model by 5.76 points for ROUGE-1, 2.62 points for ROUGE-2, and 2.74 points for ROUGE-L. Similarly, in Table 3, on the En2ViSum dataset, which has slightly similar syntactic structures and semantic information, the ROUGE-1, ROUGE-2, and ROUGE-L scores of our model improve by 6.33, 5.58, and 7.51 points compared to the mBART model. In addition, we use two reverse datasets, the En2JaSum dataset and the Ja2EnSum dataset, to validate our model. For the En2JaSum dataset in Table 4, the ROUGE-1, ROUGE-2, and ROUGE-L scores of our model improve by 3.95, 4.22, and 3.58 points compared to the mBART model. For the reverse dataset Ja2EnSum in Table 5, our model outperforms the mBART model by 3.81 points for ROUGE-1, 5.82 points for ROUGE-2, and 6.38 points for ROUGE-L. There is a significant improvement in the performance of our model compared to the mBART model. This shows that our model can build a strong correlation between two languages by adopting the knowledge distillation framework, and it has improved in semantic and context understanding.

Compared to the strong baseline MCLAS model, in Table 2, on the En2ArSum dataset, we see that our model is just 0.39, 1.07, and 2.35 points lower on ROUGE-1, ROUGE-2, and ROUGE-L. Similarly, in Table 3, on the En2ViSum dataset, the ROUGE-1 scores of our model improve by 1.438 points. For the En2JaSum dataset in Table 4, the ROUGE-1 scores of our model improve by 0.52 points. It shows that our model can capture and retain critical information in the generated summaries. For the reverse dataset Ja2EnSum in Table 5, our model outperforms the MCLAS model 1.44 points for ROUGE-2. Compared to [20], the partial ROUGE values of our model on the En2ViSum dataset and the Ja2EnSum dataset increased. In particular, the ROUGE-1 score on the low-resource dataset En2ViSum improved by 0.41. In contrast, the partial ROUGE values on the other datasets decreased only slightly. This demonstrates that our model can correctly capture the contextual relationships between words in a summary, generating a more coherent cross-linguistic summary. We can see that our model performance has some improvement on some datasets. It shows that our model is competitive.

In conclusion, our approach can improve the quality of cross-lingual summaries on language pairs with separate syntactic structures and vocabulary morphologies in low-resource scenarios.

4.5. Ablation Experiment

We conduct the following ablation experiments on our model to investigate the importance of different model components for improving the quality of cross-lingual summaries.

4.5.1. Attention Weight and Word Distribution

We adopt a knowledge distillation to distill the attention weight and word distribution of the monolingual teacher model into the cross-lingual student model. To verify the rationality of distilled content, we set the following variables on the En2ArSum dataset: (1) no distilled content (non-KD), (2) no distilled attention weight (our model—attention weight), and (3) no distillation word distribution (our model—word distribution), (4) simultaneous distillation attention weight and word distribution (our model). The ablation experiments results of knowledge distillation are shown in Table 6.

Table 6. Results of ablation tests of attention weight and word distribution.

Method	ROUGE-1	ROUGE-2	ROUGE-3
Non-KD	31.41	14.27	22.98
Our model—attention weight	32.59	14.31	23.75
Our model—word distribution	32.74	14.28	23.92
Our model	35.89	16.20	25.21

The analysis of the experimental results in Table 6 shows that distilling only the attention weights or word distribution improves the performance of the model compared to no distilled content. This is because distilling the attention weights of the teacher model can help the student model better understand the critical information in the original text, thus improving the accuracy of the generated summaries. The distillation of word distribution can better guide the student model to learn the content and structure of the text, thus improving the readability of the generated summaries. Our model distills both attention weights and word distributions, resulting in the cross-lingual model having significant performance improvements and the cross-lingual summaries becoming more accurate and fluent.

4.5.2. Similarity Metric

We use a similarity metric via optimal transport-based contrastive learning, constructing a bidirectional semantic alignment between two different language summaries. To verify the rationality of the similarity metric, we set the following variables on the En2ArSum dataset: (1) no similarity metric (non-Sim), (2) average pooling (average pooling), (3) bidirectional semantic alignment of similarity metric (Bid-Sim), (4) Bid-Sim with contrast learning (our model). The experimental results are shown in Table 7.

Table 7. Results of ablation tests in text similarity.

Method	ROUGE-1	ROUGE-2	ROUGE-3
Non-Sim	32.31	14.79	23.74
Average pooling	32.43	14.80	23.78
Bid-Sim	33.16	14.92	24.05
Our model	35.89	16.20	25.21

The analysis of the experimental results in Table 7 shows that the quality of cross-lingual summaries can be improved using only the average pooling. Compared with the average pooling, Bid-Sim captures the bidirectional semantic similarity between two texts and mitigates the influence of language differences, improving the ROUGE score. As seen from lines 3 and 4, our model further improves the quality of the generated cross-lingual summaries. We think that because contrast learning helps the model to focus on the differences between languages it can better capture text bidirectional semantics.

