


Article

A Multitask Cross-Lingual Summary Method Based on ABO Mechanism

Qing Li ¹, Weibing Wan ^{1,*}  and Yuming Zhao ²

¹ School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai 201620, China; m025120313@sues.edu.cn

² Department of Automation, Shanghai Jiao Tong University, Shanghai 200240, China; arola_zym@sjtu.edu.cn

* Correspondence: wbwan@sues.edu.cn

Abstract: Recent cross-lingual summarization research has pursued the use of a unified end-to-end model which has demonstrated a certain level of improvement in performance and effectiveness, but this approach stitches together multiple tasks and makes the computation more complex. Less work has focused on alignment relationships across languages, which has led to persistent problems of summary misordering and loss of key information. For this reason, we first simplify the multitasking by converting the translation task into an equal proportion of cross-lingual summary tasks so that the model can perform only cross-lingual summary tasks when generating cross-lingual summaries. In addition, we splice monolingual and cross-lingual summary sequences as an input so that the model can fully learn the core content of the corpus. Then, we propose a reinforced regularization method based on the model to improve its robustness, and build a targeted ABO mechanism to enhance the semantic relationship alignment and key information retention of the cross-lingual summaries. Ablation experiments are conducted on three datasets of different orders of magnitude to demonstrate the effective enhancement of the model by the optimization approach; they outperform the mainstream approaches on the cross-lingual summarization task and the monolingual summarization task for the full dataset. Finally, we validate the model's capabilities on a cross-lingual summary dataset of professional domains, and the results demonstrate its superior performance and ability to improve cross-lingual sequencing.

Keywords: cross-lingual summary; pre-trained language model; abstractive summarization



Citation: Li, Q.; Wan, W.; Zhao, Y. A Multitask Cross-Lingual Summary Method Based on ABO Mechanism. *Appl. Sci.* **2023**, *13*, 6723. <https://doi.org/10.3390/app13116723>

Academic Editor: Vincent A. Cicirello

Received: 20 April 2023

Revised: 24 May 2023

Accepted: 29 May 2023

Published: 31 May 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Cross-lingual summarization (CLS) involves extracting the core content of a document in one language and expressing it in another language [1]. This task requires addressing both redundant information and language differences, which usually involves multiple steps. However, decomposing cross-lingual text summarization into multiple tasks does not guarantee optimal results for each individual neural network. In addition, the distinct linguistic and syntactic structures of different languages need to be considered. Directly applying an end-to-end model across languages can result in the loss of important linguistic nuances and syntactic patterns, potentially leading to the omission of key information or the incorrect placement of sentences within the summary. Therefore, we must address two problems: how to integrate multiple tasks into an effective workflow [2], and how to capture the connections between different languages to generate summaries with reasonable content.

To improve the quality of summaries, recent studies have focused on reducing variability between tasks [3,4]. Transformer-based neural networks [5] have demonstrated the ability to share feature representations of languages between hidden layers in a language-independent way. Based on this idea, this paper abstracts the translation task into a cross-lingual summarization task with the same input sequence to output sequence

length ratio, and integrates it with the monolingual summarization (MS) task in the same model. The overall objective of the model is simplified to “monolingual summarization task + cross-lingual summarization task”. The optimized cross-lingual pre-training model [6] is used as the basis, and the corpus with monolingual and cross-lingual summaries is used as input to allow all information to be shared between tasks. Integrating sequences helps the model gain a comprehensive understanding of the core content in the source language and facilitates better alignment of information across different languages. Furthermore, given the presence of unique nuances and expressions in each language, direct translation is often insufficient to capture these subtleties. By incorporating both monolingual and cross-lingual summary sequences, the model can effectively capture and comprehend these language-specific nuances. This becomes particularly crucial in the context of cross-language summarization, as it enables the model to capture the essence of the source document in its original language and generate a summary in the target language with efficiency.

While end-to-end approaches are known for their improved generalization performance in handling noisy text inputs and reducing issues such as error accumulation [7], those based on pre-trained models often lack flexibility and fail to fully exploit the benefits of pre-training. Consequently, they frequently encounter overfitting problems during downstream task training. To tackle these challenges, we propose novel reinforced regularization methods. By introducing randomness during training, the reinforced regularization method effectively provides regularization, while the inclusion of sparse softmax encourages sparsity, preventing the model from excessively relying on specific features or classes. This combination further enhances the model’s robustness.

Given that the model integrates both monolingual and multilingual summarization tasks, it is necessary to configure the corresponding generation methods to enhance the quality of the generated summaries. However, many previous summarization approaches, such as top-k [8] or pointer networks [9], primarily focus on providing raw output content without ensuring the coherence of the generated statements. Unfortunately, these methods are insufficient for capturing semantic relationships between languages in cross-language summarization tasks. To address this limitation, we have devised the ABO mechanism. This mechanism leverages flexible keyword tags to preserve important semantic information in the abstract and maintain continuity between words. Additionally, we have incorporated a filtering mechanism during the generation stage to effectively mitigate issues such as word duplication and the inclusion of near-synonyms.

The main work presented in this paper can be summarized as follows:

1. We provide a multitask training strategy that combines the monolingual summary task with the cross-lingual summary task. By using the model features and combining the inputs of both tasks, we achieve hard parameter sharing of the overall process, which eliminates task differences and reduces the semantic loss from segmentation tasks, thereby improving the performance of the cross-lingual summary model;
2. We optimize the model for multiple input and output features. To enhance the regularization ability of the fine-tuned model and reduce the risk of overfitting in downstream tasks, we improve the consistency of the model output by averaging the weights of the forward network with different dropout probabilities. Additionally, we incorporate a sparse set of softmax filtering predictors in the regularization process to improve the output accuracy. Furthermore, we streamline the pre-trained model parameters and customize the cross-lingual word list to reduce the training cost;
3. We design a targeted sequence generation and filtering mechanism based on the proposed model and method. We combine the monolingual summary sequences annotated with word tags to form consecutive fragments which effectively solve the problem of sequential alignment and loss of important information in different languages. Moreover, we use an external word list to ensure the occurrence of keywords in the source text in the summary, ensuring the generation of key information

and alleviating the problem of multiple meanings of words and that of words out of vocabulary (OOV).

The chapters in this paper are organized as follows: Section 2 provides an introduction to the related work on cross-lingual summarization and generation mechanisms. Section 3 elaborates on the methods proposed in this paper, including the strengthened regularization method, the ABO mechanism, and the multitask fusion method. Section 4 describes the experimental details and analysis of the results. The experiments cover ablation experiments with different sample magnitudes, model comparison experiments with full data sets, and single-language summary experiments. In Section 5, the process of constructing a cross-lingual summarization data set in the professional field and the results of the model ablation experiment on this data set are presented. Finally, Section 6 summarizes the main work presented in this paper and outlines future directions for research.

2. Related Work

2.1. Cross-Lingual Summarization

The primary task of cross-lingual summarization is to extract the essential content of a document in one language and express it in another language. Previous methods have typically combined translation tasks and single-text summarization tasks in different orders. There are two main approaches: (1) translating the corpus first and then extracting the abstract from the translated content [10–12]; (2) extracting the abstract first and then translating it into the target language [13–15]. Such a pipeline method is heavily influenced by the performance of the model and lacks any connection between tasks, which can easily lead to content deviation.

