ACROSS: An Alignment-based Framework for Low-Resource Many-to-One Cross-Lingual Summarization

Peiyao Li[1] **Zhengkun Zhang**[1] **Jun Wang**[2] **Liang Li**[3] **Adam Jatowt**[4] **Zhenglu Yang**[1]*
[1]TKLNDST, CS, Nankai University, China
[2]Shandong Key Laboratory of Language Resource Development and Application,
College of Mathematics and Statistics Science, Ludong University
[3]Nayuan Technology Co., Ltd. [4]University of Innsbruck, Austria
{peiyaoli, zhangzk2017, junwang}@mail.nankai.edu.cn
leo.li@nayuan.net, adam.jatowt@uibk.ac.at, yangzl@nankai.edu.cn

Abstract

This research addresses the challenges of Cross-Lingual Summarization (CLS) in low-resource scenarios and over imbalanced multilingual data. Existing CLS studies mostly resort to pipeline frameworks or multi-task methods in bilingual settings. However, they ignore the data imbalance in multilingual scenarios and do not utilize the high-resource monolingual summarization data. In this paper, we propose the **A**ligned **CROS**s-lingual **S**ummarization (ACROSS) model to tackle these issues. Our framework aligns low-resource cross-lingual data with high-resource monolingual data via contrastive and consistency loss, which help enrich low-resource information for high-quality summaries. In addition, we introduce a data augmentation method that can select informative monolingual sentences, which facilitates a deep exploration of high-resource information and introduce new information for low-resource languages. Experiments[1] on the CrossSum dataset show that ACROSS outperforms baseline models and obtains consistently dominant performance on 45 language pairs.

1 Introduction

Given a source document, Cross-Lingual Summarization (CLS) aims to generate a summary in a different language. Therefore, CLS helps users quickly understand news outlines written in foreign, unknown to them, languages. Early CLS approaches typically use pipeline frameworks (Leuski et al., 2003; Orasan and Chiorean, 2008), which are intuitive but suffer from the problem of error cascading. Researchers have recently turned to end-to-end models (Zhu et al., 2019, 2020; Bai et al., 2021) that are immune to this problem. However, these studies are limited to bilingual learning and do not conform to the reality of multilingual scenarios.

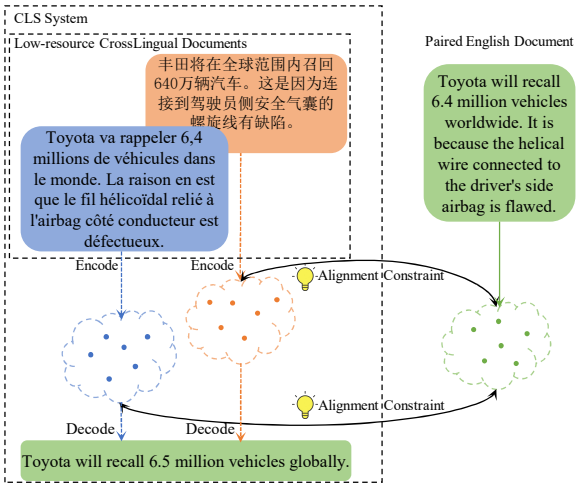


Figure 1: The schematic diagram of ACROSS. We try to build a strong alignment relationship between cross-lingual inputs and the corresponding monolingual input. We select English as the target language. We constrain French and Chinese documents to have the same representation as paired English document in the learning process. Finally, the CLS system can give the target English summary independently.

Given that the real-world news is written in diverse languages and that only a few researchers have explored the multilingual scenarios, we investigate the many-to-one CLS scenario to meet realistic demands. As stated before, CLS data can be viewed as low-resource since parallel CLS data is significantly less abundant than monolingual data (Zhu et al., 2019). The low-resource characteristic of CLS data is further amplified in multilingual scenarios. However, directly training an end-to-end model does not perform well due to the ineffective use of high-resource data and the scarcity of low-resource data. The foremost challenges are how to model cross-lingual semantic correlations in multilingual scenarios and introduce new knowledge to low-resource languages.

To tackle the above challenges, we investigate a novel yet intuitive idea of cross-lingual alignment.

*Corresponding author.
[1]https://github.com/Youggls/ACROSS-ACL23

Findings of the Association for Computational Linguistics: ACL 2023, pages 2458–2472
July 9-14, 2023

The cross-lingual alignment method can extract deep semantic relations across languages. As portrayed in Figure 1, the materials in three languages (i.e., French, Chinese, and English) express similar semantics. We can align all these languages for deep cross-lingual semantic knowledge, which is crucial for refining crosslingual materials over different languages for generating high-quality summaries. Moreover, we also consider devising a novel data augmentation (DA) method to introduce new knowledge to low-resource languages.

To investigate the two hypotheses, we introduce a novel many-to-one CLS model for low-resource learning called **A**ligned **CROS**s-lingual **S**ummarization (ACROSS), which improves the performance of low-resource scenarios by effectively utilizing the abundant high-resource data. This model conducts cross-lingual alignments both at the model and at the data levels. From the model perspective, we minimize the difference between the cross-lingual and monolingual representations via contrastive and consistency learning (He et al., 2020; Pan et al., 2021; Li et al., 2021, 2022). This helps to facilitate a solid alignment relationship between low-resource and high-resource language. From the data perspective, we propose a novel data augmentation method that selects informative sentences from monolingual summarization (MLS) pairs, which aims to introduce new knowledge for low-resource language.

We conducted experiments on the CrossSum dataset (Hasan et al., 2021), which contains cross-lingual summarization pairs in 45 languages. The results show that ACROSS outperforms the baseline models and achieves strong improvements in most language pairs (2.3 average improvement in ROUGE scores).

Our contributions are as follows:

- We propose a novel many-to-one summarization model that aligns cross-lingual and monolingual representations to enrich low-resource data.
- We introduce a data augmentation method to extract high-resource knowledge which is later transferred and which facilitates low-resource learning.
- An extensive experimental evaluation validate the low-resource CLS performance of our model in both quantitative and qualitative ways.

2 Related Work

Early CLS research typically used pipeline methods, such as the translate-then-summarize (Leuski et al., 2003; Ouyang et al., 2019) or summarize-then-translate methods (Orasan and Chiorean, 2008; Wan et al., 2010; Yao et al., 2015; Zhang et al., 2016), which are sensitive to error cascading that causes their subpar performance.

