Contrastive Aligned Joint Learning for Multilingual Summarization

Danqing Wang, Jiaze Chen, Hao Zhou, Xipeng Qiu †, Lei Li
ByteDance AI Lab
{wangdanqing.122,chenjiaze,zhouhao.nlp,lileilab}@bytedance.com
† Fudan University
xpqiu@fudan.edu.cn

Abstract

Multilingual text summarization requires the ability to understand documents in multiple languages and generate summaries in the corresponding language, which poses more challenges on current summarization systems. However, this problem has been rarely studied due to the lack of large-scale supervised summarization data in multiple languages. In this paper, we first provide a large-scale multilingual summarization corpus MLGSum consisting of 1.1 million articles and summaries in 12 different languages. Based on it, we develop a unified summarization model to understand the document and generate summaries in different languages. We use the contrastive learning strategy to train our multilingual summarization system (CALMS), which consists of two training objectives, contrastive sentence ranking (CSR) and sentence aligned substitution (SAS). The two training objectives are designed to share salient information extractive ability and align sentence-level representation across different languages. Experimental results indicate that CALMS achieves significant improvement over monolingual models in all languages. We further transfer CALMS to other languages and find that it will also benefit similar languages. Our code and dataset are available at https://github.com/brxx122/CALMS.

1 Introduction

Automatic text summarization aims at providing a brief summary for a long document. It requires the ability to understand document-level input, catch the main idea of it, and generate a fluent text. Recently, monolingual summarization has witnessed great success with the development of new neural systems (Zhong et al., 2020; Wang et al., 2020) and the availability of monolingual pre-training language models (Kenton and Toutanova, 2019; Liu and Lapata, 2019; Liu et al., 2019; Lewis et al., 2020b). , Inspired by the success of monolingual pre-trained models, researchers further pre-train these models with multiple languages to get the multilingual versions (Huang et al., 2019; Liu et al., 2020; Lewis et al., 2020a), which provide the abilities of understanding and generation in different languages. The multilingual pre-training model can be used as the initialization and finetuned for downstream summarization tasks.

However, the pre-training phase for language models usually focuses on predicting masked tokens or denoising the noisy input, both of which are token-level tasks. It lacks the ability to align sentence-level information among languages and to distinguish which information is the most critical for the document-level input. Most previous multilingual summarization models focus on training one model for different language or partly share encoder/decoder layers (Wang et al., 2018; Lin et al., 2018; Scialom et al., 2020). Cao et al. (2020) and Lewis et al. (2020a) try to train one model for all languages, but they find that although low-resource languages can benefit from the larger training data, the performance of rich-resource languages has been sacrificed. Thus, we want to investigate the following question: *Can we design a unified multilingual summarization model that can benefit both high-resource and low-resource languages?*

In this paper, we design a neural model with the contrastive aligned joint learning strategy for multilingual summarization (CALMS) with two new training objectives: contrastive sentence ranking (CSR) and sentence aligned substitution (SAS). CSR samples sentences from the document and constructs positive and negative pairs based on their saliency. By contrastively learning what is more important, the model is supposed to obtain the ability to distinguish salient information from the document. In order to align sentence-level information among languages, SAS replaces sentences with an-

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 2739–2750

other language and generates the summary based on the noisy input.

We conduct the experiments in five languages: English, Chinese, German, French, and Russian. The experimental results show that CALMS outperforms the monolingual baseline significantly. Further promotion will be gained by finetuning on the specific language. We also transfer our model to 7 languages (Hindi, Spanish, Indonesian, Turkish, Vietnamese, Ukrainian, Portuguese) and achieve great improvements, which indicates our model obtains a better initialization for summarization and can be a better solution for low-resource summarization. We additionally propose a new large-scale multilingual summarization dataset with 12 languages for future multilingual summarization research.

We highlight our contributions as follows:

(1) We design a neural model with the contrastive aligned learning strategy for multilingual summarization (CALMS), which improves summarization performance in both rich-resource and low-resource languages.

(2) We propose two new training strategies to distinguish important information from the document and align sentence-level information across languages.

(3) In order to investigate multilingual summarization, we create a 1.1 million multilingual summarization dataset MLGSum with 12 languages. The experimental results on 5 main languages show that our model significantly outperforms the monolingual summarization model. The extensive experiments on 7 other languages indicate our model can transfer to other similar languages with a good performance.

2 Related Work

Multilingual Summarization Abstractive summarization aims at generating a shorter version of the document while maintaining the most important information. With the large success brought by pre-trained language models in English abstractive summarization (Liu and Lapata, 2019; Lewis et al., 2020b; Zhang et al., 2020), several works focus on summarization in multiple languages. Nguyen and Daumé III (2019) constructs a small cross-lingual dataset with English summaries for non-English articles, and Scialom et al. (2020) proposes MLSUM with 5 languages as the extended version of English summarization dataset CNN/DailyMail (Hermann et al., 2015). Cao et al. (2020) use a Transformer-based model with 6 layers encoder and decoder to combine auto-encoder training, translation and summarization. Different from Cao et al. (2020), we focus on document-level multilingual summarization, which means understanding of long input in different languages is more important for our model. Besides, we propose a large-scale multilingual dataset with 12 languages and each document-summary pair is in the same language.