The Transformer-based deep neural network can improve the effectiveness of each component of the task. For instance, Cao [16] introduced a learnable linear mapper to the multitask framework to enhance the isomorphism between different languages and improve the cross-lingual transfer ability of the model. Similarly, Zhu [17] added a translation layer to the Transformer model to compute candidate translation words in the source document and added high-scoring candidates into summary generation to improve the summary quality. Luo [18] introduced a cross-attention module to the Transformer encoder to establish interdependence between languages and enhance cross-lingual information correlation. Another approach proposed by Zhu [19] involved combining cross-lingual summarization tasks with single-language summarization tasks and translation tasks. This was achieved by sharing the encoder and using two independent decoders to process the two different tasks, enabling the model to learn the input of both tasks and improve the performance of cross-lingual summarization models. Bai [20] spliced the MS output from the decoder and the CLS output sequentially while sharing the CLS encoder. This approach enhanced the interaction between different languages, implicitly considering cross-lingual alignment, semantic similarity, and patterns between summaries in different languages, which facilitate knowledge transfer from high-resource languages to low-resource languages. Liang [21] employed a conditional variational autoencoder [22] with a shared encoder and decoder for multitask learning of machine translation (MT), MS and CLS tasks. The authors constructed two local-level latent variables for translation and summarization, respectively, and a global-level latent variable for CLS.

Previous studies have primarily focused on enhancing the model's cross-lingual learning capability during training for various tasks. However, an essential aspect that impacts summary quality is how the model transforms discrete predicted word distributions into continuous sequences during the generation phase.

2.2. Auxiliary Generation Method

Although auxiliary generation methods have been shown to improve the performance of CLS models by incorporating other tasks, the cost of training with multiple auxiliary tasks is often high. To address this issue, feature enhancement methods extract and optimize the internal features of the CLS process and apply them to the model to enhance

its semantic understanding, cross-linguistic alignment, and text generation capabilities. This approach enhances the interaction between features within the CLS task and is not dependent on specific tasks, but exploiting internal features requires the development of complex algorithms.

Zhu [17] selected keywords from the source text, obtained their translation distributions through a probabilistic bilingual dictionary, and then used the output distribution of the Transformer along with the translation distribution to generate summaries. This method extends the influence of source text keywords on generating the final summary. Duan [23] added a comparison attention mechanism to the Transformer. This attention is calculated from the parameters and weights inside the Transformer to increase attention to irrelevant information between the target language reference summary and the source language text. The method extends the influence of model parameters and weights on generating the final summary. Jiang [24] first extracted key cues, such as keywords and named entities, from the source language text, transformed the source language text into a text graph using a cue-guided algorithm, and then constructed a graph encoder and a cue encoder to encode the text graph and key cues, respectively. The respective outputs were passed into the decoder and finally the output distribution and translation distribution were used together to generate the summary. This method extends the influence of key cues on generating the final summary.

3. Methods

The model proposed in this paper is based on the standard seq2seq structure. It utilizes the baseline model to perform two different tasks consecutively, sharing all parameters within the model. In the fine-tuning process, a reinforced regularity approach is incorporated. The input sequence passes through a forward network twice, with different drop probabilities, which is approximated as passing through two distinct sub-networks. The prediction word distribution is then filtered by sparsifying the output, and the distribution results are obtained by calculating the *KL* scatter [25]. Regularization is provided by reinforced regularization, which introduces randomness during training. On the other hand, sparse softmax promotes sparsity, preventing the model from overly relying on specific features or classes. These regularization techniques help the model generalize better and enhance its robustness to unseen examples. In the generation phase, the ABO mechanism is added, which combines the core semantic content of monolingual summaries and the cross-lingual summary sequences generated by screening the near-sense candidates. Important semantics are preserved through the use of selected tag words, while the correctness of linguistic sequences is maintained by leveraging potential connections between these tag words. The overall flow of the model is illustrated in Figure 1.

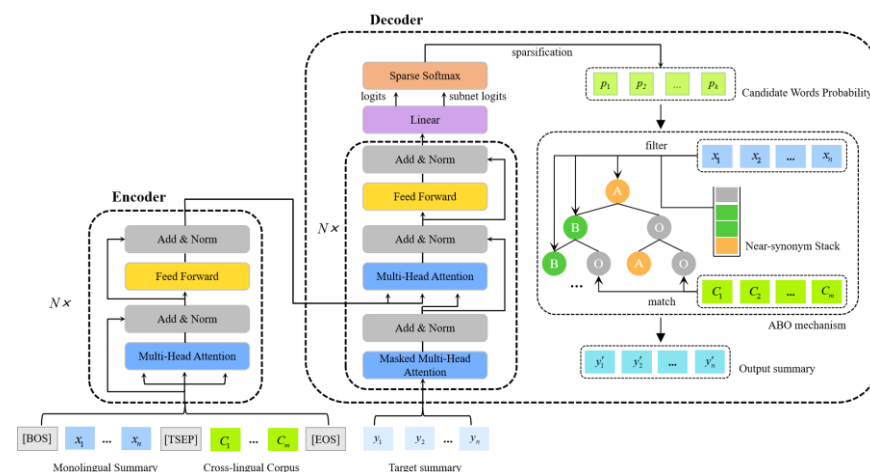


Figure 1. Model structure schematic.

In the figure, a linguistic corpus $C = \{C_1, C_2, \dots, C_n\}$ is provided, which corresponds to a monolingual summary sequence $X = \{x_1, x_2, \dots, x_n\}$ and a target summary sequence $Y = \{y_1, y_2, \dots, y_n\}$. Input sequence starts with $\{[BOS], x_1, x_2, \dots, x_n, [TESP], C_1, C_2, \dots, C_n, [BOS]\}$ and the final target output sequence is obtained after being processed by the encoder, ending with $\{[BOS], y'_1, y'_2, \dots, y'_n, [EOS]\}$. The identifier $[TESP]$ represents the truncate separate character.

3.1. Reinforced Regularization Method

Fully trained large language models typically have a high number of model parameters, which can provide substantial prior knowledge, but the limited data resources available for downstream tasks can result in overfitting during model fine-tuning. Therefore, it is necessary to introduce an appropriate regularization strategy to reduce overfitting and improve the model's generalization performance. The baseline model does not incorporate dropout regularization during the pre-training phase because it may inhibit the model's fitting effect and reduce its learning ability. However, during the fine-tuning phase, the model may suffer from overfitting when facing a single downstream task, so adding a more effective regularization method can significantly improve the fine-tuning performance of the model. As a result, a reinforced regularization method is added during fine-tuning.

During training, performing forward calculations with different sampling processes for the same input x is equivalent to the same data sample going through two sub-networks and obtaining two different distributions, P_1 and P_2 . The final weighting of the cross-entropy of the two components is shown in the following equation:

$$L_i^{(CE)} = -\log P_\theta(y_i|x_i) - \log P'_\theta(y_i|x_i). \quad (1)$$

To keep the output of the path network consistent across dropouts, the KL scatter is calculated.

$$L_i^{(KL)} = \frac{1}{2} [KL(P_\theta(y_i|x_i) \parallel P'_\theta(y_i|x_i)) + KL(P'_\theta(y_i|x_i) \parallel P_\theta(y_i|x_i))]. \quad (2)$$

The final loss after two loss calculations is the weighted sum of the two losses, which is calculated as shown below:

$$L_i = L_i^{(CE)} + \alpha L_i^{(KL)}. \quad (3)$$

In the equation above, α represents the weight coefficient of the KL loss and is the only hyperparameter. By weighting the two losses with α , the model space is further regularized, compensating for the inconsistency of dropout in training and testing, and improving the model's generalization ability.

