Improving the Factual Consistency of Abstractive Summarization: Model Self-Improvement Contrastive Learning

Anonymous ACL submission

Abstract

Abstractive summarization models often produce summaries that are inconsistent with the content of the original text. Contrastive learning is an effective strategy for improving the factual consistency of abstractive summarization. However, the success of contrastive learning depends largely on the construction of the dataset. Existing contrastive learning methods usually directly consider the gold summaries of summary datasets as factual summaries, and ignore the factual consistency problem of the gold summaries. They mainly focus on the generation of hallucinated summaries, i.e., negative samples, and the construction for negative samples is usually based on the gold summaries rather than from the perspective of the model itself. The quality of the positive and negative samples of these methods is not high enough, which will affect the effect of contrastive learning. Therefore, this paper proposes Model Self-Improvement Contrastive Learning: a method to improve the factual consistency of abstractive summarization. This method begins with fine-tuning the model itself, considering its already acquired knowledge of generating summaries. It focuses on the inference aspect of the generation phase, and delves deeper into the content that may cause factual errors. At the same time, it takes into account the factual consistency of both the positive and negative samples, constructs the negative samples in a targeted manner and improves the positive samples. It then further improves the factual consistency of the model through contrastive learning.

1 Introduction

Abstractive summarization is a significant research direction in the field of natural language processing. Its goal is to extract key information from a given document and generate a concise, factual, and key information-containing summary. Despite the significant progress made by large pre-trained language models in abstractive summarization (Lewis et al., 2019; Devlin et al., 2018), existing summarization models often produce summaries inconsistent with the original text(Maynez et al., 2020; Cao et al., 2018). To address this issue, researchers have proposed numerous methods. One approach to fine-tune the model using training data with higher factual consistency, to aid the model in learning factuality (Goyal and Durrett, 2021; Wan and Bansal, 2022; Chaudhury et al., 2022). Another major research direction is to improve the structure or learning method of the model, enabling it to better understand and generate factual information (Cao and Wang, 2021; Chen et al., 2021; LIN and ZHOU, 2023; Li et al., 2023), where contrastive learning is an effective learning method to enhance factual consistency. Cao and Wang (2021) proposed a novel contrastive learning formula, designing deletion, replacement, rearrangement, and hallucination strategies to create negative samples based on reference summaries. The model is then trained to better distinguish these summaries using contrastive learning. Wan and Bansal (2022) generated negative samples by applying a series of rule-based transformations to the sentences in the source document, such as content replacement and sentence negation. Contrastive learning was introduced to help the model better distinguish between factual summaries and hallucinated summaries. These studies indicate that contrastive learning can help the model better distinguish between factual and non-factual information, effectively improving the model's factual consistency. The factual consistency of the training data is crucial for the model's factual consistency. However, existing contrastive learning methods directly take the gold summaries as positive samples, assuming they are factually accurate, and the construction of negative samples is usually based on transformations of the gold summaries.

In this work, we propose a Model Self-

Improvement Contrastive Learning method that combines data filtering and re-ranking methods to specifically enhance the quality of positive and negative samples, and then greatly improves the factual consistency of the summarization model by introducing contrastive learning. Our method takes into account both the factual consistency of positive and negative samples and whether the negative samples are consistent with the actual errors made by the model. We use data filtering methods to select summaries with higher factual consistency as positive samples, and construct negative samples from the perspective of the model itself. We delve into the inference stage of its generation phase based on the summarization knowledge already obtained by fine-tuning the model itself, and use beam search to dig deep into the content that may cause factual errors, and construct negative samples in a targeted manner. Finally, we reinforce the correct knowledge of the model through contrastive learning, thereby improving factual accuracy. The experimental results on the XSum dataset (Narayan et al., 2018) and the PEGASUS model (Zhang et al., 2020) show that our method effectively improves the factual consistency of abstractive summarization. In summary, our contributions are as follows:

(1) We have analyzed the characteristics and shortcomings of the existing contrastive learning methods in constructing samples, and proposed a Model Self-Improvement Contrastive Learning method. Starting from the perspective of the model itself, we specifically improved the quality of positive and negative samples.