4.5.3. Vector Space Mapping

Since the teacher and student models generate summaries in different languages, the representation of two summaries lies on different vector spaces. To verify the rationality of the vector space mapping method, we set the following variables on the En2ArSum dataset: (1) GeoMM [29], (2) Gromov–Wasserstein (GW) [30],(3) ICP [31],(4) UMH (our model). The experimental results are shown in Table 8.

Table 8. Results of ablation tests in vector space mapping.

Method	ROUGE-1	ROUGE-2	ROUGE-3
GeoMM	34.72	15.48	23.07
GW	34.91	15.75	23.66
ICP	35.14	16.03	24.54
Our model	35.89	16.20	25.21

The analysis of the experimental results in Table 8 shows that the UMH method maps vectors from different spaces to the same vector space better than other space vector mapping methods. This is because the UMH method can maximize the quality of word translation between languages when mapping each language into a common space.

4.6. Case Study

We take the example of a cross-lingual summary from English to Arabic. We compare our model with three models. The results of the summaries generated by each model are shown in Figure 6.

In Table 2, on the En2ArSum dataset, we choose two samples of high-quality summaries generated by the baseline model. As can be seen, except for the MCLAS model and [20], the other baseline models are missing keywords, e.g., “cooperation” in sample 1 and “traffic control” and “temporary” in sample 2. Our model and MCLAS can extract most of the keywords. However, in sample 1, MCLAS generates the wrong person’s name. Our model and [20] can extract keywords from text and generate text summaries with complete semantics. Therefore, the quality of the summaries generated by our model is competitive, even sometimes better than MCLAS regarding summary readability.

Sample 1

Source: Former US Secretary of State Henry Kissinger and Clinton Secretary of State Madeleine Albright stressed the importance of the US-China relationship in New York City on Monday night, expressing optimism about the two countries' cooperation with the US in the direction of the transition to the Trump administration.

(**Translation:**前美国国务卿亨利 • 基辛格和克林顿国务卿马德琳 • 奥尔布赖特周一晚上在纽约市强调了美中关系的重要性，对两国与美国向特朗普政府过渡方向的合作表示乐观。)

mBART :وزيرا خارجية أمريكيان يؤكدان على أهمية العلاقات مع الصين.(**Translation**: Two US Secretary of State emphasized the importance of relations with China.)

mBART+MADP$_D$:اثنين من وزراء الخارجية السابقين : أهمية أمريكا الوسطى والتفاؤل بشأن ولاية ترامب.(**Translation:**Two former secretaries of state: The importance of Central America and optimism about the Trump presidency.)

MCLAS:كيسنجر أولبرايت التأكيد على أهمية العلاقات بين الولايات المتحدة والصين ، متفائل بشأن التعاون خلال الفترة الانتقالية من حكومة ترامب .(**Translation:**Kissinger Albright emphasized the importance of China US relations and expressed optimism about cooperation during the Trump administration's transitional period.)

Nguyen, T et al.(2021):كيسنجر أولبرايت شدد على أهمية العلاقات بين الولايات المتحدة والصين ، وأعرب عن تفاؤله بشأن التعاون مع حكومة ترامب خلال الفترة الانتقالية .(**Translation:**Kissinger and Albright emphasized the importance of China US relations and expressed optimism about cooperation with the Trump administration during the transitional period.)

Our Model:**كيسنجر أولبرايت** على أهمية العلاقات بين الولايات المتحدة والصين ، متفائل بشأن **التعاون** مع حكومة ترامب .(**Translation:**Kissinger and Albright emphasized the importance of China US relations and expressed optimism about cooperation with the Trump administration.)

Monolingual summary: Kissinger and Albright emphasized the importance of US-China relations and expressed optimism for cooperation in the transition to the Trump administration.

Reference:كيسنجر أولبرايت على أهمية العلاقات بين الولايات المتحدة والصين ، وأعرب عن تفاؤله بشأن التعاون خلال الفترة الانتقالية إلى حكومة ترامب .

Sample 2

Source: On Monday morning, a group of nearly 50 Tibetan antelopes was seen gathering by the side of the road at a section of the Qinghai-Tibet Highway. As a precautionary measure, the staff of the nature reserve implemented temporary traffic control to ensure safety.

(**Translation:**周一早上，在青藏公路的一个路段，看到一群近 50 只藏羚羊聚集在路边。作为预防措施，自然保护区的工作人员实施了临时交通管制，以确保安全。)

mBART:الموظفين في التبت الظباء الأغنام بالقرب من الطريق السريع لتنفيذ الرقابة .(**Translation**: Xizang staff are monitoring near the expressway.)

mBART+MADP$_D$:الظباء التبتية تجمعوا على طول الطريق السريع تشينغهاى التبت ، مما دفع الموظفين لمراقبة حركة المرور .(**Translation:**Xizang antelopes gathered along the Qinghai Tibet Expressway, prompting staff to monitor traffic.)