The above method performs data augmentation while maintaining the input semantics, which boosts the confidence level of the primary categories but increases the training cost by adding numerous non-target classes in the prediction stage. To address this issue, we sparsify the output to provide a positive gain for the regular method via the category-invariant property.

Typically, the softmax function is capable of mapping multiple neuron outputs to the $(0, 1)$ interval, thereby enabling the numerical assignment of approximate probabilities. The common exponential form of softmax [26] is calculated as

$$p_i = \text{softmax}(s_i) = \frac{e^{s_i}}{\sum_{j=1}^n e^{s_j}}. \quad (4)$$

In the above equation, s_i represents the score of an output result, p_i represents the corresponding probability, and n denotes the total number of output results. The softmax function transforms the set of n real-valued scores into a probability distribution.

The decoder part uses softmax for two main functions: (1) calculating the normalized attention weights, and (2) computing the prediction probability distributions. However, traditional softmax cannot assign a probability of 0 to any predictor, making it impossible

to exclude low-probability predictors. To address this, sparse softmax is utilized instead, where the hyperparameter k is manually set to fix the category and complete the initial screening of low-probability words.

$$p_i = \begin{cases} \frac{e^{S_i}}{\sum_{j \in N_k} e^{S_j}} & i \in N_k \\ 0 & i \notin N_k \end{cases} \quad (5)$$

The output logits of the fully connected layer are denoted as S_i . These logits are sorted from largest to smallest, and the set of subscripts of the first k elements is denoted as N_k . At this point, sparse softmax only retains the probability values of the first k elements after sorting when calculating the probability. The values of the remaining elements after k will be directly set to 0.

As shown in Figure 2, same sequence is input into two forward networks, each with different dropout probabilities, resulting in distinct scores after computation. Before passing through the softmax function, the computed results from the two networks need to be sorted. Additionally, manual parameters are used to sparsify the distributions. When the hyperparameter k is set to 3, sparse softmax considers the top 3 logits after sorting as the target class and calculates their corresponding probabilities, while the probabilities of logits in non-target classes are set to 0 directly. Based on the results, the sparse output prevents the probability distribution from being wasted on unlikely outputs and significantly enhances the accuracy of the output. In tasks with low output ambiguity, the sparse output typically requires producing only one or a few fixed sets containing the correct answers.

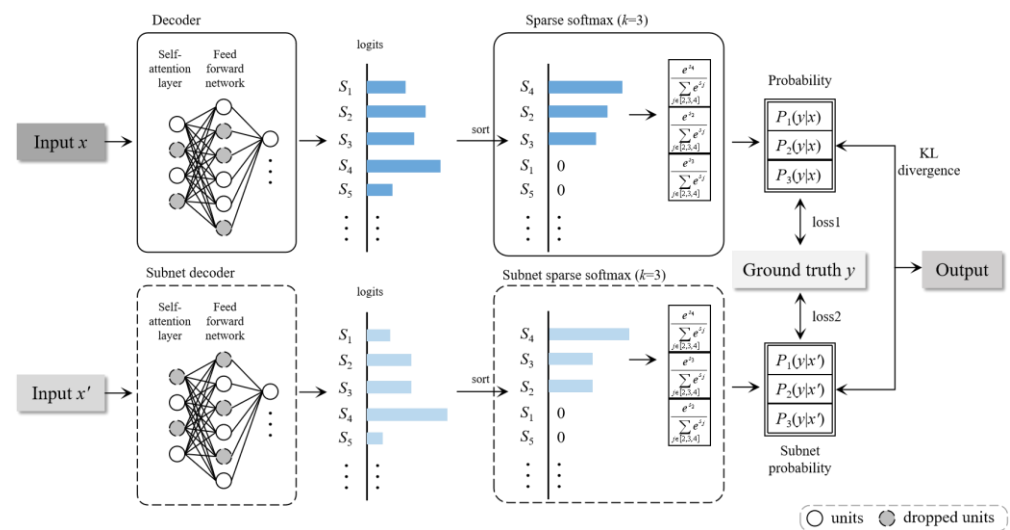


Figure 2. Schematic diagram of the enhanced regularity function.

Although the aforementioned approach involving reinforced regularity includes two forward calculations during training, it effectively alleviates the problem of long-tailed multicategorical distribution, improves the confidence level of the main categories, and significantly reduces the search space of the model. As a result, the pre-trained model’s robustness in downstream tasks is improved while maintaining computational efficiency.

3.2. ABO Mechanism

The most common problem faced by generative cross-lingual summary models is the semantic drift between different languages. This occurs when the generation of a summary in another language results in a representation that is contrary to the original due to issues such as sequence length or near-synonyms, even though a better representation has been obtained with the monolingual summary of that corpus. To tackle this issue, we filter

out cross-lingual summaries that exhibit semantic errors by comparing them with the monolingual summaries generated by the same model.

Before generating cross-lingual summaries, the model tags the monolingual summary sequences. The token with the largest L2 paradigm among the candidate words is selected as the anchor word (A), and the neighboring words are gradually identified through pointers. Assuming that the set of candidate words is $P = \{p_1, p_2, \dots, p_n\}$, A tag word for p^A , then p^A is

$$p^A = \underset{i}{\operatorname{argmax}} \|p_i - \bar{p}\|_2. \quad (6)$$

When a word is identified during the traversal of the sequence that is not present in the original corpus, it is marked as an out-of-text word (O), and all the words that have been traversed are marked as bound words (B). O -labeled words refer to the words generated to summarize the semantics of paragraphs in monolingual summaries, and preserving them is crucial for maintaining semantic coherence in cross-lingual summarization. Additionally, multiple consecutive B -labeled words serve as the basis for forming a smooth abstract word order. During marking, a binary tree with fixed combination rules is maintained, with A as the root node, and only B and O nodes allowed under A nodes, B and O nodes under B nodes, and A and O nodes under O nodes.

During cross-lingual summary generation, the same approach is used to select the anchor word tagged as A , but it is restricted to selecting a word that has the translated word tagged as A in the monolingual summary or its near-synonym as the anchor word in the cross-lingual summary sequence. Otherwise, the candidate sequence is considered to be missing the core word and discarded. To prevent unmatched cases, we use a two-stage out-stack operation to determine whether core words are present in the sequence. The O -tagged words in monolingual summaries are copied into the sequences of cross-lingual summaries through the translation word list. Using the properties of the A -tagged words and the O -tagged words, we ensure that the core words appear in the cross-lingual summary and maintain the basic semantic structure. For the B -tagged words, we ensure the closeness of meaning and alignment of the word order through filtering. Specifically, we maintain a list of proximity words for B -tags in monolingual summaries and traverse the cross-lingual summary candidate sequence, skipping if a word is found and discarding if it is not found by more than 2-g. The specific implementation is described in Algorithm 1.