Contrastive learning in Summarization The goal of contrastive training is to let the model distinguish specific features by constructing positive and negative pairs. For summarization, it is often used to find a better summary. (Shi et al., 2019) randomly replaces a sentence in the ground-truth summary with a random sentence to form the negative sample. Wu et al. (2020) constructs negative samples on different aspects of summary qualities and propose a new summary evaluation method by contrastive learning. Zhong et al. (2020) use a pre-trained extractive model to select several candidates as negative samples and take the ground-truth as the positive. In this work, we dynamically sample several sentences from the document during the training phase and construct the positive and negative pair based on their similarity with the ground-truth summary.

Multilingual Pre-training for Generation Several works try to expand the successful unsupervised pre-training English language model to multiple languages for multilingual understanding and generation (Lample and Conneau, 2019; Huang et al., 2019; Liu et al., 2020; Xue et al., 2020). mBART (Liu et al., 2020) denoises full texts in multiple languages and pre-trains the complete encoder-decoder model, which works well on both sentence-level and document-level machine translation. mT5 (Xue et al., 2020) is the multilingual version of T5 (Kale and Rastogi, 2020) for text-to-text. MARGE (Lewis et al., 2020a) is trained with the multi-lingual multi-document paraphrasing objective, which reconstructs text in one language by retrieving a set of related texts in other languages.

3 Method

Given a document $D = \{x_1, x_2, \cdots, x_M\}$ with M words, the goal of abstractive summarization is to generate a summary with N words $Y = \{y_1, y_2, \cdots, y_N\}$, where $M > N$. For multilin-

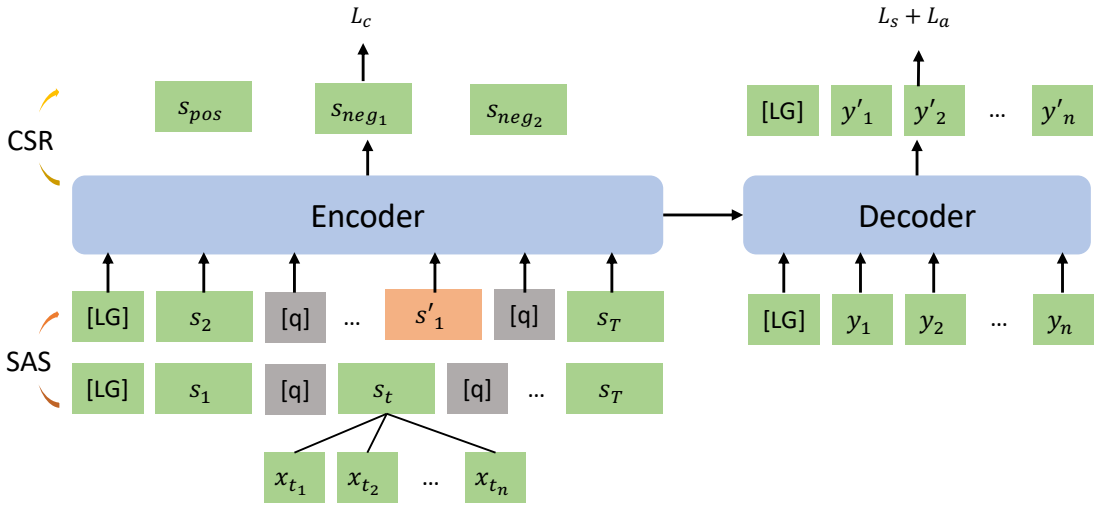

Figure 1: Model Overview. The input and output in green are the same language while the orange is another language. The input document includes T sentences separated by the delimiter '[q]', and each sentence s_t consist of t_n token x. The language indicator '[LG]' is added at the beginning of the encoder and the decoder. CSR selects $q = 3$ sentences from the document and constructs the contrastive pair with one positive and two negative examples. SAS replaces sentence s_1 with the translated s'_1.

gual summarization, the model should be able to deal with inputs in multiple languages and generate the summary in the same language. Formally, for each language l_k in the collection with K languages $L = \{l_1, l_2, \cdots, l_K\}$, the training objective can be defined as

$$\mathcal{L}^{(l_k)} = -\sum_{i=1}^{N_k} logP(Y_i^{(l_k)}|D_i^{(l_k)}), \qquad (1)$$

where $D_i^{(l_k)}$ and $Y_i^{(l_k)}$ are the i-th sample for the language l_k and N_k is the size of examples in l_k.