Thanks to the development of the transformer-based methods (Vaswani et al., 2017), researchers introduced teacher-student frameworks (Shen et al., 2018; Duan et al., 2019) wherein the CLS task can be approached via an encoder-decoder model. Thereafter, the multi-task framework started to be popular in this field (Zhu et al., 2019, 2020; Bai et al., 2021). Recently, researchers have begun to investigate how to fuse translation and summarization tasks into a unified model to improve the performance on the CLS tasks (Liang et al., 2022; Takase and Okazaki, 2022; Bai et al., 2022; Nguyen and Luu, 2022; Jiang et al., 2022). For example, Bai et al. (2022) considered compression so that their model can handle both the CLS and translation tasks at different compression rates.

Focusing on multi-task learning, these multi-task studies attempt to improve CLS performance using machine translation (MT) and MLS tasks in bilingual settings. However, such approaches still establish implicit connections among languages and leave aside the information of high-resource data.

Hasan et al. (2021) recognized the limitations of the above-mentioned scenarios. They proposed a new dataset, CrossSum, in multilingual scenarios and introduced a method balancing the number of different language pairs in a batch, which could alleviate the uneven distribution of training samples and balance performance in different languages. However, deep semantical correlations across languages as well as abundant information from high-resource data have not been investigated.

In contrast to the aforementioned methods, ACROSS introduces cross-lingual alignment and a novel data augmentation method, which can improve low-resource performance from both model and data perspectives.

3 Aligned CROSs-lingual Summarization

In this section, we explain the details of ACROSS. ACROSS introduces alignment constraints at both

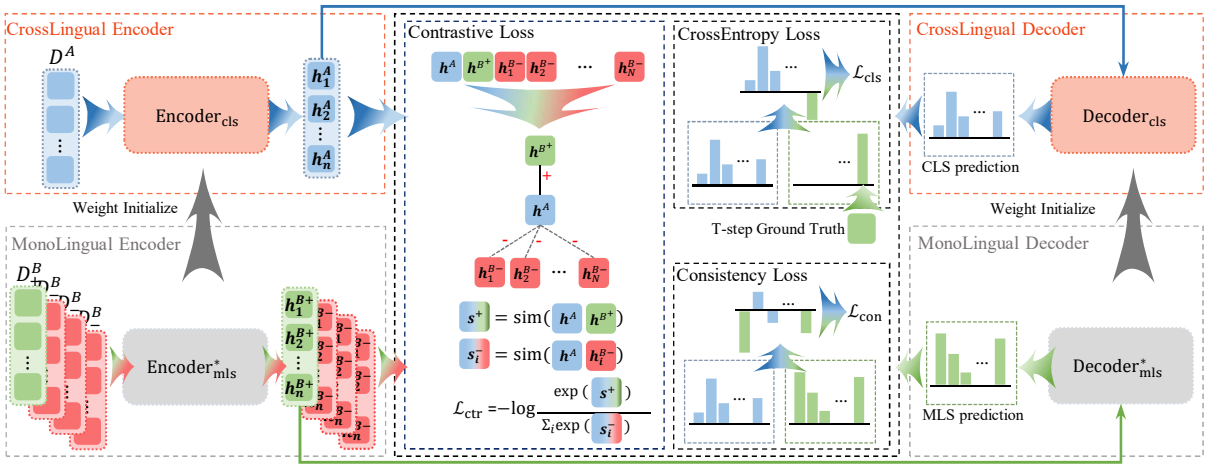


Figure 2: The framework of ACROSS. For CLS input D^A, the paired MLS input D^B_+ and the negative samples D^B_- are fed into the pretrained MLS model. ACROSS uses contrastive loss at the encoder side and consistency loss at the decoder to minimize the representations of the two languages.

the encoder and decoder sides. Figure 2 illustrates the overall framework.

3.1 Preliminary

Mono-Lingual Abstractive Summarization. Given a document $D^A = \{x_1^A, x_2^A, ..., x_n^A\}$ written in language A, a monolingual abstractive summarization model induces a summary $S^A = \{y_1^A, y_2^A, ..., y_m^A\}$ by minimizing the loss function as follows:

$$\mathcal{L}_{\text{abs}} = -\sum_{t=1}^{n} \log P(y_t^A | y_{<t}^A, D^A, \boldsymbol{\theta}_{\text{mls}}), \quad (1)$$

where n and m are the lengths of the input document and output summary, respectively, and $\boldsymbol{\theta}_{\text{mls}}$ is the parameter of the monolingual summarization model.

Cross-Lingual Abstractive Summarization. Different from monolingual abstractive summarization models, a cross-lingual abstractive summarization model generates a summary $S^B = \{y_1^B, y_2^B, ..., y_m^B\}$ in language B when given a source document $D^A = \{x_1^A, x_2^A, ..., x_n^A\}$ in language A. The loss function of the CLS model can be formulated as:

$$\mathcal{L}_{\text{cls}} = -\sum_{t=1}^{n} \log P(y_t^B | y_{<t}^B, D^A, \boldsymbol{\theta}_{\text{cls}}), \quad (2)$$

where $\boldsymbol{\theta}_{\text{cls}}$ is the parameter of the CLS model.

3.2 Cross-Lingual Alignment

Cross-Lingual Contrastive Learning for Encoder. Multilingual transformer treats all languages equally, which leads to the representation of different languages being distributed over different spaces, eventually making it difficult for CLS tasks to take advantage of the high-resource monolingual data. Therefore, we should encourage the model to improve cross-lingual performance with a strong monolingual summarization capability. With the help of contrastive learning, ACROSS can align the cross-lingual input representation to the monolingual space, thus realizing the idea mentioned above.

Firstly, given a cross-lingual summarization and the paired monolingual document tuple: (D^A, D^B_+, S^B), we need to randomly choose a negative document set $\mathcal{N} = \{D_1^B, D_2^B, ..., D_{|\mathcal{N}|}^B\}$ in the dataset. Then, we can obtain the representation of D^A with a Transformer Encoder and a pooling function $\mathcal{F}$ as follows:

$$\boldsymbol{h}^A = \mathcal{F}(\text{Encoder}_{\text{cls}}(D^A)). \quad (3)$$

Similarly, we can obtain the representation of D^B_+ with a pretrained Encoder of the monolingual summarization model as:

$$\boldsymbol{h}^B = \mathcal{F}(\text{Encoder}^*_{\text{mls}}(D^B)). \quad (4)$$

Finally, the contrastive learning objective is constructed to minimize the loss as follows:

$$\mathcal{L}_{\text{ctr}} = -\log \frac{e^{\text{sim}(\boldsymbol{h}^A, \boldsymbol{h}^B_+)/\tau}}{\sum_{i \in \text{idx}(\mathcal{N})} e^{\text{sim}(\boldsymbol{h}^A, \boldsymbol{h}^B_i)/\tau}}, \quad (5)$$

where τ is a temperature hyper-parameter and $\text{sim}(\cdot)$ denotes a similarity function that can measure the distance of two vectors in an embedding space[2].