To generate a summary sequence, labeling the summary words in the source corpus is necessary. Therefore, predicting the label distribution during training is crucial. We calculate the loss function by computing the cross-entropy loss between the predicted label distribution and the ground truth label distribution for all samples. The cross-entropy loss quantifies the dissimilarity between the predicted and true label distributions. By minimizing this loss, the model's predictions are improved and aligned with the actual labels, enhancing its performance. The calculation method is as follows:

$$L^{(TAG)} = -\frac{1}{NS} \sum \sum_i z'_i \log z_i. \quad (7)$$

In the above formula, N is the length of the input corpus sequence, S is the batch size, z_i is the tagged word, and z'_i is the predicted tag word. To account for the sparsity of label predictions, we average the loss over the batch size ($\frac{1}{NS}$). This ensures that the loss remains consistent and independent of the batch size, allowing for fair comparisons between different batch sizes. Combined with Equation (3), the loss is calculated as

$$L_i = L_i^{(CE)} + \alpha L_i^{(KL)} + L_i^{(TAG)}. \quad (8)$$

In Figure 3, the O appearing in the second and third levels of the binary tree is not a real node but a determination step during the construction of monolingual summary generation to check whether the next word exists in the original text. Besides the probability

distribution output by the decoder, the “filter” in the figure also includes other conventional determination methods, such as basic grammar rules, conventional phrase combinations, and sequence length restrictions.

Algorithm 1 ABO mechanism

Algorithm implementation:

Input: monolingual summarization sequence S_m , tag set $T(a, b, o)$, cross-lingual summarization candidate token set C_s , synonym dictionary S_d

Output: cross-lingual summarization output sequence S_c

```

1.  for  $s \in S_m$  do
2.    if MATCH_TAG( $s, o$ ) then
3.       $s' \leftarrow$  TRANSLATE( $s$ )
4.      add( $s'$ ) to the set  $S_c$ 
5.    else
6.       $S_t \leftarrow$  SEARCH( $s, S_d$ )
7.    end if
8.    for  $t \in$  TRANSLATE( $S_t$ ) do
9.      if MATCH_TAG( $t, a$ ) or MATCH_TAG( $t, b$ ) then
10.       PUSH( $t$ )
11.      else
12.       Break
13.      end if
14.    end for
15.  end for
16. for  $c \in C_s$  do
17.  if !EQUAL( $c, s'$ ) or !GRAMMER_RULE_FILTER( $c, S_c$ ) then
18.   remove  $c$  from  $C_s$ 
19.  else if !JUDGE_INTERSECTION( $c, S_c$ ) then
20.    $c \leftarrow$  pop( $T$ )
21.   break
22.  else
23.   add( $c$ ) to the set  $S_c$ 
24.  end if
25. end for
26. return  $S_c$ 

```

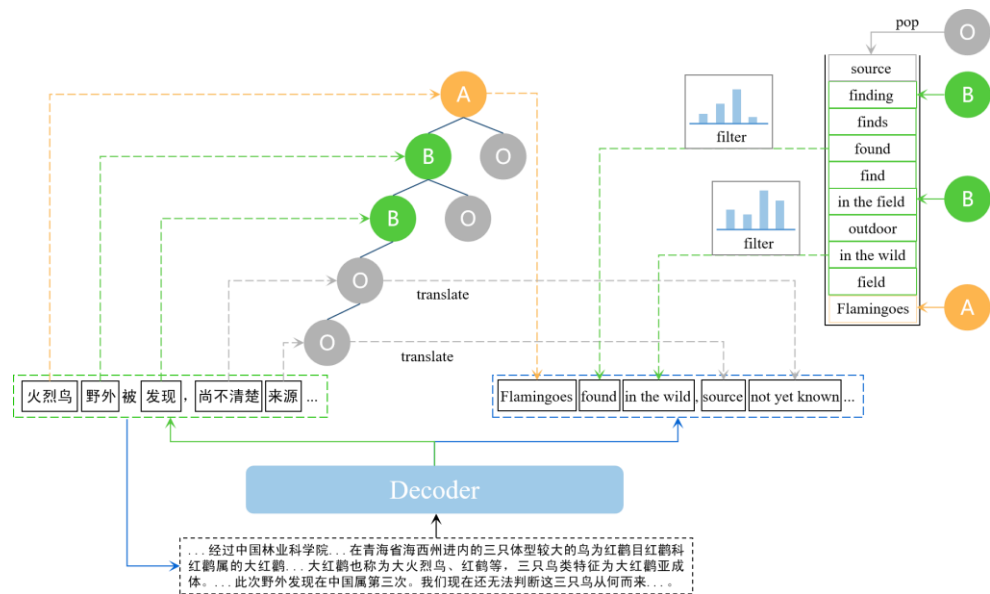


Figure 3. ABO mechanism diagram.

3.3. Multitask Fusion Approach

To fully leverage crucial information from the corpus using multiple tasks, a model with hard parameter sharing is used to embed the data representation of various tasks in the same semantic space [27]. As a result, the decoder is shared by multiple tasks, and the target sequence is replaced by a combination of multitask sequences. The overall loss calculation becomes

$$L_{sum} = \beta_1 L_i^{(1)} + \beta_2 L_j^{(2)}. \quad (9)$$

In the above equation, $L_i^{(1)}$ is the loss function of the monolingual summary task which includes the weighted sum with added *KL* scatter, as mentioned above. The term $L_j^{(2)}$ includes the joint probability of both the translation task and the cross-lingual summary task. Given a monolingual corpus C_i , the model generates the summary content X_i by performing the monolingual summary task. The target sequence consists of a probability distribution of $p(y_t|y < t, X_i, C_i)$. The expression is as follows:

$$L_j^{(2)} = \sum_{t=1}^n \log P(y_t|y < t, X_i, C_i). \quad (10)$$

In Equation (9), β_1 and β_2 are the weights that adjust the multitask loss. However, since the overall model can converge slowly due to the larger loss value after weighting, the learning rate needs to be reduced. To address this, the weights are dynamically adjusted by calculating the state of each task, setting such weights in the following way:

$$\beta_n = \frac{1}{\|\nabla_{\theta} L_{\theta}^{(n)}\|}. \quad (11)$$

In actual calculations, $L_{\theta}^{(n)}$ is fixed after one derivative to maintain numerical stability of the overall gradient.

4. Experiments

4.1. Datasets

In the cross-lingual summarization experiments, our main focus is on both Chinese–English and English–Chinese summary approaches. Therefore, in our experiments, we mainly use two datasets, En2ZhSum and Zh2EnSum [19]. En2ZhSum is converted by the back-translation method using the general text summarization datasets CNN/DailyMail [28] and MSMO [29]. The dataset is divided into 364,687 training data pairs, 3000 evaluation data pairs, and 3000 test data pairs. Zh2EnSum is created by converting the summary part of the large open-domain Chinese dataset LCSTS [30] into English by the same method. It contains 1,693,713 Chinese-to-English training samples, 3000 evaluation data pairs, and 3000 test data pairs.

For the ablation experiments, we first conducted experiments on the methods proposed in this paper, retaining the settings of minimum, medium, and maximum to reflect the differences among methods while controlling the number of samples. The specific sample size settings are listed in Table 1.