In this section, we propose a contrastive aligned joint learning strategy for all languages to share the salient information extraction and align sentence-level representations across languages. We propose two extra training objectives for our CALMS and describe them in detail below.

3.1 Multilingual Summarization

To understand and generate text in multiple languages, it is important to have a good multilingual language model. Without loss of generality, we use mBART (Liu et al., 2020) as the model initialization. It is a powerful Transformer-based multilingual pre-trained model trained on monolingual document corpus in 25 languages with denoising training objectives. It provides a shared vocabulary across languages and a good multilingual language model. We fully share model parameters among different languages by jointly training on all summarization data in different languages. A language indicator is used to indicate the language of each example. Thus, the multilingual summarization loss for K languages is written as:

$$\mathcal{L}_s = -\sum_{k=1}^{K}\sum_{i=1}^{N_k} logP(Y_i^{(l_k)}|D_i^{(l_k)}), \qquad (2)$$

3.2 Contrastive Sentence Ranking

Different from pre-trained denoising tasks, the output is much shorter than the input in the summarization task. Therefore, it is important for the summarization model to catch the salient information from the document during the finetuning phrase. We design a contrastive training strategy, contrastive sentence ranking (CSR), to help the model distinguish salient information, which is independent of languages. Inspired by content selection in extractive summarization (Shi et al., 2019; Zhong et al., 2020), we take sentences to construct positive and negative pairs. However, instead of pre-constructing contrastive summaries pairs for the dataset, we dynamically sample sentences from the document during the training phase.

Specifically, for a document D with T sentences $D = \{s_1, s_2, \cdots, s_T\}$, we randomly sample q sentences as candidates and calculate n-gram overlaps between the ground-truth summaries and these candidates. The candidate with the highest overlaps will be viewed as positive and the others are negative. By dynamically sampling, the model is able to explore the whole document. Besides, we can

change the negative sample number for each language to alleviate the imbalance between the data. Each time the data loader takes an example from the dataset, it will construct a positive-negative pair and save the corresponding sentence masks. These masks will be used to get sentence representation from the document's hidden state in the last layer of the encoder.

The model is trained with margin-based triplet loss, which is defined as:

$$\mathcal{L}_c = \frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{N_k} max\left(0, \sum_{j=1}^{q-1} s_{i,neg_j}^{(l_k)} - s_{i,pos}^{(l_k)} + \epsilon\right), \quad (3)$$

where $s_{i,pos}^{(l_k)}$ is the score of the positive candidate of the i-th example in language l_k, and $s_{i,neg_j}^{(l_k)}$ is the j-th negative candidate for i-th example. We use a linear layer with sigmoid function to get the score from the masked hidden state of the last layer of the encoder. ϵ is a hyper-parameter for the margin distance.

3.3 Sentence Aligned Substitution

Training with multiple languages makes it possible to share the representative space across languages and obtain a universal representation for summarization. Lin et al. (2020) randomly replaces words with a different language during the pre-training phase for machine translation. However, the input for summarization is longer than sentence-level machine translation and the single word replacement shows little influence (Kedzie et al., 2018). Thus, we propose sentence aligned substitution (SAS) for summarization.

We take lead sentences rather than randomly sampling from the document because these sentences are more important in the summarization task. We use an extra translation tool [1] to translate our sentences into another language to get the aligned information. To get rid of the lead bias, we randomly insert the translated sentences back into the original document. The training objective can be defined as:

$$\mathcal{L}_a = -\sum_{k=1}^{K}\sum_{i=1}^{N_k} logP\left(Y_i^{(l_k)} | \mathcal{R}\left(D_i^{(l_k)}\right)\right), \quad (4)$$

where $\mathcal{R}$ is the sentence replacement function. For the document in language l_k, its lead sentences are replaced with the rest languages $l_{k'}$ in ratio r.

[1] https://translate.google.com/

Finally, The training objectives of CALMS can be written as:

$$\mathcal{L} = \mathcal{L}_s + \mathcal{L}_c + \mathcal{L}_a. \quad (5)$$

Figure 1 demonstrates the overview of our model. CSR takes the output of the encoder for its margin loss, while SAS replaces sentences before encoding.

4 Experiment

In this section, we describe the multilingual summarization dataset used in our experiment and the experimental settings.

4.1 Dataset

We construct a large-scale summarization dataset MLGSum with 12 languages for the multilingual summarization task. We collect articles from news websites with multiple languages, such as BBC[2] and france24[3], and select faz[4] to extend our dataset with German text. We take the brief introduction written by editors as summaries[5]. We illustrate a short French example in Table 1.