(2) We have conducted an in-depth study on the characteristics and preferences of various cutting-edge factual consistency evaluation metrics in the context of data filtering, providing a new perspective for the use of data filtering methods.

(3) We have used multiple cutting-edge factual evaluation metrics to evaluate the summaries generated by our model, and the results showed that our method effectively improved the factual consistency of the model.

2 Related Work

2.1 Improvement and Evaluation of Factual Consistency

Current models can generate highly fluent summaries, but the generated summaries often contain factual errors (Maynez et al., 2020; Fang et al., 2020). In response to the issue of factual consistency, some research has been conducted, mainly divided into two categories: One category is to directly improve the factual consistency of the generated summary by improving the structure of the model or introducing additional information and constraints, enabling the model to generate summaries with higher factual consistency (Wan and Bansal, 2022; Chaudhury et al., 2022; Fang et al., 2020; Kryściński et al., 2019). The other category is to evaluate the factual consistency of the summary, by designing effective evaluation metrics or methods, enabling the model to better detect or correct factual errors (Kryściński et al., 2019; Laban et al., 2022), indirectly helping the model to enhance the factual consistency of the generated summary.

2.2 Enhancing the Factual Consistency of Training Data

Enhancing the factual consistency of training data to aid the model in generating summaries with higher factual accuracy is an important research direction. Some researchers have already explored in this area. Goyal and Durrett (2021) indicates that fine-grained human-annotated data can help train more factual summary models. Wan and Bansal (2022) trained a corrector module to remove hallucinations present in the gold summaries of the XSum training dataset, thereby improving the factual consistency of the training data. Chaudhury et al. (2022) demonstrated that filtering training data based on the scores of factual consistency evaluation metrics can assist in training more realistic summary models. They also trained a corrector module using gold summaries and hallucination summaries to correct factual errors in generated summaries.

2.3 Constrastive Learning

Constrastive learning has always been a popular method in representation learning, and in recent years, some researchers have applied it to natural language processing. Fang et al. (2020) Constrastive learning is used to train language models with stronger representation capabilities. Cao and Wang (2021) A novel formula for constrastive learning was proposed, which is based on reference summaries. Four types of strategies were designed to create negative samples: deletion, replacement, rearrangement, and hallucinations. Then, constrastive learning is used to train the summary model that can better distinguish them. Wan and

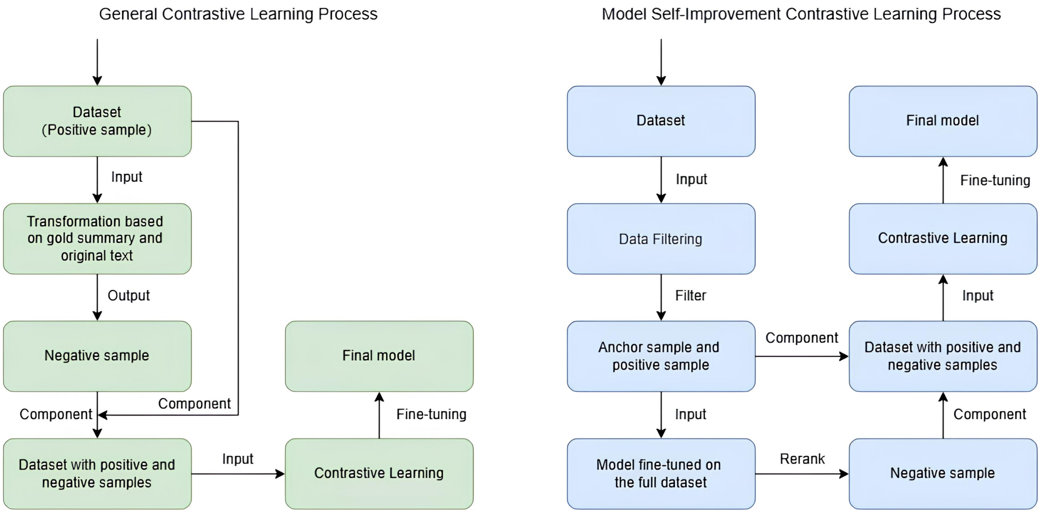
General Contrastive Learning Process
Dataset (Positive sample)
Input
Transformation based on gold summary and original text
Output
Negative sample
Component
Component
Dataset with positive and negative samples
Input
Contrastive Learning
Fine-tuning
Final model
Model Self-Improvement Contrastive Learning Process
Dataset
Input
Data Filtering
Filter
Anchor sample and positive sample
Input
Model fine-tuned on the full dataset
Rerank
Negative sample
Component
Dataset with positive and negative samples
Component
Input
Contrastive Learning
Fine-tuning
Final model