[2]We use cosine similarity as the similarity function.

Cross-Lingual Consistency Learning for Decoder. Consistency learning aims to model consistency across the models' predictions, which can help child models gain improvement from the pre-trained parent model. By constraining the output probability distributions of decoders, the CLS child model can be aligned to the MLS pre-trained parent model.

Given a tuple composed of a CLS document and its paired monolingual document (D^A, D^B_+), we can obtain the output distribution of the CLS model at each decoding step as follows:

$$P(y^B_t|y^B_{<t}, D^A, \boldsymbol{\theta}_{\text{cls}}) = \text{Model}_{\text{cls}}(y^B_{<t}, D^A). \quad (6)$$

Similarly, we can construct the output distribution of the MLS model at each decoding step as:

$$P(y^B_t|y^B_{<t}, D^B_+, \boldsymbol{\theta}^*_{\text{mls}}) = \text{Model}^*_{\text{mls}}(y^B_{<t}, D^B_+), \quad (7)$$

where $\boldsymbol{\theta}^*_{\text{mls}}$ denotes frozen parameters and $\text{Model}^*_{\text{mls}}$ means that the parameters of the MLS model are frozen during training. Then, we can bridge the distribution gap between the CLS and MLS models by minimizing the following consistency loss function as:

$$\mathcal{L}_{\text{con}} = \sum_{t=1}^{n} \text{JS-Div}[P(y^B_t|y^B_{<t}, D^A, \boldsymbol{\theta}_{\text{cls}}), P(y^B_t|y^B_{<t}, D^B_+, \boldsymbol{\theta}^*_{\text{mls}})], \quad (8)$$

where JS-Div denotes Jensen–Shannon divergence (Lin, 1991), which is used to measure the gap between the pretrained and child models.

Training Objective of ACROSS. We jointly minimize CLS, consistency, and contrastive loss during the training period. The final training objective of ACROSS is formulated as:

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{\text{cls}} + \beta \cdot \mathcal{L}_{\text{ctr}} + \gamma \cdot \mathcal{L}_{\text{con}}, \quad (9)$$

where α, β, and γ are hyper-parameters used to balance the weights of the three losses.

3.3 Data Augmentation for Cross-Lingual Summarization

Data augmentation is a widely used technique in low-resource scenarios (Sennrich et al., 2016a; Fabbri et al., 2021). In Seq2Seq tasks, it often leverages translation to increase the amount of data in low-resource scenarios. However, in the CLS task, the direct translation of monolingual data from a high-resource language to a low-resource language might lose some valuable information. The distribution of information in the input document is uneven, making some sentences potentially more important than others. Therefore, directly translating all sentences into a low-resource language and using them for training the model may not be conducive to CLS.

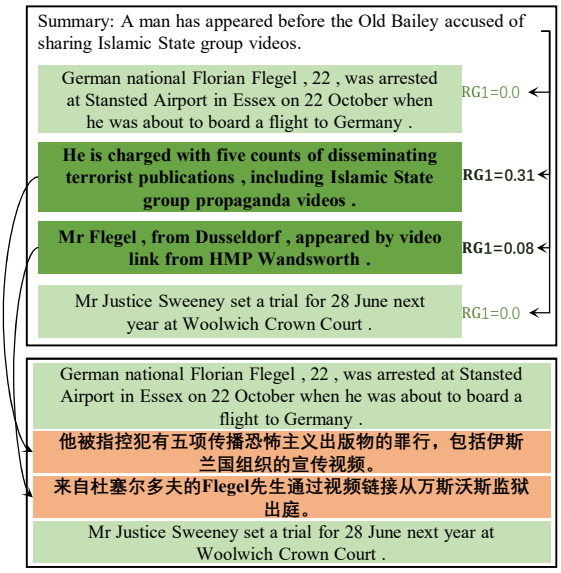


Figure 3: An example of our data augmentation method. This example shows the process from an English monolingual summarization pair to a Chinese-English summarization pair. Each green block contains an English sentence, while each orange block contains a Chinese sentence. The sentences corresponding to the dark green blocks have higher ROUGE scores, and these sentences will be translated into Chinese.

Considering the characteristics of the summarization task, we propose an importance-based data augmentation method based on ROUGE scores. First, an input document D^B_i is split into several sentences $\mathcal{S} = \{s_1, s_2, ..., s_k\}$. Then, the ROUGE score is calculated for each sentence and summary S^B. The ROUGE score of each sentence is represented as $\mathcal{R} = \{r_1, r_2, ..., r_k\}$. Next, the sentences corresponding to the top $a\%$ ROUGE scores are selected and translated into the low-resource language. Finally, the translated sentences are reassembled with other sentences to form a pseudo document $D^{A_p}_i$, so that pseudo-low-resource summarization pairs $(D^{A_p}_i, S^B)$ can be generated.

Figure 3 shows an example of the process from an English monolingual summarization pair to a Chinese-English summarization pair[3]. The two

[3]We set a=50.

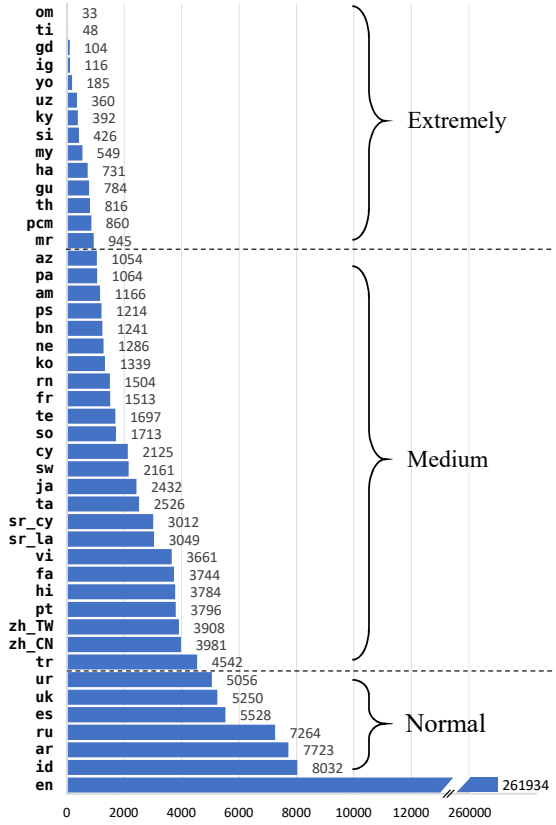


Figure 4: Distribution of the number of training samples over different languages to English. English-to-English summarization data accounts for more than 70% of the entire dataset.

sentences with the highest ROUGE scores are the second and third sentences; hence, these two sentences are translated into Chinese.