Based on the language size, we divide MLGSum into two parts: the first part includes five high-resource languages: German(De), English(En), Russian(Ru), French(Fr), and Chinese(Zh), which will be used to train our CALMS. The second part has limited training data, which includes Hindi(Hi), Spanish(Es), Indonesian(Id), Turkish(Tr), Vietnamese(Vi), Ukrainian(Uk), and Portuguese(Pt). The data of each language is split into train/dev/test by 95%/5%/5%. Compared with multilingual Gigaword used by Cao et al. (2020), whose average document/summary length is 33.1/8.6, our document and summary are longer. This asks for document-level understanding and generation. The detailed information is listed in Table 2.

4.2 Settings

We use mBART (Liu et al., 2020) as the multilingual initialization. It is the multilingual version of BART-large (Lewis et al., 2020b), which is

[2] https://www.bbc.com/

[3] https://www.france24.com/

[4] https://www.faz.net/

[5] The summary is tagged with 'story-body__introduction' in BBC, 't-content__chapo' in france24, 'atc-IntroText' in faz. The data is from Jan, 2010 to Sep, 2020. We remove document length smaller than 50 or longer than 5000 and summaries shorter than five words. We provide the url of HTML page for each example.

Article Arsenal, le leader de la Premier League, a été sévèrement corrigé, samedi 14 décembre, par Manchester City (6-3) qui prend la deuxième place du classement à seulement trois points des hommes d'Arsène Wenger. [q] Il s'agit de la huitième victoire de City à domicile où il est invaincu cette saison, et ce contre la meilleure équipe à l'extérieur. [q] Les Londoniens ont commencé à prendre l'eau dès l'entame du match, Sergio Agüero ayant besoin de 14 minutes seulement pour ouvrir la marque et inscrire son 13e but de la saison en championnat.
(*Premier League leaders Arsenal were severely corrected on Saturday 14 December by Manchester City (6-3) who took second place in the standings just three points behind Arsène Wenger's men. [q] It was This is City's eighth home win where they are undefeated this season, against the best away team. [q] Londoners started to get wet from the start of the match, with Sergio Agüero needing just 14 minutes to open the scoring and score his 13th league goal of the season.*)

Summary Irrésistible à domicile depuis le début de la saison, Manchester City a étrillé Arsenal (6-3) lors du match au sommet de la Premier League. [q] Les Mancuniens reviennent à trois points des Gunners en haut du classement.
(*Irresistible at home since the start of the season, Manchester City crushed Arsenal (6-3) in the game at the top of the Premier League. [q] The Mancuniens are three points behind the Gunners at the top of the standings.*)

Table 1: A Fr example of our dataset. The text in brackets is the corresponding English translation. The sentences are separated by '[q]'.

Language	Size	Doc.	Summ.	Train
De	494,514	457	27	445,062
En	191,365	476	24	172,228
Ru	87,125	499	24	78,412
Fr	85,030	463	36	76,527
Zh	65,203	799	56	58,682
Hi	59,145	565	28	53,230
Es	43,162	703	30	38,845
Id	35,495	360	21	31,945
Tr	26,539	342	20	33,047
Vi	26,539	847	34	23,885
Uk	33,214	444	21	29,892
Pt	20,945	927	34	18,850
Total	1,168,276	573.5	29.6	1,060,605

Table 2: The dataset statistic. Doc. and Summ. refer to the average length of the document and the summary. Train is the size of the training set.For non-space language like Zh and Ja (with '*'), it is calculated by the character number.

a Transformer-based architecture (Vaswani et al., 2017) with 12 layers of encoder and 12 layers of the decoder. The hidden size is 1024 with 16 attention heads. mBART covers 25 languages and shares the vocabulary with the sentencepiece tokenizer (Kudo and Richardson, 2018), which includes 250,000 subword tokens. We follow the language indicators with mBART, and change its position to the beginning of the source and target sequence. We replace $[q]$ in the dataset with the delimiter $< /s >$ to separate sentences.

We use the first part of our dataset as training languages: De, En, Ru, Fr, Zh. We mix the training examples and do global shuffling to avoid local overfitting on a specific language. For CSR, we random sample $q = 3$ sentences from the document to construct the positive-negative pairs and let the margin $\epsilon = 1.0$. For SAS, we translate sentences to the other four languages with equal probability and substitute sentences with a ratio $r = 0.2$.

We use fairseq[6] (Ott et al., 2019) to implement the architecture. We limit the max tokens to 2048 for each GPU and set the gradient accumulation to 4. The Adam optimizer (Kingma and Ba, 2015) is used with a learning rate of 3e-5 for unified training and 1e-5 for finetuning on the specific language. The other parameters are the same as previous work (Liu et al., 2020). The joint training takes around 7 epochs and each epoch needs 5 hours on two 32G Tesla V100. During inference, we use trigram blocking to avoid repetition.

[6]https://github.com/pytorch/fairseq

4.3 Models

Here, we describe the models used in our experiments. We first introduce several baseline models and take the strong mBART monolingual model for each language as the main competitor for our unified multilingual summarization model.

Lead2 Lead-K is the common strong baseline for summarization tasks. We select the first two sentences based on the average summary length.