Figure 1: Flowchart of General Contrastive Learning and Model Self-Improvement Contrastive Learning.

Bansal (2022) By applying a series of rule-based transformations to sentences in the source document, such as content replacement and sentence negation, to generate negative samples, constrastive learning is introduced to help the model better distinguish between factual summaries and hallucination summaries.

3 Method

3.1 Data Filtering

In this study, we employed various cutting-edge factual consistency evaluation metrics to filter the dataset, aiming to select data with high factual consistency scores. To choose the best filtered data to construct positive samples for contrastive learning, we conducted an in-depth analysis of the characteristics and preferences when filtering data with different metrics. Subsequently, we fine-tuned the model using the data filtered out by different metrics. The fine-tuned model was then used to predict the test dataset. Afterward, we evaluated the generated summaries using various factual consistency evaluation metrics to observe the effect of fine-tuning the model using data filtered with different metrics. In this way, we selected the best filtering metric.

3.2 Model Self-Improvement Contrastive Learning

In order to further enhance the factual consistency of the model based on data filtering, we introduce Model Self-Improvement Contrastive Learning. The fundamental idea of the Model Self-Improvement Contrastive Learning is to start from fine-tuning the model itself, taking into account the abstractive summarization knowledge it has already acquired, focusing on the inference link in the generation stage, and deeply exploring the content that may cause factual errors. At the same time, we considered the factual consistency of positive and negative samples, constructed negative samples in a targeted manner, and improved positive samples. We constructed a dataset containing the original text, factual summary, and hallucination summary using data filtering and re-ranking techniques, and then further improved the model's factual consistency through contrastive learning.Its process differs from general contrastive learning as shown in Figure 1. The key steps are as follows:

(1) Firstly, we acquire the positive samples. We select the best performing filtering metric and use it to filter out the data with high factual consistency scores. The original text of this data is used as the anchor sample, and the golden summary is used as the factual summary, i.e., the positive sample, ensuring the factual consistency of the positive sample.

(2) Then, we construct the negative samples. We use the model fine-tuned on the complete dataset to predict the filtered data. During the summary generation, we select ten candidate summaries using the Beam Search method. Then, we use the filtering metric to calculate the factual scores of these ten summaries, re-rank them, and select the summary with the lowest score as the hallucinated summary, i.e., the negative sample. Although this

sample is not the direct output of the model, it is generated by the actual knowledge learned by the model, which better reflects the types of factual errors that the model is prone to make when generating summaries. Thus, we have constructed a contrastive learning dataset that includes the original text, factual summary, and hallucinated summary. This dataset's factual summary has a high factual consistency, and the hallucinated summary is automatically generated by the fine-tuned summarization model, which better reflects the actual errors made by the model when generating summaries.