4 Experiment Setup

4.1 Dataset

We conduct experiments using the previously mentioned CrossSum dataset (Hasan et al., 2021). CrossSum is a multilingual CLS dataset that contains cross-lingual summarization data in 45 languages. Moreover, it realistically reflects the skewness of data distribution in practical CLS tasks. Figure 4 portrays the degree of imbalance of the dataset. As we can see, English monolingual summaries constitute over 70% of the English target summaries, while there are less than 30% summaries of other 44 languages to English. We classify languages with less than 1,000 training samples as extremely low-resource scenarios, between 1,000 and 5,000 as medium low-resource scenarios, and larger than 5,000 as normal low-resource scenarios.

4.2 Baselines

We compare our model with the following baselines:

Multistage: a training sampling strategy proposed by Hasan et al. (2021). Thi s method balances the number of different language pairs in a batch, thus alleviating the uneven distribution of training samples in different languages.

NCLS+MT: a method based on the multi-task framework proposed by Zhu et al. (2019). The model uses two independent decoders for CLS and MT tasks. As the original NCLS+MT model can only handle bilingual CLS task, we replace its encoder with a multilingual encoder.

NCLS+MLS: a method also proposed by Zhu et al. (2019). Its difference with NCLS+MT is that the multi-task decoder is used for MLS task.

4.3 Experimental Settings

For training, the MLS model is trained on the English-English subset of the CrossSum dataset and its parameters are initialized using mT5 (Xue et al., 2021). Thereafter, we initialize the CLS model using the pre-trained MLS model. We set dropout to 0.1 and the learning rate to 5e-4 with polynomial decay scheduling as well as a warm-up step of 5,000. For optimization, we use the Adam optimizer (Kingma and Ba, 2015) with ϵ = 1e-8, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay = 0.01. The hyper-parameters α, β, and γ are set to 1.0, 1.0, and 2.0, respectively. The size of the negative sample set is 1,024. The temperature hyper-parameter τ is set to 0.1. To stabilize the training process, we choose the gradient norm value to be 1.0. The vocabulary size is 250,112, and BPE (Sennrich et al., 2016b) is used as the tokenization strategy. We limit the max input length to 512 and the max summary length to 84. We train our model on 4 RTX A5000 GPUs for 40,000 training steps, setting the batch size to 256 for each step. For inference, we use the beam-search decoding strategy (Wiseman and Rush, 2016) and set the beam size to 5.

5 Experiment Results

5.1 Main Results

We evaluate ACROSS on the standard ROUGE metric (Lin, 2004), reporting the F1 score (%) of ROUGE-1, ROUGE-2, and ROUGE-L. Table 1 presents the main results of ACROSS and other

Model	Extremely			Medium			Normal			Overall		
	RG1	RG2	RGL	RG1	RG2	RGL	RG1	RG2	RGL	RG1	RG2	RGL
NCLS+MLS-small	24.46	6.21	18.97	25.76	7.09	20.06	28.57	8.52	22.20	25.78	7.07	20.05
NCLS+MT-small	25.69	7.15	20.19	26.68	7.49	20.74	29.17	8.90	22.68	26.75	7.64	20.86
Multistage-small	25.78	7.07	19.97	27.13	7.87	21.03	29.94	9.56	23.16	27.04	7.89	20.99
Multistage-base	28.00	8.51	21.97	30.10	9.90	23.36	33.16	11.94	25.84	29.90	9.82	22.34
ACROSS-small	28.20	8.43	22.06	29.34	8.99	22.64	31.94	10.58	24.82	29.24	9.01	22.70
ACROSS-base	**31.01**	**10.46**	**24.29**	**33.86**	**12.35**	**26.56**	**36.11**	**14.11**	**28.49**	**33.34**	**12.16**	**26.27**

Table 1: The main results of different models on CrossSum dataset (%). The bold values indicate the best results in *base* settings. The underlined values indicate the best results in *small* settings. ACROSS improves significantly in different low-resource settings and metrics.

models on different low-resource settings. *base* and *small* refer to different mT5 settings. *base* model contains a 12-layer encoder and 12-layer decoder with 768-dimensional hidden representations. *small* model contains an 8-layer encoder and 8-layer decoder with 512-dimensional hidden representations. As discussed in Section 4.1, we classify languages as extremely, medium and normal low-resource scenarios. It should be clarified that although we have artificially divided languages into different low-resource scenarios, any CLS language pair is actually low-resource compared to the English-English data volume.

Comparision with Multistage. Compared with the Multistage-base, ACROSS-base obtains 1.95, 2.45, and 2.17 ROUGE-2 improvements for extremely, medium and normal low-resource scenarios, respectively. Furthermore, ACROSS-base reaches 3.01, 3.76, and 2.95 ROUGE-1 improvements for extremely, medium and normal low-resource scenarios, respectively. The ROUGE-L scores for extremely, medium and normal low-resource scenarios are also improved by 2.32, 3.2, and 2.65, respectively. As shown in Figure 5, ACROSS-base outperforms Multistage-base significantly under the different language test sets. The ROUGE2 scores for more than 30 languages have an increase of more than 2, which represents a stable improvement of ACROSS.

Moreover, ACROSS-small surpasses or is compared to Multistage-base under some metrics (improving ROUGE-1 by 0.2 and ROUGE-L by 0.09 in extremely low-resource scenarios).

In addition, we can see in Figure 5 that English-English's ROUGE-2 score improves by only 0.77, which illustrates that the improvement of ACROSS comes mainly from the better alignment between other languages and English, rather than from

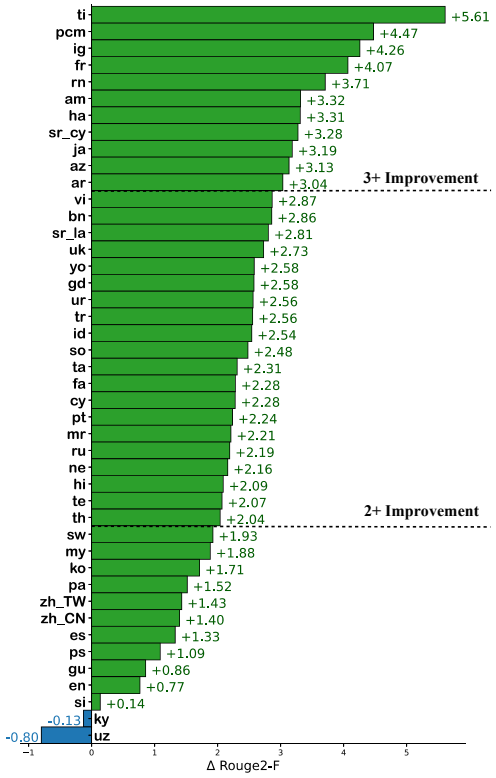


Figure 5: The improvement of ACROSS-base compared to mt5-base in ROUGE2-F on different languages to English (%).

the improvement of the ability to do summarization on English. Considering the actual data size, ACROSS significantly overperforms baselines in low-resource CLS scenarios. Additionally, we demonstrate ROUGE-1 and ROUGE-L results in Appendix A.