Table 1. Dataset sizes of multiple low-resource scenarios for CLS datasets.

Scenarios	Minimum (Pairs)	Medium (Pairs)	Maximum (Pairs)	Full Dataset
Zh2En	5000	25,000	50,000	1,693,713
En2Zh	1500	7500	15,000	364,687

4.2. Experimental Settings

In this paper, we combined the approach mentioned above with the Transformer structural pre-training model to construct an end-to-end model. Before the experiment, we

first simplified the model's input. After taking into account the statistics, we found that the parameters of the model input and output layers accounted for 65% of the total number of parameters. However, since our experiments only included Chinese and English, we removed most of the other language contents from the word list. Following the processing method of mBERT [31], we streamlined the word list and added some commonly used Chinese words in a targeted manner. Finally, the overall word list included 11,000 English words, 30,000 Chinese words, and 100 special symbols, reducing the number of words to 20% of the original word list. We then processed the sentencepiece word splitter of the original model by replacing the first 40,000 words and removing other irrelevant content. After manual debugging, we set the hyperparameters $\alpha = 4$ for the reinforcement regular and $k = 10$ for the Sparse softmax. Since the model uses Adafactor [32], a larger initial learning rate was chosen, and we set the initial learning rate of the model to be 2×10^{-4} . We used different prefix task identifiers to distinguish the monolingual summarization task from the cross-lingual summarization task and trained the model by alternating the two tasks. This part of the automatic evaluation compares the performance differences of several models by the standard ROUGE [33] method and shows the ROUGE-1, ROUGE-2 and ROUGE-L scores.

4.3. Ablation Study

This section of the experiments explores the performance improvement of the proposed methods on different scales of cross-linguistic summary datasets, En2ZhSum and Zh2EnSum, by controlling the joint models. The experimental models are divided into three parts: Base, which is the base model; Base + RR, which is the model with the addition of reinforced regularity; and Base + RR + ABO, which is the model with the addition of reinforced regularity and the ABO mechanism. Base + RR + ABO uses the original monolingual summary sentence in the dataset. The experimental results are compared in the table below.

The results in Table 2 indicate that the model achieved a quantitative improvement with the addition of the reinforcement regularity for a small sample size. This is because during training, the same samples are forward computed twice, which is formally equivalent to augmenting the overall dataset, and repeated learning reduces the risk of overfitting the pre-trained model on a small number of samples. Additionally, the sparse output reduces the search range of the predicted words. When combined with the regularization method, it provides higher reliability for a small number of prediction categories under artificial constraints. Thus, the overall reinforced regularity approach improves the effectiveness of the pre-trained model on specific downstream tasks.

Table 2. Results of ablation experiments with different sample size data sets.

Scenarios	Model	Zh2EnSum			En2ZhSum		
		RG1	RG2	RGL	RG1	RG2	RGL
Minimum	Base	21.61	5.88	16.19	33.04	11.45	18.04
	Base + RR	23.53	6.45	18.25	34.14	12.38	20.17
	Base + RR + ABO	25.37	7.50	20.47	35.69	14.36	23.65
Medium	Base	27.22	9.91	23.33	35.38	14.54	24.12
	Base + RR	28.46	12.11	25.36	36.34	17.01	26.20
	Base + RR + ABO	30.21	14.57	27.18	37.22	18.67	28.23
Maximum	Base	29.35	11.78	25.23	36.35	18.44	26.14
	Base + RR	31.10	12.93	26.51	37.24	19.87	27.26
	Base + RR + ABO	33.08	14.72	28.14	38.11	21.25	28.35

After the addition of the ABO mechanism to the model, the generated summaries showed more improvement in the ROUGE-2 and ROUGE-L scores. This is mainly because the ABO mechanism preserves the proximity of candidate words when selecting anchor

words and bound words. As a result, the semantic span between anchor words and bound words is not suddenly increased after filtering by basic grammar rules. Additionally, the proximity of candidate words and phrases of bound words can be used as the basis for the calculation of the same subsequence in 2-g and above. Therefore, the summaries generated by the model incorporating the ABO mechanism usually have higher fluency.

4.4. Cross-Lingual Summary Comparison Experiments

This section performs cross-lingual summary comparison experiments using the full dataset of two cross-lingual datasets, En2ZhSum and Zh2EnSum, which are divided into automatic and manual evaluations.

4.4.1. Contrast Model

TLTran: Transformer-based Late Translation. The model is a pipeline approach with a monolingual summary model as the main body. The method first uses the source document as input and generates the same language summary, and then translates the generated summary into the target language summary by the translation model.

TETran [12]: Transformer-based translation priority model. This method first translates the source document into the source document of the target language, and then extracts the target summary from the translated corpus through another trained single-language summary model of the Transformer structure.

TNCLS [19]: Transformer-based Neural Cross-Lingual Summarization (TNCLS). The model accomplishes the cross-lingual summarization task by jointly training encoders and decoders for different languages and different input sequence lengths.

CLS + MS [20]: Combining Cross-Lingual Summarization with Monolingual Summarization. A shared encoder is used to encode the input utterances, and the decoder for the Cross-Lingual Summarization task is connected to the Monolingual Summarization decoder at the same time to train both tasks in a unified manner.

CLS + MT: Combining Cross-Lingual Summarization (CLS) with a translation task (Machine Translation) [17]. The decoder for the translation task part is trained by an additional translation corpus and alternatively trained using a shared encoder while adding the decoder for the Cross-Lingual Summarization task.

4.4.2. Experimental Results Analysis

Table 3 shows that our model outperforms other cross-lingual summarization methods on both datasets. It is known that pipelined models usually cannot achieve the same results as end-to-end models since multitask integration often shares some of the parameters, which mitigates the loss of transformation between different tasks [34]. Although models that integrate multiple tasks have significant advantages over pipelined models, CLS + MS and CLS + MT still share parameters in the encoder part, but can only pass single-task content in the attention part of the encoder. The single-language summary task during training allows the overall model to learn the content of the single-language summary part repeatedly, which helps the model learn the corpus more accurately. Additionally, the reinforced regularization method proposed in this paper can avoid training-time oscillations caused by random noise in the new input sequence before generating the cross-lingual digest. This stabilizes the training process of the model and fully utilizes the performance of the pre-trained model. The ABO mechanism proposed in this paper can also avoid this problem since the anchor words and bound words in ABO can easily form consecutive clauses that potentially affect the order of the generated summaries. This motivates the formal alignment of the generated summaries with the target summaries, avoiding the problem when the whole is divided by subwords, which can lead to the unbalanced number of generated words in both languages and information loss. The display of the generated results is shown in Figure 4.

Table 3. Comparison table of cross-lingual summarization experiments.