Monolingual Model We train a monolingual model for each language as our baseline. We use a standard Transformer with 12 layers of encoder and decoder with 1024 hidden states and 16 heads and randomly initialize it. The number of parameters is the same as the mBART. We use an independent vocabulary for each language and tokenize them with the sentencepiece model trained on the corresponding language corpus. For the mBART model, we follow the setting of Liu et al. (2020) to finetune it on the monolingual summarization task.

Model	Settings	De		En		Ru		Fr		Zh		Avg
		R-1	Delta	R-1	Delta	R-1	Delta	R-1	Delta	R-1	Delta	Delta
Mono	Lead2	26.35	-	22.54	-	17.21	-	**37.61**	-	29.74	-	-
	Transfromer	24.27	-	31.76	-	14.07	-	25.34	-	29.52	-	-
	mBART	25.92	-	38.89	-	21.52	-	35.75	-	38.25	-	-
Multi	mTransformer	23.91	-2.01	31.65	-7.24	15.07	-6.45	32.26	-3.49	31.65	-6.60	-5.16
	mBART	26.13	0.21	39.78	0.89	21.90	0.38	36.24	0.49	38.91	0.66	0.53
	CALMS	**26.38**	**0.46**	39.83	0.94	22.04	0.52	37.00	1.25	38.83	0.58	0.75
Finetune	mBART	26.01	0.09	39.87	0.98	21.57	0.05	36.02	0.27	38.93	0.68	0.41
	CALMS	26.33	0.41	**39.88**	**0.99**	**22.21**	**0.69**	36.88	1.13	**39.02**	**0.77**	**0.80**

Table 3: The main results. R-1 is the F1 score of ROUGE-1, and Delta is the difference between models and monolingual model initialized with mBART. Avg is the average delta for five languages. The best results are bolded.

Multilingual Model We jointly training summarization in five languages. For Transformer, we use the same shared vocabulary with mBART. We directly finetune mBART on the multilingual summarization task with the language indicator. For CALMS, we add the training objectives CSR and SAS, and the loss is defined as Equal 5. After jointly training, we directly evaluate the unified model on the test set of five languages.

Finetuning We finetune the unified mBART model and CALMS on the specific language for several steps and evaluate it on the test set. The training data for finetuning is the same as the jointly training phrase.

5 Results

We present the main quantitative results and design several qualitative analyses in this section. To better illustrate the improvement, we use the delta between different models and the strong baseline monolingual mBART in five languages for analysis.

For evaluation, we use the automatic summarization metric ROUGE(Lin, 2004)[7]. Since the original ROUGE is only designed for English, we map tokens in other languages to the digit and then calculate ROUGE. For the non-space language such as Chinese, we take each character as a token. We report the F-1 score of ROUGE-1 in the main paper and leave other scores in the appendix.

5.1 Main Results

In Table 3, we show our main results in five languages. We focus on the following questions: 1) Does a unified summarization for all languages perform better than the individual model for each language? 2) Does CALMS perform better on multilingual summarization compared with the unified mBART? 3) Does finetuning on the specific language benefit?

Monolingual v.s Multilingual For Transformer, the joint model performs worse on rich-resource De and En, while it gains improvement on Ru, Fr, and Zh. It indicates that the unified multilingual model without multilingual pre-training sacrifices the rich-resource languages and improve the low-resource languages. However, with the pre-training multilingual language model mBART, the unified model outperforms the monolingual ones on all five languages. This demonstrates that not only low-resource languages can benefit from the larger training data, but also high-resource languages can further be improved by multilingual joint training. Multilingual language models help the model to share the latent space across languages to some extend.

mBART v.s CALMS We directly evaluate the jointly training models on five languages in the test set. Compared with the unified mBART, our CALMS outperforms on all five languages, especially in Fr. For the average delta, CALMS outperforms the monolingual mBART by 0.75 ROUGE-1. The result demonstrates that CALMS is an effective and efficient solution for multilingual summarization. It can handle different languages with one unified model and improve performance on all languages without sacrificing rich-resource languages.

Does Finetuning benefit? Finetuning on CALMS makes the model further move on to the

[7]https://github.com/bheinzerling/pyrouge

specific language and get better results, such as En, Ru, and Zh. However, for De and Fr, it is better to directly evaluate the multilingual model, which indicates further finetuning may cause overfitting on several languages. It is similar to unified mBART, where the finetuning fails on De and Fr and benefits on En and Zh.

	De	En	Ru	Fr	Zh
CALMS	26.38	39.83	22.04	37.00	38.83
CALMS w/o CSR	26.24	39.73	22.01	36.95	38.89
CALMS w/o SAS	26.33	39.62	22.12	36.85	38.93
CALMS w/o pre-train	23.83	31.54	15.30	32.30	31.76

Table 4: Ablation study on CALMS on ROUGE-1. CALMS w/o CSR indicates removing CSR loss from CALMS. CALMS w/o mBART indicates randomly initialize the model and train with CSR and SAS.