Comparision with Multi-Task Methods. Compared with the two multi-task methods (i.e., NCLS+MT and NCLS+MLS), we find that the two methods do not perform as well as Multistage and have a greater gap with ACROSS. Compared with NCLS+MT and NCLS+MLS, the ROUGE-

Model	RG1	RG2	RGL
Multistage	27.04	7.89	20.99
ctr+con+DA	**29.24**	**9.01**	**22.70**
con+DA	29.13	8.88	22.60
con	28.88	8.66	22.27
DA	27.63	8.28	21.51

Table 2: Ablation results (%). *ctr+con+DA* refers to the original ACROSS, *con+DA* represents ACROSS with removed contrastive loss, and *con* denotes ACROSS without contrastive loss and data augmentation. *DA* means that ACROSS uses only data augmentation during training without contrastive loss and consistency loss.

Model	RG1	RG2	RGL
Multistage	27.04	7.89	20.99
Random	28.36	8.62	22.26
Uninformative	28.24	8.56	22.09
Truncation	28.94	8.77	22.52

Table 3: Performance of different selection methods (%). *Informative* means selecting the sentences corresponding to the top 50% of ROUGE values. *Uninformative* is the opposite, and it selects the sentence with the lowest ROUGE value. *Random* denotes randomly selecting 50% of the sentences from the document. *Truncation* denotes that only the first half of the document is selected for translation.

Model	FL	IF	CC
Multistage	4.10	3.57	3.68
ACROSS	**4.43**	**3.96**	**4.04**

Table 4: Human Evaluation of ACROSS and Multistage on Chinese-English and French-English, the best results are in bold.

1, ROUGE-2, and ROUGE-L scores of ACROSS are enhanced by more than 3, 1, and 2, respectively. This phenomenon reveals that multi-task approaches that rely on MT and MLS learning may be not effective in multilingual scenarios. ACROSS turns to be more suitable for the scenarios with imbalanced resources.

5.2 Analysis

Ablation Study. We next conduct the ablation study in small settings. We summarize the experimental results in Table 2 as below:

- *ctr+con+DA* performs better than *con+DA*, suggesting that although con can significantly improve performance, the aligned representation is also beneficial for CLS tasks.
- The complete model produces better results compared with *DA*. Except for Multistage, *DA* performs worse than the models adding other losses, which implies that the excellent performance of ACROSS does not merely come from data augmentation.
- Comparing *DA* and *con*, we can see that the aligned model and representation are crucial for a successful CLS task.

Analysis of Data Augmentation. We conduct experiments on different selection approaches to evaluate the performance of our proposed DA method.

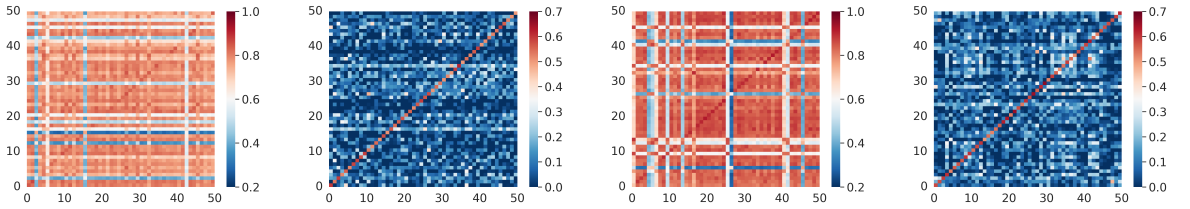
As recorded in Table 3, *Informative* performs best compared to the other methods, which indicates that the DA method can help ACROSS learn more important information in the CLS task. The *Truncation* performs inferior, because the more important sentences in the news report tend to be in the relatively front position. The results also validate the effectiveness of the DA method in selecting more important sentences for translation.

Generally speaking, the results tell us that the DA method is beneficial for the CLS task, and translating important sentences is useful for cross-lingual alignment.

Human Evaluation. Due to the difficulty of finding a large number of users who speak low-resource languages, we only conduct the human evaluation on 20 random samples from Chinese-English and French-English test sets. We compare the summaries generated by ACROSS with those generated by Multistage. We invite participants to compare the auto-generated summaries with ground truth summaries from three perspectives: fluency (FL), informativeness (IF), and conciseness (CC). Each sample is evaluated by three participants.

The results shown in Table 4 indicate that ACROSS is capable of generating fluent summaries and these summaries are also informative and concise according to the human feelings.

Visualization of Alignment. To further demonstrate the alignment result of ACROSS, we visualize the similarity between CLS inputs and the paired English inputs in Chinese-English and French-English test sets. We randomly sample 50 cross-lingual inputs from the test set and obtain the



(a) mT5, Chinese&English (b) ACROSS, Chinese&English (c) mT5, French&English (d) ACROSS, French&English

Figure 6: Visualization of Alignment Effect. The four figures are the heatmaps of different models and language pairs. The closer the color of the point is to dark red, the higher the similarity between the two corresponding inputs is. A clear diagonal line in the 6b and Figure 6d indicates that the paired inputs have a higher similarity. In contrast, Figure 6a and Figure 6c have many unexpected lines, meaning the model cannot distinguish the paired inputs from any other negative pairs.

representations of these cross-lingual inputs and the paired English inputs. Then, we calculate the cosine similarity of the two languages to construct the similarity matrix. Finally, we plot the heat map of the similarity matrix.

In Figure 6b and Figure 6d, the clear diagonal indicates the paired inputs have significantly higher similarities. In comparison, other unpaired inputs have lower similarities. In Figure 6a and Figure 6c, we can observe that the similarity distribution is characterized by more confusion.

In summary, ACROSS can effectively align cross-lingual and English inputs, demonstrating through the experiments that aligned representations are more useful for CLS tasks in multilingual settings.