MCLAS:الظباء التبتية يتجمعون في تشينغهاى التبت السريع ، محمية طبيعية مؤقتة لمراقبة حركة المرور لحماية الموظفين(**Translation:**Tibetan antelopes gather in the fast-paced Shanghai of Qinghai Tibet, a temporary traffic controlled nature reserve to protect employees)

Nguyen, T et al.(2021):التبت الظباء على طول الطريق السريع بين تشينغهاى والتبت ، حيث الموظفين في المناطق المحمية المؤقتة لمراقبة حركة المرور .(**Translation:**Tibetan antelopes along the Qinghai Tibet Highway are being monitored for traffic by staff in temporary protected areas.)

Our Model:الظباء التبتيين تجمعوا على طول الطريق السريع ، والموظفين في المناطق المحمية المؤقتة لمراقبة حركة المرور السريع.(**Translation:**Tibetan antelope gather along the highway, and reserve staff conducts temporary traffic control.)

Monolingual summary: Tibetan antelopes gathered by the Qinghai-Tibet Highway, prompting temporary traffic control from nature reserve staff.

Reference:الظباء التبتية تجمعوا على طول الطريق السريع تشينغهاى التبت ، مما دفع موظفي المحميات الطبيعية المؤقتة لمراقبة حركة المرور .

Figure 6. Case studies of English-to-Arabic summarization (Nguyen, T et al. (2021) from the [20]).

5. Conclusions

This paper proposes a knowledge distillation model to construct strong correlations between two languages with separate syntactic structures and vocabulary morphologies in low-resource settings. We construct a monolingual teacher model and a cross-lingual student model, respectively. The attention weight and word distribution from the teacher model are transferred to the student model. Furthermore, we propose a similarity metric based on contrast learning to estimate cross-lingual differences effectively. Extensive experiments show that our approach has an improved accuracy and performance in low-resource settings for languages with separate syntactic structures and vocabulary morphologies.

Our research addresses the challenges of cross-linguistic associations, especially by enhancing the collaboration of editors in editing overseas reports. By providing relevant experimental results demonstrating the effectiveness of our model in low-resource environments, we aim to emphasize the great strides our approach has made in facilitating effective multilingual collaboration. However, one potential weakness of our work is that it may not perform as well in languages with more complex syntaxes and vocabularies. We plan to attempt to address this issue in future research endeavors.

Author Contributions: Conceptualization, X.Y. and J.Y.; methodology, X.Y.; software, X.Y.; validation, X.Y.; formal analysis, X.Y., J.Y. and B.Z.; investigation, X.Y. and B.Z.; resources, J.Y. and L.L.; data curation, X.Y. and Q.B.; writing—original draft preparation, X.Y.; writing—review and editing, X.Y.,

B.Z. and J.Y.; visualization, X.Y.; supervision, J.Y. and L.L.; project administration, J.Y. All authors have read and agreed to the published version of this manuscript.

Funding: The work was supported by the National Natural Science Foundation of China (62062055); Basic Scientific Research Expenses Program of Universities directly under Inner Mongolia Autonomous Region (JY20220249); and Teaching Reform Project for Postgraduate Education under the Inner Mongolia University of Technology (YJG2021016).