Ablation Study We conduct the ablation study on each training strategy in Table 4. We jointly train each model and directly evaluate the test set without finetuning.

As it shown, both CSR and SAS contribute to our CALMS. Compared with CALMS w/o CSR and CALMS w/o SAS, we find De, Ru, Zh are more affected by removing CSR, while SAS is more important for En and Fr. When we remove mBART, the performance degrades significantly. This is because the multilingual pre-training language model not only provides a good initialization for multilingual representation but also have a strong generation ability as a language model, which has been proved in monolingual summarization with BART(Lewis et al., 2020b).

CALMS without pre-trained mBART can also be viewed as a jointly training mTransfromer with CSR and SAS. Compared with results in Table 3, we can find that the two training strategies improve performance in Ru, Fr and Zh, but the rich-resource languages De and En have been hurt. It implies that, without multilingual pretrained model, it is difficult for the multilingual model to recover from the denosing task SAS.

Transfer to other languages Does CALMS really help to learn a unified model for multilingual summarization? In order to answer this question, we further transfer the unified model to other languages. We finetune our CALMS trained on five languages to another 6 languages: Pt, Es, Uk, Tr, Vi, and Hi. Among them, Pt, Uk, and Id are not covered by the pre-training training phrase of mBART. We use '[UNK]' as the language indicator. For comparison, we also take the monolingual summarization model of each language as the baseline, which is similar to monolingual models described in 4.3. The results are listed in Table 5.

Family	Lang	Transformer	mBART	CALMS
Romance	Pt*	15.93	24.82	**25.89**
	Es	21.51	29.37	**29.77**
Slavic	Uk*	11.09	18.62	**19.23**
Turkic	Tr	13.45	**21.97**	21.68
Vietic	Vi	18.82	**30.88**	30.75
Indo-Aryan	Hi	25.53	**33.36**	32.98
Malayo-Polyn	Id*	18.61	27.17	**28.00**
Average	-	17.85	26.60	**26.90**

Table 5: Finetuning on CALMS trained on five languages. The family indicates language family and lang is the abbreviation of language. The Transformer and mBART are monolingual summarization model trained on each language. The languages with '*' are not covered by the pre-training corpus of mBART.

As the table shows, CALMS outperforms the monolingual Transformer and mBART in Pt, Es, Uk, and Id. Among these languages, Pt and Es is the same language family as Fr, while Uk and Ru both belong to Slavic. It indicates that our multilingual summarization model CALMS can help similar languages to get a better result against the monolingual model trained on its limited training data. For Id, it is not covered by the pre-training phase and our CALMS also shows better results on it. However, for other languages that far away from the training languages, CALMS has no obvious advantage over the monolingual model.

5.2 Analysis

In this sections, we conduct several in-depth explorations on the two training objectives CSR and SAS.

Negative Sample Number We explore how the candidate number q influences our model. Similar to above, we take the ROUGE-1 improvement against the mBART monolingual model to normalize the improvement. For the document with sentences fewer than q, we repeat the negative examples several times. After training the unified model,

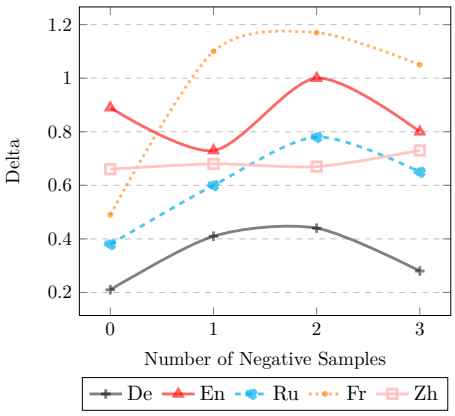

Figure 2: The negative example number of contrastive learning. 0 indicates the mBART unified model. The y-axis is the ROUGE-1 delta between our CALMS and the mBART monolingual model.

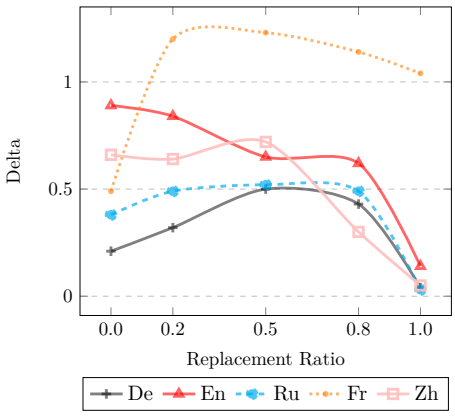

Figure 3: The replacement ratio of SAS. 0 indicates no replacement and 1 indicates each example will be subsituted. The y-axis is the ROUGE-1 delta between our CALMS and the mBART monolingual model.

we directly evaluate them without finetuning.