(3) Finally, we introduce contrastive learning. Through contrastive learning, we encourage the model to prefer factual summaries when given the context of a document, further enhancing the model's factual consistency. The contrastive loss function we use when fine-tuning the model is as follows:

$$loss_c = a \cdot loss_f - (1-a) \cdot loss_h \qquad (1)$$

Here, $loss_f$ is the loss of the factual summary, $loss_h$ is the loss of the hallucinated summary, and a is an adjustable parameter used to control the weights of the factual summary and hallucinated summary in the contrastive loss. The design of this contrastive loss function encourages the model to not only minimize the loss of the factual summary but also maximize the loss of the hallucinated summary during the training process. This encourages the model to generate summaries that are more in line with the facts and avoid generating hallucinated summaries.

4 Experiment

4.1 Datasets

Document	[...]A full road closure will be in place in Cardiff city centre from 12:30 GMT to 17:30 with the kickoff at 14:30 at the Principality Stadium.Arriva Trains Wales passengers are advised a new queuing system will be in place at Cardiff Central for the match.[...]
Summary	Rugby fans are advised to be aware of travel restrictions and road closures due to Wales' Six Nations clash against Italy in Cardiff of Saturday.

Table 1: Examples of hallucinations in the gold summaries of the XSum dataset.

Our experiments were conducted on the XSum dataset (Narayan et al., 2018) which is highly suitable for evaluating abstractive single-document summarization systems. This dataset comprises 226,711 news articles from the British Broadcasting Corporation (BBC) and a one-sentence summary for each article. The official random division includes 204,045 (90%) training documents, 11,332 (5%) validation documents, and 11,334 (5%) test documents. It has been reported Maynez et al. (2020) that over 70% of the gold summaries in this dataset exhibit hallucinations, making it an ideal dataset for studying the factual consistency of abstractive summarization. Table 1 displays examples of hallucinations in the gold summaries from the XSum dataset.

4.2 Baselines

The following baselines are used for comparison with the results obtained by our model:

DAE: Goyal and Durrett (2021) explored whether the factual errors made by synthetic data and generative models are consistent, and the role of fine-grained human-annotated data in training more factual summarization models. It shows that there are significant differences in the factual errors exhibited by different datasets, and fine-grained human-annotated data can help train more factual summarization models, thereby improving the model's factual consistency.

FactPEGASUS: Wan and Bansal (2022) proposed a summarization model FactPEGASUS that incorporates factuality into the entire training process. During the pre-training process, it explored the combination of ROUGE and FactCC as selection criteria, so that the model can learn important and factually accurate sentences from the input document. During fine-tuning, it introduced three supplementary components to enhance factuality during fine-tuning, greatly improving the model's factual consistency.

CLIFF: Based on the reference summary, four types of strategies were designed to create negative samples: deletion, replacement, rearrangement, and hallucinations (Cao and Wang, 2021). Then, a training objective based on contrastive learning was proposed, allowing the model to better distinguish between factual summaries and erroneous summaries, thereby improving the model's factual consistency. We use the better-performing MASKENT model as the baseline.