Case Study. We finally implement the case study of a sample from the Chinese-English test set. The **Baseline** employed here is the Multistage model. The words and characters in red are important and overlap with **Ground Truth**. On the opposite, the words in green are errors. As shown in Figure 7, compared to Multistage, ACROSS can cover details in a better and more detailed way (e.g., using some proper nouns and phrases). For example, *asthma* and *processed meat* are present in the generated summary by ACROSS; yet, the summary generated by the baseline does not involve these important terms, and it also contains factual consistency errors. Taking another example, in the summary generated by the baseline, the terms *fruit and vegetables, including cabbage, broccoli, and kale* appear, while these terms are not mentioned in the original text.

The above examples suggest that ACROSS improves the performance of CLS based on the ability of strong MLS under the guidance of alignment.

Source: 70克大约是一根香肠再加一片火腿。 根据法国**研究人员**的调查发现, 如果一周**吃**四份以上的**加工肉食品**就会增加健康风险。但专家说, 两者之间的联系并没有得到证明, 需要做更多的调查。专家还建议，人们应该遵循一种更健康的饮食结构，例如每天吃的红肉和加工肉食品不要超过70克。参加这项试验的人中有一半是哮喘病人，然后观察他们的**哮喘症状**。试验显示，如果他们吃了过多的加工肉, 症状就会加重。 **Translation:** 70 grams is about one sausage plus one slice of ham. According to a survey by French **researchers**, **eating** more than four servings of **processed meat** a week increases health risks. But experts say the link between the two has not been proven and more investigation is needed. Experts also recommend that people follow a healthier diet, such as eating no more than 70 grams of red and processed meat per day. Half of the people who took part in the trial were asthmatics, and their **asthma symptoms** were then observed. Tests showed that if they ate too much processed meat, symptoms worsened.
ACROSS: **Eating** lots of **processed meat** could increase the risk of an **asthma** attack, according to **researchers**.
Baseline: A link between **eating** a lot of **fruit and vegetables, including cabbage, broccoli and kale**, has been suggested by French **researchers**.
Ground Truth: **Eating processed meat** might make **asthma** symptoms worse, say **researchers**.

Figure 7: Case study. The words in red are important and overlap with Ground Truth. The green words are errors.

6 Conclusion

In this work, we propose ACROSS, a many-to-one cross-lingual summarization model. Inspired by the alignment idea, we design contrastive and consistency loss for ACROSS. Experimental results show that with the ACROSS framework, CLS model improves the low-resource performance by effectively utilizing high-resource monolingual data. Our findings point to the importance of alignment in cross-lingual fields for future research. In

the future, we plan to apply this idea to combine CLS in multimodal scenarios, which might enable the model to better serve realistic demands.

Acknowledgements

This work was supported in part by National Natural Science Foundation of China under Grant No. 62106091, and in part by the Shandong Provincial Natural Science Foundation under Grant No. ZR2019MF062.

Limitations

Considering that English is the most widely spoken language, we select it as the high-resource monolingual language in this study. While ACROSS is a general summarization framework not limited to a certain target language, it deserves an in-depth exploration of how ACROSS works on other high-resource languages.

Additionally, we employ mT5 as our backbone because it supports most languages in CrossSum. The performance of ACROSS after replacing mT5 with other models, such as mBART(Liu et al., 2020), FLAN-T5(Chung et al., 2022), will be investigated in the future.

Ethical Consideration

Controversial Generation Content. Our model is less likely to generate controversial content(e.g., discrimination, criticism, and antagonism) since the model is trained on a dataset from the BBC News domain. Data in the news domain is often scrutinized before being published, and thus the model is not likely to generate controversial data.

Desensitization of User Data. We use the Amazon Mechanical Turk crowdsourcing platform to evaluate three artificial indicators (i.e., fluency, informativeness, and conciseness). For investigators, all sensitive user data is desensitized by the platform. Therefore, we also do not have access to sensitive user information.

References

Yu Bai, Yang Gao, and Heyan Huang. 2021. Cross-lingual abstractive summarization with limited parallel resources. In *Proceedings of ACL-IJCNLP*, pages 6910–6924.

Yu Bai, Heyan Huang, Kai Fan, Yang Gao, Yiming Zhu, Jiaao Zhan, Zewen Chi, and Boxing Chen. 2022. Unifying cross-lingual summarization and machine translation with compression rate. In *Proceedings of SIGIR*, pages 1087–1097.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Xiangyu Duan, Mingming Yin, Min Zhang, Boxing Chen, and Weihua Luo. 2019. Zero-shot cross-lingual abstractive sentence summarization through teaching generation and attention. In *Proceedings of ACL*, pages 3162–3172.

Alexander Fabbri, Simeng Han, Haoyuan Li, Haoran Li, Marjan Ghazvininejad, Shafiq Joty, Dragomir Radev, and Yashar Mehdad. 2021. Improving zero and few-shot abstractive summarization with intermediate fine-tuning and data augmentation. In *Proceedings of NAACL-HLT*, pages 704–717.

Tahmid Hasan, Abhik Bhattacharjee, Wasi Uddin Ahmad, Yuan-Fang Li, Yong-Bin Kang, and Rifat Shahriyar. 2021. Crosssum: Beyond english-centric cross-lingual abstractive text summarization for 1500+ language pairs. *CoRR*, abs/2112.08804.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of CVPR*, pages 9726–9735.

Shuyu Jiang, Dengbiao Tu, Xingshu Chen, Rui Tang, Wenxian Wang, and Haizhou Wang. 2022. Cluegraphsum: Let key clues guide the cross-lingual abstractive summarization. *CoRR*, abs/2203.02797.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of ICLR*.

Anton Leuski, Chin-Yew Lin, Liang Zhou, Ulrich Germann, Franz Josef Och, and Eduard H. Hovy. 2003. Cross-lingual c*st*rd: English access to hindi information. *ACM Trans. Asian Lang. Inf. Process.*, pages 245–269.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven Chu-Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. In *Proceedings of NeurIPS*, pages 9694–9705.

Zhaocong Li, Xuebo Liu, Derek F. Wong, Lidia S. Chao, and Min Zhang. 2022. Consisttl: Modeling consistency in transfer learning for low-resource neural machine translation. *CoRR*, arXiv/2212.04262.

Yunlong Liang, Fandong Meng, Chulun Zhou, Jinan Xu, Yufeng Chen, Jinsong Su, and Jie Zhou. 2022. A variational hierarchical model for neural cross-lingual summarization. In *Proceedings of ACL*, pages 2088–2099.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain.

J. Lin. 1991. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, pages 145–151.

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*, pages 726–742.

Thong Thanh Nguyen and Anh Tuan Luu. 2022. Improving neural cross-lingual abstractive summarization via employing optimal transport distance for knowledge distillation. In *Proceedings of AAAI*, pages 11103–11111.

Constantin Orasan and Oana Andreea Chiorean. 2008. Evaluation of a cross-lingual romanian-english multi-document summariser. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco*.