As Figure 2 shown, The x-axis is the negative sample number, which is $q-1$. When we take two negative examples for contrastive training, most languages get the best results. However, when it comes to three, the performance slips significantly. This is because it is more likely to construct the same contrastive pair during the dynamic sampling due to the limited length of the document. Compared with other languages, the negative sample number has little impact on Zh.

Replacement Ratio We also investigate different replacement ratios r as Figure 3 shown. When $r = 1.0$, it means that we always replace lead sentences with the other language. For $r = 0.0$, we do not replace any sentences, which is the jointly training mBART model. Same as above, we evaluation the unified model directly.

For En, with the ratio increases, the performance degrades, because SAS enforces the model to obtain a more unified representation for all languages by sacrificing the English bias. When the ratio is greater than 0.5, performance begins to degrade in all languages. The Delta is almost 0 when the ratio comes to 1. This indicates that the unified model no longer has the advantage over the individual model. In this case, all the lead sentences will be inserted into the document in different languages. It will mislead the model to ignore the lead bias and the learned language indicator. The ratio between 0.2 and 0.5 is appropriate for all five languages.

CSR for Individual Different from SAS which designed for aligning multiple languages, CSR aims at distinguishing important information. It can also be used on individual models. Thus, we add CSR to the mBART monolingual model for each language and set $q = 3$. The results are listed in Table 6.

	De	En	Ru	Fr	Zh
Individual	25.92	38.89	21.52	35.75	38.25
Individual + CSR	26.00	39.25	21.20	36.57	38.63

Table 6: CSA for individual model. The Individual model is trained on mBART for each language.

We find that De, En, Fr, and Zh all benefit from the original monolingual model, especially Fr. However, the performance degrades for Ru. From Figure 2, we can find that Ru is sensitive to the negative sample number, and Table 2 illustrates Ru have the longest article compared with De, En, and Fr (Zh is calculated by characters). Small q will lead to indistinguishable contrastive pairs during randomly sampling especially for long input, which will cause the performance decline.

6 Conclusion

We propose a contrastive aligned joint learning strategy CALMS. It is an effective and efficient solution for multilingual summarization that can handle different languages with one unified model. The experimental results show that CALMS outperforms the monolingual summarization model in all five training languages, and it can further transfer to similar languages and achieve improvement against monolingual mBART via finetuning. We also provide a multilingual summarization dataset MLGSum with 12 languages for future research.

Acknowledgements

We thank all the anonymous reviewers for their valuable suggestions. We thank the Internet Archive Projects[8] to share the archive's collections for research purposes.

Ethics Consideration

We collect the dataset from three news websites: BBC, france24, and faz. BBC provides news in more than 40 languages and each article is written by native authors. France24 is an international news website with 4 languages and faz is a German website. All of these websites have a highlight written by the editor at the beginning of the news article to summarize the main idea, which can be viewed as the summary. This information can be easily extracted through the HTML tag ('story-body__introduction' in BBC, 't-content__chapo' in france24, 'atc-IntroText' in faz). We collect MLGSum mainly from BBC and use france24 to expand French, English, and Spanish. Faz is used for German.

Similar to XSum (Narayan et al., 2018) and Newsroom (Grusky et al., 2018), we provide the Wayback archived URL of each article and the processing script to release MLGSum. The Wayback Machine[9] is an initiative of the Internet Archive, building a digital library of Internet sites that archive billions of web pages. We search news articles ranging from 2010 to 2020 for the above websites. We emphasize that the intellectual property and privacy rights of the articles belong to the original authors and the corresponding website. We carefully check the terms of use, privacy policy, and copyright policy[10] of the Internet Archive and the dataset construction is consistent with all terms.

We emphasize that we meet the usage requirements: "*Access to the Archive's Collections is provided at no cost to you and is granted for scholarship and research purposes only*" and "*abide by all applicable laws and regulations, including intellectual property laws, in connection with your use of the Archive*". We certify that our use of any part of the Archive's Collections will be limited to non-infringing or fair use under copyright law. If any authors or publishers express a desire for their documents not to be included in MLGSum, we will remove that portion from the dataset.

[8] https://archive.org/projects/

[9] http://web.archive.org/

[10] https://archive.org/about/terms.php

References

Yue Cao, Xiaojun Wan, Jinge Yao, and Dian Yu. 2020. Multisumm: Towards a unified model for multilingual abstractive summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11–18.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 708–719.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1684–1692.

Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. 2019. Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2485–2494.

Mihir Kale and Abhinav Rastogi. 2020. Text-to-text pre-training for data-to-text tasks. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 97–102.

Chris Kedzie, Kathleen McKeown, and Hal Daumé III. 2018. Content selection in deep learning models of summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1818–1828.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR (Poster)*.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*.