Model	Zh2EnSum			En2ZhSum		
	RG1	RG2	RGL	RG1	RG2	RGL
TLTran	30.36	12.36	29.21	31.13	12.91	25.13
TETran	21.62	10.52	18.93	24.35	11.77	23.39
TNCLS	33.21	16.24	29.04	34.40	22.34	27.05
CLS + MS	35.35	16.67	30.28	36.82	23.75	28.71
CLS + MT	36.09	16.71	30.16	37.13	23.22	28.75
ours	37.56	17.98	32.48	38.95	24.50	30.33

Source	长时间看手机感到眼睛酸胀痛? ...屏幕对视力到底有怎样的影响? ...新闻频道今日为你解答! 经常用手机的你, 有感到视力受影响吗? Look at the phone for a long time and feel sore eyes? ...What exactly does the screen have on vision?... News Channel will answer for you today! Do you, who use mobile phones a lot, feel that your eyesight is affected ?
Reference	今日关注: 大屏幕手机会造成视力下降吗? Today's attention: Will large-screen mobile phones cause vision loss ?
TLTran	When I look at my phone for a long time, my eyes feel sore.
TETran	My eyes feel sore and swollen from looking at my phone for long periods of time.
TNCLS	I feel sore eyes when I look at my mobile phone for a long time.
CLS+MS	Have you felt that your vision is affected by using your mobile phone a lot?
CLS+MT	Looking at the mobile phone for a long time makes your eyes sore. Do you feel your vision affected .
ours	How long does it affect your eyes to look at your mobile phone ? News Channel will answer it for you today.

Figure 4. Cross-lingual summarization generation example.

4.4.3. Human Evaluation

This section of the manual evaluation aims to compare the usefulness of the generated summaries. We randomly selected 20 samples from each of the En2ZhSum and Zh2EnSum test sets, and three evaluators rated the summaries on a scale of 1 (worst) to 5 (best) based on their informativeness (IF), conciseness (CC), and fluency (FL). The average score for each group was calculated based on the total sample size.

Table 4 shows that the model proposed in this paper outperforms other models in terms of completeness, indirection, and fluency. In terms of fluency, the ABO mechanism combines consecutive fragments, allowing the model to select words that match back and forth, resulting in a subjectively fluent sentence. Regarding conciseness, the model in this paper is similar to the CLS + MS and CLS + MT models, as both set the output with the same sequence length in the encoder part and use alternating tasks in the decoder part. However, our model uses a unified encoder and decoder to integrate multiple tasks, enabling the model to learn the corpus content more comprehensively. Additionally, compared to the multitask framework that uses multiple decoders, our model relies more on the transfer of hidden states between encoders and decoders, resulting in a more holistic content transfer, which provides a clear advantage in terms of completeness and fluency.

Table 4. Cross-lingual summary of human evaluation results.

Models	IF	CC	FL
TLTran	3.23	3.31	3.53
TETran	3.45	3.22	3.33
TNCLS	3.46	3.68	3.61
CLS + MS	3.53	3.83	3.91
CLS + MT	3.68	3.76	3.77
ours	3.86	4.05	4.21

4.5. Monolingual Summary Comparison Experiments

To demonstrate that our model does not suffer from performance degradation in single-language summarization tasks due to partial repetition during parallel training on multiple tasks, we conducted separate experimental validations for text summarization tasks in Chinese and English. As the ABO mechanism does not incorporate the translation process with the near-synonym word list in single-language summarization, we searched for anchor words and bound words in the original text to ensure summary generation fidelity compared to the original text.

To showcase the practical effects of other optimization methods on pre-trained models, we conducted incremental and non-incremental experimental models in addition to the monolingual summary comparison experiments. The initial model parameter settings in the incremental experimental model were the same as those in the cross-lingual summarization experiments. The baseline models in both Base and Base + RR groups were not fine-tuned for the cross-lingual summarization task. The non-incremental experimental model retained the same part of the monolingual summarization model used by TLTran and TETran, and only the shared encoder and monolingual summarization decoder were kept in the CLS + MS model. The TNCLS and CLS + MT models do not contain a separate monolingual summarization part, so this experiment was not included. The experimental results are shown in the figure below.

Figure 5 illustrates that the model with enhanced regularization has significantly improved the single-language summarization task compared to the baseline model. For the baseline that does not use dropout in the pre-training stage, enhanced regularization can better ensure the consistency between the model and downstream tasks. In incremental experiments, our model has significantly improved the longest subsequence calculation index compared to Base and Base + RR due to the continuous fragments generated by the combination of ABO labels. Compared with simple beam search and other methods, it can better establish the connection between n-grams. In non-incremental experiments, TLTran is a Transformer monolingual summary model trained from scratch. Due to data and training condition limitations, its effect lags far behind other fine-tuned pre-training models. NCLS + MS, as a combination of mBert and Transformer decoder, cannot share the overall parameters, which results in insufficient interaction of the model in the training of cross-lingual summarization and monolanguage summarization tasks. However, our model fully learns the information of the two language summaries and maintains the sharing of the parameter space in the interactive training, ensuring a positive impact of both tasks on the model.

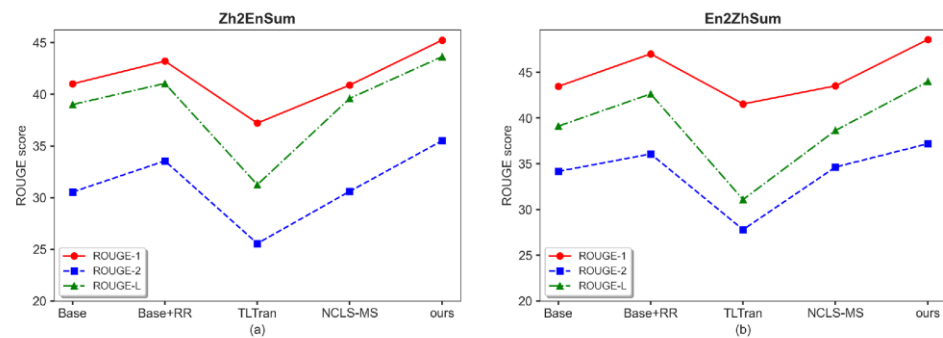


Figure 5. Monolingual summarization experiments comparison chart. (a) Shows the experimental results of single-language summarization on the Zh2EnSum data set. (b) Shows the experimental results of single-language summarization on the En2ZhSum data set.

5. Application

We created a cross-lingual summarization dataset that includes English and Chinese using proprietary domain data. The data set is sourced from a combination of internal maintenance manuals from an automobile company, instructional materials, and the text component of FETA Car-Manuals [35]. The 357 PDF documents were subjected to OCR recognition, and the text component was extracted. Each document in the dataset contains a varying number of text paragraphs with different topics, resulting in a total of 4472 text paragraphs that were manually separated. As only some text paragraphs contain headings, the process of summarization required the selection of sentences based on the three most frequently occurring keywords in texts without headings or with headings that were too short. After counting the word frequency, the corresponding sentences were manually screened and the selected sentences were combined and spliced to create a summary of the corresponding paragraph. The length of the combined abstract was kept at approximately 15% of the length of the original text. The following Figure 6 is the length statistics of the processed summary dataset.