Jessica Ouyang, Boya Song, and Kathy McKeown. 2019. A robust abstractive system for cross-lingual summarization. In *Proceedings of NAACL-HLT*, pages 2025–2031.

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. 2021. Contrastive learning for many-to-many multilingual neural machine translation. In *Proceedings of ACL-IJCNLP*, pages 244–258.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of ACL*, pages 86–96.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of ACL*, pages 1715–1725.

Shiqi Shen, Yun Chen, Cheng Yang, Zhiyuan Liu, and Maosong Sun. 2018. Zero-shot cross-lingual neural headline generation. *IEEE ACM Trans. Audio Speech Lang. Process.*, pages 2319–2327.

Sho Takase and Naoaki Okazaki. 2022. Multi-task learning for cross-lingual abstractive summarization. In *Proceedings of LREC*, pages 3008–3016.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceddings of NIPS*.

Xiaojun Wan, Huiying Li, and Jianguo Xiao. 2010. Cross-language document summarization based on machine translation quality prediction. In *Proceedings of ACL*, pages 917–926.

Sam Wiseman and Alexander M. Rush. 2016. Sequence-to-sequence learning as beam-search optimization. In *Proceedings of EMNLP*, pages 1296–1306.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of NAACL-HLT*, pages 483–498.

Jin-ge Yao, Xiaojun Wan, and Jianguo Xiao. 2015. Phrase-based compressive cross-language summarization. In *Proceedings of EMNLP*, pages 118–127.

Jiajun Zhang, Yu Zhou, and Chengqing Zong. 2016. Abstractive cross-language summarization via translation model enhanced predicate argument structure fusing. *IEEE ACM Trans. Audio Speech Lang. Process.*, pages 1842–1853.

Junnan Zhu, Qian Wang, Yining Wang, Yu Zhou, Jiajun Zhang, Shaonan Wang, and Chengqing Zong. 2019. NCLS: neural cross-lingual summarization. In *Proceedings of EMNLP-IJCNLP*, pages 3052–3062.

Junnan Zhu, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2020. Attend, translate and summarize: An efficient method for neural cross-lingual summarization. In *Proceedings of ACL*, pages 1309–1321.

A Appendix

Analysis of Alignment Methods. To further show the effectiveness of ACROSS, we conduct an experiment to analyze the alignment methods. We replace the alignment methods of the encoder and decoder. As Table 5 shows, replacing any part of the original alignment methods will make the model perform worse. In particular, replacing the consistency and contrastive loss at the same time significantly reduces the model's performance, which reinforces the rationality of our different loss designs.

Model	RG1	RG2	RGL
ctr+con	**29.24**	**9.01**	**22.70**
ctr+ctr	26.58	8.28	21.23
con+con	28.43	8.89	22.32
con+ctr	26.23	8.01	21.18

Table 5: Effective of alignment method. *ctr+con* refers to the original ACROSS model framework. *con+ctr* denotes that the encoder uses consistency loss, and the decoder uses contrastive loss. *con+con* and *ctr+ctr* indicate the two alignment methods of the model using all consistency loss and all contrastive loss, respectively.

Data Augmentation Settings. We use Helsinki-NLP[4] as our translation model. In practice, we select the sentences corresponding to the top 50% of ROUGE scores. Furthermore, we set the beam size to 4, lenght-penalty to 1.0 and min-length to 10 for decoding.

ROUGE-1 & ROUGE-L Improvement for ACROSS-base. To show the improvement of our model on different metrics, we plot the improvement compared to Multistage-base, similar to Figure 5. As Figure 8 and 9 shows, ACROSS-base also has a significant and stable improvement on ROUGE-1 and ROUGE-L among different languages.

Analysis of Translation Ratio. We also analyze the impact of the translation ratio α on the final results. As table 6 shows, ACROSS-100% performs worse instead, probably because translating all sentences introduces too much extraneous noise instead.

Improvement in Different Resource Scenarios. We also analyze the improvement in different resource scenarios. As table 7 shows, low-resource languages get a more significant improvement compared to high-resource languages.

[4]https://huggingface.co/Helsinki-NLP

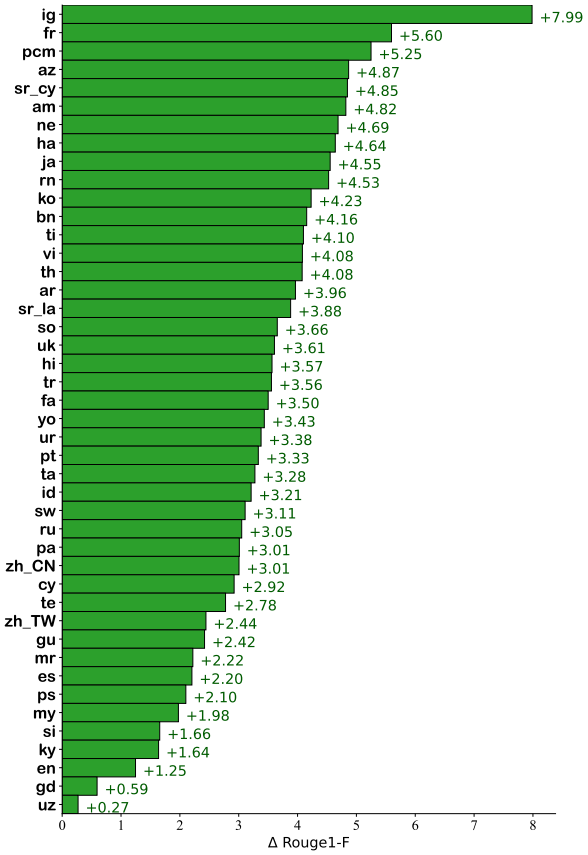


Figure 8: The improvement of ACROSS-base compared to Multistage-base in ROUGE1-F on different languages to English.

Form for Human Evaluation. Figure 10 shows the form we gave to participants, on the case of French-English summarization evaluation. Participants were asked to compare the auto-generated summaries with ground truth summaries from three perspectives: fluency, informativeness, and conciseness from one to five. And each participant will be informed that their scores for the different summaries will appear in our study as an evaluation metric.

Model	RG1	RG2	RGL
baseline	27.04	7.89	20.99
ACROSS-50%	**29.24**	**9.01**	**22.70**
ACROSS-100%	29.04	8.90	22.58

Table 6: Impact of translation ratio on the final result. ACROSS-50% denotes $\alpha = 50$, ACROSS-100% denotes $\alpha = 100$.