Mike Lewis, Marjan Ghazvininejad, Gargi Ghosh, Armen Aghajanyan, Sida Wang, and Luke Zettlemoyer. 2020a. Pre-training via paraphrasing. *Advances in Neural Information Processing Systems*, 33.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020b. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

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

Junyang Lin, Xu Sun, Shuming Ma, and Qi Su. 2018. Global encoding for abstractive summarization. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 163–169.

Zehui Lin, Xiao Pan, Mingxuan Wang, Xipeng Qiu, Jiangtao Feng, Hao Zhou, and Lei Li. 2020. Pre-training multilingual neural machine translation by leveraging alignment information. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2649–2663.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3721–3731, Hong Kong, China. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Khanh Nguyen and Hal Daumé III. 2019. Global voices: Crossing borders in automatic news summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 90–97.

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

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. Mlsum: The multilingual summarization corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8051–8067.

Jiaxin Shi, Chen Liang, Lei Hou, Juanzi Li, Zhiyuan Liu, and Hanwang Zhang. 2019. Deepchannel: Salience estimation by contrastive learning for extractive document summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6999–7006.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Danqing Wang, Pengfei Liu, Yining Zheng, Xipeng Qiu, and Xuan-Jing Huang. 2020. Heterogeneous graph neural networks for extractive document summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6209–6219.

Li Wang, Junlin Yao, Yunzhe Tao, Li Zhong, Wei Liu, and Qiang Du. 2018. A reinforced topic-aware convolutional sequence-to-sequence model for abstractive text summarization. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 4453–4460.

Hanlu Wu, Tengfei Ma, Lingfei Wu, Tariro Manyumwa, and Shouling Ji. 2020. Unsupervised reference-free summary quality evaluation via contrastive learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3612–3621.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuan-Jing Huang. 2020. Extractive summarization as text matching. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6197–6208.

A Appendices

We present ROUGE-2 and ROUGE-L in Table 7 and Table 8 for models in Table 3.

Different from ROUGE-1, monolingual models show an advantage over multilingual models on ROUGE-2 and ROUGE-L for De, which indicates that the multilingual models have difficulty in catching long patterns of German. However, the situation is the opposite for the French. The other trends are similar with analysis in Section 5.1.

Model	Settings	De		En		Ru		Fr		Zh		Avg
		R-2	Delta	R-2	Delta	R-2	Delta	R-2	Delta	R-2	Delta	Delta
Mono	Lead2	9.87	-	4.27	-	2.39	-	16.05	-	12.10	-	-
	Transfromer	12.30	-	9.89	-	4.16	-	7.09	-	14.51	-	-
	mBART	12.57	-	15.65	-	8.38	-	17.21	-	21.22	-	-
Multi	mTransformer	8.49	-4.08	9.67	-5.98	4.08	-4.30	11.8	-5.41	15.26	-5.96	-5.15
	mBART	11.75	-0.82	16.06	0.41	8.57	0.19	17.25	0.04	21.78	0.56	0.08
	CALMS	**11.94**	**-0.63**	16.18	0.53	8.67	0.29	**17.29**	**0.08**	21.68	0.46	0.15
Finetune	mBART	11.64	-0.93	15.39	-0.26	8.41	0.03	17.03	-0.18	21.78	0.56	-0.16
	CALMS	11.90	-0.67	**16.36**	**0.71**	**8.82**	**0.44**	17.25	0.04	**21.85**	**0.63**	**0.23**

Table 7: The main results of R-2, which is the F1 score of ROUGE-2. Delta is the difference between models and monolingual model initialized with mBART. Avg is the average delta for five languages. The best results are bolded.

Model	Settings	De		En		Ru		Fr		Zh		Avg
		R-2	Delta	R-2	Delta	R-2	Delta	R-2	Delta	R-2	Delta	Delta
Mono	Lead2	24.18	-	17.05	-	14.81	-	30.52	-	22.75	-	-
	Transfromer	22.55	-	24.1	-	12.63	-	20.91	-	24.76	-	-
	mBART	23.18	-	29.98	-	19.18	-	29.50	-	31.86	-	-
Multi	mTransformer	20.33	-2.85	23.51	-6.47	13.06	-6.12	24.85	-4.65	25.30	-6.56	-5.33
	mBART	22.80	-0.38	30.51	0.53	19.22	0.04	29.48	-0.02	31.86	0.00	0.03
	CALMS	**22.91**	**-0.27**	30.62	0.64	19.35	0.17	29.63	0.13	31.83	-0.03	0.13
Finetune	mBART	22.70	-0.48	30.28	0.30	19.01	-0.17	29.31	-0.19	31.91	0.05	-0.10
	CALMS	22.87	-0.31	**30.66**	**0.68**	**19.51**	**0.33**	**29.65**	**0.15**	**32.12**	**0.26**	**0.22**

Table 8: The main results of R-L (ROUGE-L). The other notations are the same with Table 7