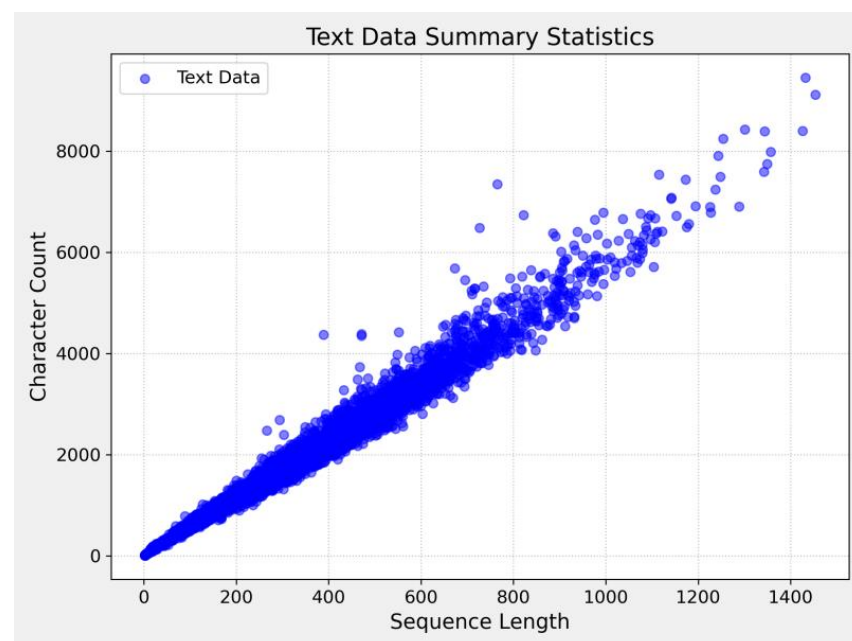


Figure 6. Statistical chart of processed data word count and sequence length.

In the cross-lingual summary section, we utilized a translation tool to batch translate the text. Furthermore, we also extracted the professional vocabulary from this dataset. Table 5 below presents the final statistics of the document.

Table 5. Statistics of the CarManualSum.

	CarManualSum (cn2en)			CarManualSum (en2cn)		
	Train	Valid	Test	Train	Valid	Test
Number of samples	3460	500	500	3460	500	500
Average words in text	133	133	132	1034	1040	1031
Average words in summary	21	21	20	124	125	124

We intend to conduct experiments on this dataset to validate the effectiveness of the proposed method in improving the model’s performance on professional datasets. However, due to the limited number of datasets and the vast professional vocabulary, initializing the training model for comparison experiments poses a significant challenge. Thus, we only used the fine-tuned baseline model and the proposed method for ablation experiments in this test section. The experimental settings employed here are consistent with those in Section 4.2. To maximize the utilization of the limited dataset and mitigate the risk of model overfitting, we employ a fivefold cross-validation during the training process. The evaluation methodology remains the same, utilizing ROUGE-1, ROUGE-2, and ROUGE-L scores. The experimental results are shown in the following table.

Based on Table 6 above, it is evident that the proposed method enhances the performance of the model on specialized domain datasets. Even with a limited number of datasets, the proposed method effectively maintains the model’s ability to generate high-quality summaries. Notably, the utilization of the ABO mechanism enables the model to generate cohesive and fluent summaries efficiently, without requiring extensive training data.

Figure 7 demonstrates the examples of model generation based on the CarManualSum dataset. The datasets used in our study can be accessed through the following links: <https://github.com/leesin5079/Car-manual-CLS-dataset-1000> (accessed on 18 April 2023).

Sample 1	
Source	This manual contains the latest information at the time of printing...the hazards identified in the manual...and how to avoid or reduce the hazards.... The fit and adjustment of the head restraint is also very critical and if not properly fitted and adjusted, occupants are more likely to suffer neck/spine injuries in a crash. So until all occupants' head restraints are installed and properly adjusted,...
Reference	该手册提供最新车辆信息，注意安装和调整头枕以避免颈部受伤。外侧座椅位置的头枕可调节。 This brochure provides current vehicle information, emphasizing proper headrest installation and adjustment to prevent neck injuries. The outboard seating position headrests are adjustable for personalized comfort and enhanced safety.
ours	车辆手册包含最新信息，注意安装和调整头枕，避免颈部受伤。 The vehicle manual contains the latest information, pay attention to installing and adjusting headrests to avoid neck injuries.
Sample 2	
Source	When servicing a vehicle, attention must be paid to the size and structural identification of individual components to ensure proper replacement of parts. ... For example, when replacing a front end pull-up rod, a mild steel material is required. ..., you need to pay attention to the size of the opening on the side of the body. ..., when replacing the front wheel housings, mild steel front wheel housings need to be replaced. When replacing front wheel well extensions, attention also needs to be paid to materials and properties to ensure the correct durability and function of the part,...
Reference	在维修车辆时，需要注意组件的尺寸、结构和材料特性识别，以确保更换零件的正确性和耐用性。 When repairing vehicles, attention needs to be paid to the size, construction and identification of material properties of components to ensure correctness and durability of replacement parts.
ours	维修车辆时注意组件的尺寸、结构识别和材料特性。 Pay attention to the dimensions, structural identification, and material characteristics of components when repairing vehicles.

Figure 7. The content of the data set and the display of our model generation results.

Table 6. Results of the ablation experiment on the CarManualSum data set.

Model	CarManualSum (cn2en)			CarManualSum (en2cn)		
	RG1	RG2	RGL	RG1	RG2	RGL
Base	25.53	14.74	25.50	26.85	15.83	24.45
Base + RR	27.66	17.31	28.69	29.97	19.94	27.38
Base + RR + ABO	31.54	20.06	30.77	34.43	22.23	31.56

6. Conclusions

Owing to the phenomenon of information misalignment resulting from disparities in language and syntactic structures, the process of generating cross-language summaries often encounters notable challenges such as substantial semantic loss and errors in semantic expression. To address these issues, this paper presents a cross-language summarization model founded on the ABO mechanism. Primarily, a novel multitasking approach for cross-language summarization is introduced, which simplifies the diverse subtasks involved in this process by integrating monolingual summarization and cross-language summarization. This strategy facilitates the sharing of parameters within the unified model, thereby enhancing alignment across different languages. Additionally, a reinforced regularization method for model characteristics is proposed. This method elevates the performance of the model in cross-language text summarization tasks by enhancing regularization through the amalgamation of sub-networks with varying dropout probabilities, which introduces a controlled level of randomness. Furthermore, it employs sparsification of output categories to prioritize the inclusion of valid information. In the stage of summary generation, an innovative ABO mechanism is devised and incorporated to strengthen the correlation between summaries in different languages. This mechanism encompasses a predictive labeling and filtering mechanism to mitigate semantic loss and discrepancies in word order within the summaries generated by the model. The effectiveness of the proposed approach is empirically evaluated through experiments conducted on publicly available datasets, namely Zh2EnSum and En2ZhSum. Moreover, a proprietary domain-specific cross-language summary dataset, CarManualSum, is constructed to provide further insights into the performance of the method. Nonetheless, there are certain limitations associated with the proposed approach. Although the enhanced regularity approach and the ABO mechanism contribute to improvements in cross-language summarization, they currently lack the flexibility required to adapt to the majority of pre-trained models. Moreover, variations in results between the public dataset and the specialized domain dataset are observed. Consequently, future research will concentrate on exploring the effectiveness of the proposed method across different models and downstream tasks. Furthermore, efforts will be made to broaden the research scope by constructing corpora encompassing a wider range of languages, thereby enhancing the efficacy of the model for cross-language summarization within diverse language scenarios.