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Liu, Y.; Gu, J.; Goyal, N.; Li, X.; Edunov, S.; Ghazvininejad, M.; Zettlemoyer, L. Multilingual denoising pre-training for neural machine translation. *Trans. Assoc. Comput. Linguist.* **2020**, *8*, 726–742. Available online: https://arxiv.org/pdf/2001.08210v2.pdf (accessed on 13 August 2022). [CrossRef]
2. Xue, L.; Constant, N.; Roberts, A.; Kale, M.; Al-Rfou, R.; Siddhant, A.; Raffel, C. mT5: A massively multilingual pre-trained text-to-text transformer. *arXiv* **2020**, arXiv:2010.11934.
3. Liang, Y.; Meng, F.; Zhou, C.; Xu, J.; Chen, Y.; Su, J.; Zhou, J. A variational hierarchical model for neural cross-lingual summarization. *arXiv* **2022**, arXiv:2203.03820.
4. Bjerva, J.; Östling, R.; Veiga, M.H.; Tiedemann, J.; Augenstein, I. What do language representations really represent? *Comput. Linguist.* **2019**, *45*, 381–389. [CrossRef]
5. Alaux, J.; Grave, E.; Cuturi, M.; Joulin, A. Unsupervised hyperalignment for multilingual word embeddings. *arXiv* **2018**, arXiv:1811.01124.
6. Wan, X.; Li, H.; Xiao, J. Cross-language document summarization based on machine translation quality prediction. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, 11–16 July 2010.
7. Zhang, J.; Zhou, Y.; Zong, C. Abstractive cross-language summarization via translation model enhanced predicate argument structure fusing. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2016**, *24*, 1842–1853. [CrossRef]
8. Linhares Pontes, E.; Huet, S.; Torres-Moreno, J.M.; Linhares, A.C. Cross-language text summarization using sentence and multi-sentence compression. In Proceedings of the Natural Language Processing and Information Systems: 23rd International Conference on Applications of Natural Language to Information Systems, NLDB 2018, Paris, France, 13–15 June 2018.
9. Zhu, J.; Wang, Q.; Wang, Y.; Zhou, Y.; Zhang, J.; Wang, S.; Zong, C. NCLS: Neural cross-lingual summarization. *arXiv* **2019**, arXiv:1909.00156.
10. Duan, X.; Yin, M.; Zhang, M.; Chen, B.; Luo, W. 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, Florence, Italy, 28 July–2 August 2019.
11. Cheng, J.; Lapata, M. Neural summarization by extracting sentences and words. *arXiv* **2016**, arXiv:1603.07252.
12. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving Language Understanding by Generative Pre-Training. 2018. Available online: https://www.mikecaptain.com/resources/pdf/GPT-1.pdf (accessed on 7 March 2023).
13. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
14. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692.
15. Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Zettlemoyer, L. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv* **2019**, arXiv:1910.13461.
16. Lample, G.; Conneau, A. Cross-lingual language model pretraining. *arXiv* **2019**, arXiv:1901.07291.
17. Huot, F.; Maynez, J.; Alberti, C.; Amplayo, R.K.; Agrawal, P.; Fierro, C.; Lapata, M. μ PLAN: Summarizing using a Content Plan as Cross-Lingual Bridge. *arXiv* **2023**, arXiv:2305.14205.
18. Wang, J.; Meng, F.; Zheng, D.; Liang, Y.; Li, Z.; Qu, J.; Zhou, J. Towards Unifying Multi-Lingual and Cross-Lingual Summarization. *arXiv* **2023**, arXiv:2305.09220.
19. Wang, J.; Meng, F.; Lu, Z.; Zheng, D.; Li, Z.; Qu, J.; Zhou, J. Clidsum: A benchmark dataset for cross-lingual dialogue summarization. *arXiv* **2022**, arXiv:2202.05599.
20. Nguyen, T.; Tuan, L.A. Improving neural cross-lingual summarization via employing optimal transport distance for knowledge distillation. *arXiv* **2021**, arXiv:2112.03473.
21. Li, P.; Zhang, Z.; Wang, J.; Li, L.; Jatowt, A.; Yang, Z. ACROSS: An Alignment-based Framework for Low-Resource Many-to-One Cross-Lingual Summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*; Association for Computational Linguistics: Cedarville, OH, USA, 2023; pp. 2458–2472.
22. Glorot, X.; Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, JMLR Workshop and Conference Proceedings, Sardinia, Italy, 13–15 May 2010; pp. 249–256.

23. Lee, S.; Lee, D.; Jang, S.; Yu, H. Toward interpretable semantic textual similarity via optimal transport-based contrastive sentence learning. *arXiv* **2022**, arXiv:2202.13196.
24. Gao, T.; Yao, X.; Chen, D. Simcse: Simple contrastive learning of sentence embeddings. *arXiv* **2021**, arXiv:2104.08821.
25. Ladhak, F.; Durmus, E.; Cardie, C.; McKeown, K. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. *arXiv* **2020**, arXiv:2010.03093.
26. Zhu, J.; Zhou, Y.; Zhang, J.; Zong, C. 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, Online, 5–10 July 2020.
27. Chi, Z.; Dong, L.; Wei, F.; Wang, W.; Mao, X.L.; Huang, H. Cross-lingual natural language generation via pre-training. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020.
28. Bai, Y.; Gao, Y.; Huang, H. Cross-lingual abstractive summarization with limited parallel resources. *arXiv* **2021**, arXiv:2105.13648.
29. Jawanpuria, P.; Balgovind, A.; Kunchukuttan, A.; Mishra, B. Learning multilingual word embeddings in latent metric space: A geometric approach. *Trans. Assoc. Comput. Linguist.* **2019**, 7, 107–120. [CrossRef]
30. Alvarez-Melis, D.; Jaakkola, T.S. Gromov-Wasserstein alignment of word embedding spaces. *arXiv* **2018**, arXiv:1809.00013.
31. Hoshen, Y.; Wolf, L. An iterative closest point method for unsupervised word translation. *arXiv* **2018**, arXiv:1801.06126.