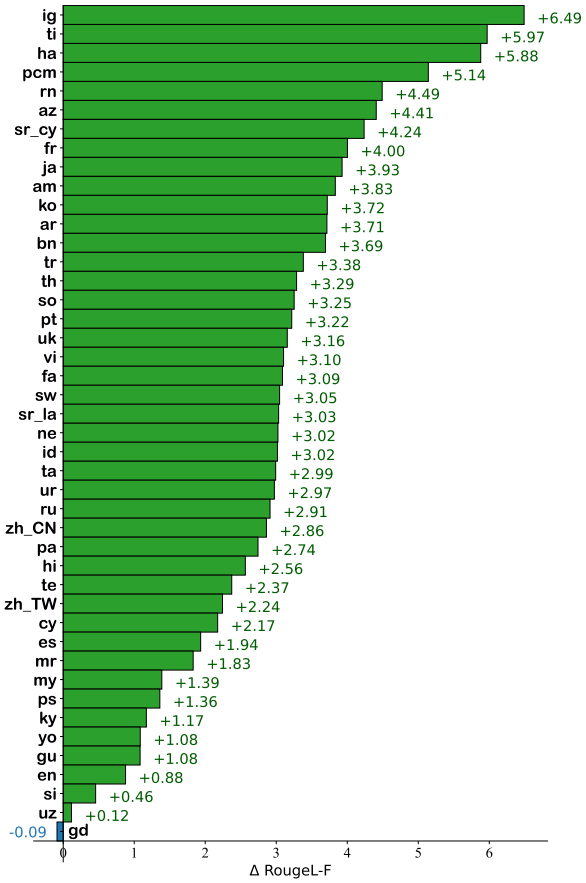


Figure 9: The improvement of ACROSS-base compared to Multistage-base in ROUGEL-F on different languages to English.

Model	Extremely	Medium	Normal
ACROSS-small	19.24%	14.23%	10.67%
ACROSS-base	22.91%	24.75%	18.17%

Table 7: Improvement of ROUGE-2 in different resource scenarios. All values in this table are percentages relative to the baseline.

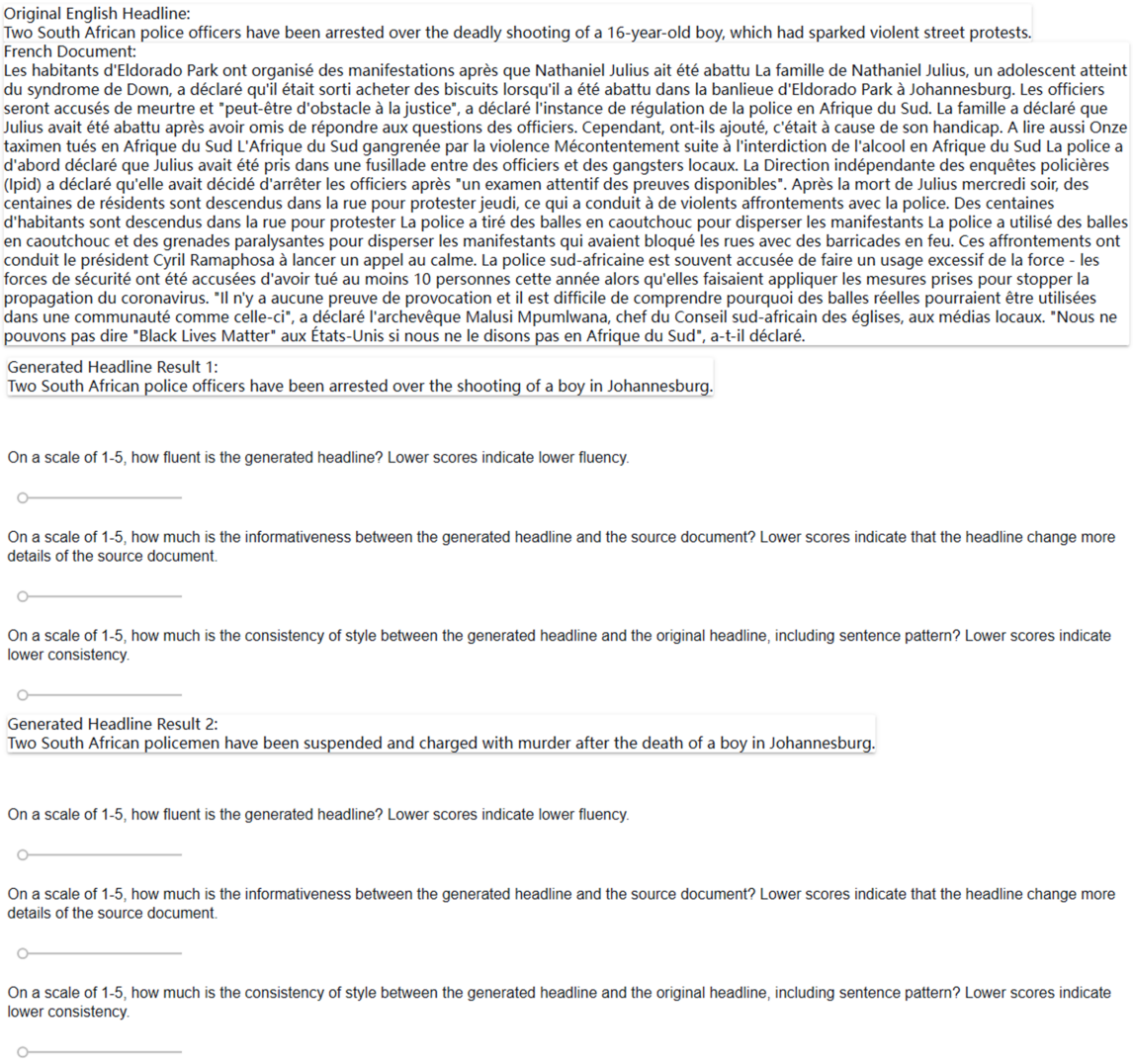


Figure 10: The form for human evaluation.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
7

☒ A2. Did you discuss any potential risks of your work?
We report a series of experimental setups in the paper, including model size, experimental data, performance on different languages, etc. Our proposed approach is also a generic generalization approach that takes into account universal scenarios.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Left blank.

☑ B1. Did you cite the creators of artifacts you used?
2

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
The dataset we use is released under license CC BY-NC-SA 4.0. The license is restricted only to those who want to modify and redistribute it, who need to use the same license as it.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
4

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
8

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
4

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
4

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

C ☑ Did you run computational experiments?

5

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
4

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
4

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
4

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
5

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

5

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
appendix a

☒ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
We used the platform provided by Amazon for human evaluation and charged 0.02$ per piece of data, which is also in line with the price of most text tasks.

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
appendix a

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.