Author Contributions: Q.L. constructed the model, designed the experiments and constructed the applied dataset, and wrote the main part of the paper. W.W. defined the research direction, corrected the writing problems, and suggested changes to the Abstract and Introduction of the paper. Y.Z. provided financial support for the dataset and experiments of the paper, and suggested changes to the Conclusions of the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Science and Technology Innovation 2030—Major Project of “New Generation Artificial Intelligence” granted by Ministry of Science and Technology of China, grant number 2020AAA0109300. and 2022 Major R&D Special 03 and 5G Projects of Jiangxi Provincial Department of Science and Technology of China, granted by Department of Science and Technology of Jiangxi Province, China, grant number 20224ABC03A15.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Raw data are available through the links in the paper.

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

References

1. Wang, J.; Meng, F.; Zheng, D.; Liang, Y.; Li, Z.; Qu, J.; Zhou, J. A survey on cross-lingual summarization. *Trans. Assoc. Comput. Linguist.* **2022**, *10*, 1304–1323. [[CrossRef](#)]
2. Mohammadzadeh, A.; Sabzalian, M.H.; Zhang, C.; Castillo, O.; Sakthivel, R.; El-Sousy, F.F. *Modern Adaptive Fuzzy Control Systems*; Springer Nature: Berlin/Heidelberg, Germany, 2022; Volume 421.
3. Wan, X. Using bilingual information for cross-language document summarization. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, OR, USA, 19–24 June 2011; pp. 1546–1555.
4. 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](#)]
5. Pires, T.; Schlinger, E.; Garrette, D. How multilingual is multilingual BERT? *arXiv* **2019**, arXiv:1906.01502.
6. Xue, L.; Constant, N.; Roberts, A.; Kale, M.; Al-Rfou, R.; Siddhant, A.; Barua, A.; Raffel, C. mT5: A massively multilingual pre-trained text-to-text transformer. *arXiv* **2020**, arXiv:2010.11934.
7. Mohammadzadeh, A.; Sabzalian, M.H.; Castillo, O.; Sakthivel, R.; El-Sousy, F.F.; Mobayen, S. *Neural Networks and Learning Algorithms in MATLAB*; Springer Nature: Berlin/Heidelberg, Germany, 2022.
8. Joachims, T. Optimizing search engines using clickthrough data. In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, AB, Canada, 23–26 July 2002; pp. 133–142.
9. See, A.; Liu, P.J.; Manning, C.D. Get to the point: Summarization with pointer-generator networks. *arXiv* **2017**, arXiv:1704.04368.
10. Leuski, A.; Lin, C.Y.; Zhou, L.; Hermann, U.; Och, F.J.; Hovy, E. Cross-lingual c* st* rd: English access to hindi information. *ACM Trans. Asian Lang. Inf. Process. TALIP* **2003**, *2*, 245–269. [[CrossRef](#)]
11. Ouyang, J.; Song, B.; McKeown, K. 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), Minneapolis, MN, USA, 1 June 2019; pp. 2025–2031.
12. Wan, X.; Luo, F.; Sun, X.; Huang, S.; Yao, J.G. Cross-language document summarization via extraction and ranking of multiple summaries. *Knowl. Inf. Syst.* **2019**, *58*, 481–499. [[CrossRef](#)]
13. Lim, J.M.; Kang, I.S.; Lee, J.H. Multi-Document Summarization Using Cross-Language Texts. In Proceedings of the NTCIR, Tokyo, Japan, 2–4 June 2004.
14. Orašan, C.; Chiorean, O.A. Evaluation of a cross-lingual romanian-english multi-document summariser. In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco, 26 May–1 June 2008.
15. 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; pp. 917–926.
16. Cao, Y.; Liu, H.; Wan, X. Jointly learning to align and summarize for neural cross-lingual summarization. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 6220–6231.
17. 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; pp. 1309–1321.
18. Luo, F.; Wang, W.; Liu, J.; Liu, Y.; Bi, B.; Huang, S.; Huang, F.; Si, L. VECO: Variable and flexible cross-lingual pre-training for language understanding and generation. *arXiv* **2020**, arXiv:2010.16046.
19. Zhu, J.; Wang, Q.; Wang, Y.; Zhou, Y.; Zhang, J.; Wang, S.; Zong, C. NCLS: Neural cross-lingual summarization. *arXiv* **2019**, arXiv:1909.00156.
20. Bai, Y.; Gao, Y.; Huang, H. Cross-lingual abstractive summarization with limited parallel resources. *arXiv* **2021**, arXiv:2105.13648.
21. 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.
22. Sohn, K.; Lee, H.; Yan, X. Learning structured output representation using deep conditional generative models. In Proceedings of the 28th International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; p. 28.
23. Duan, X.; Yu, H.; Yin, M.; Zhang, M.; Luo, W.; Zhang, Y. Contrastive attention mechanism for abstractive sentence summarization. *arXiv* **2019**, arXiv:1910.13114.
24. Jiang, S.; Tu, D.; Chen, X.; Tang, R.; Wang, W.; Wang, H. CptGraphSum: Let key clues guide the cross-lingual abstractive summarization. *arXiv* **2022**, arXiv:2203.02797.
25. Kullback, S.; Leibler, R.A. On information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. [[CrossRef](#)]
26. Bishop, C.M.; Nasrabadi, N.M. *Pattern Recognition and Machine Learning*; Springer: Berlin/Heidelberg, Germany, 2006; Volume 4.
27. Chen, Z.; Badrinarayanan, V.; Lee, C.Y.; Rabinovich, A. GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In Proceedings of the International Conference on Machine Learning, PMLR, Singapore, 10–15 July 2018; pp. 794–803.

28. Hermann, K.M.; Kocisky, T.; Grefenstette, E.; Espeholt, L.; Kay, W.; Suleyman, M.; Blunsom, P. Teaching machines to read and comprehend. In Proceedings of the 28th International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; p. 28.
29. Zhu, J.; Li, H.; Liu, T.; Zhou, Y.; Zhang, J.; Zong, C. MSMO: Multimodal summarization with multimodal output. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 4154–4164.
30. Hu, B.; Chen, Q.; Zhu, F. Lcsts: A large scale chinese short text summarization dataset. *arXiv* **2015**, arXiv:1506.05865.
31. Abdaoui, A.; Pradel, C.; Sigel, G. Load what you need: Smaller versions of multilingual bert. *arXiv* **2020**, arXiv:2010.05609.
32. Shazeer, N.; Stern, M. Adafactor: Adaptive learning rates with sublinear memory cost. In Proceedings of the International Conference on Machine Learning, PMLR, Vienna, Austria, 25 July–31 July 2018; pp. 4596–4604.
33. Lin, C.Y. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*; Association for Computational Linguistics: Barcelona, Spain, 2004; pp. 74–81.
34. Samant, R.M.; Bachute, M.R.; Gite, S.; Kotecha, K. Framework for deep learning-based language models using multi-task learning in natural language understanding: A systematic literature review and future directions. *IEEE Access*. **2022**, *10*, 17078–17097. [[CrossRef](#)]
35. Alfassy, A.; Arbelle, A.; Halimi, O.; Harary, S.; Herzig, R.; Schwartz, E.; Panda, R.; Dolfi, M.; Auer, C.; Saenko, K.; et al. FETA: Towards Specializing Foundation Models for Expert Task Applications. *arXiv* **2022**, arXiv:2209.03648.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.