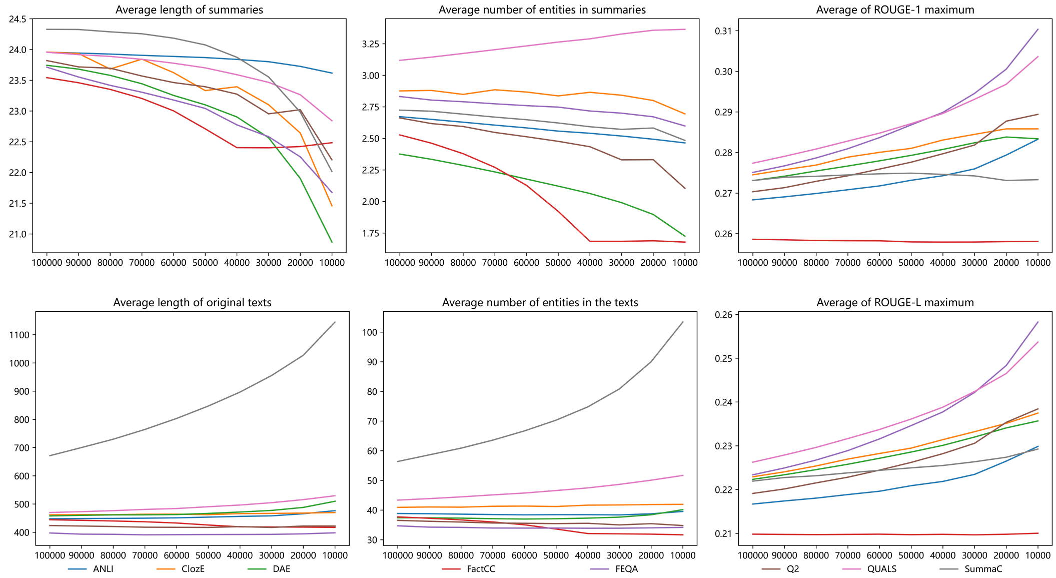

Figure 2: Characteristics of the data obtained when filtering the XSum training set using different evaluation metrics. The x-axis represents the number of data selected according to factual consistency ranking.

Model	RL	SummaC	FactCC	ClozE	DAE	ANLI	FEQA	Q2	QUALS
PEGASUS	**37.06**	25.12	24.71	70.64	64.22	86.27	14.51	28.40	-0.6786
Fine-tune with QUALS	28.26	25.17	24.34	75.58	69.90	87.81	18.11	34.23	-0.6215
Fine-tune with FEQA	28.09	25.36	24.40	74.92	70.25	88.22	**21.50**	35.18	-0.6325
Fine-tune with Q2	26.95	25.79	27.57	74.39	69.68	89.13	17.66	35.61	-0.6500
Fine-tune with ClozE	26.16	25.43	26.03	**77.85**	68.20	86.57	17.33	31.47	-0.6537
Fine-tune with DAE	26.24	25.55	30.84	73.34	**74.48**	86.86	16.85	33.47	-0.6498
Fine-tune with ANLI	26.28	25.48	25.64	73.80	68.17	88.99	17.10	32.51	-0.6572
Fine-tune with FactCC	24.95	25.07	31.34	72.69	68.94	86.83	14.89	31.54	-0.6926
Fine-tune with SummaC	33.19	**29.31**	**31.69**	75.52	70.37	**90.24**	18.68	**36.40**	**-0.5822**

Table 2: Factual consistency of summaries generated using the PEGASUS model fine-tuned with data filtered by different evaluation metrics on the XSum test dataset.

4.3 Evaluation Metrics

In this section, we summarize the commonly used cutting-edge factual consistency evaluation metrics for data filtering and evaluating the factual consistency of model-generated summaries:

ROUGE: ROUGE (Lin, 2004) is a commonly used metric for evaluating text generation tasks. Its principle is to calculate the overlap text between the golden summary and the model output. Here we use it to evaluate the fluency and information content of the generated summary.

SummaC: Laban et al. (2022) provides an efficient and lightweight method called SUMMAC-CONV, which divides the document into sentence units and aggregates the scores between each sentence and the summary, solving the input granularity mismatch between the NLI dataset (sentence level) and inconsistency detection (document level), thus enabling the NLI model to be better used for inconsistency detection and achieving good results.

FactCC: FactCC (Kryściński et al., 2019) can identify whether the summary is consistent with the text fragments in the source document. This model is trained on synthetic data generated by transforming basic facts through paraphrasing, swapping entities, numbers, pronouns, etc.

ClozE: Li et al. (2022) evaluates factual consistency through a cloze model, which is instantiated based on a masked language model, with strong interpretability and fast speed.

DAE: DAE (Goyal and Durrett, 2021) uses dependency arcs to identify non-factual markers in the data. For a factual example, all individual arcs in the summary must be factual. For non-factual examples, at least one arc must be non-factual.

ANLI: ANLI (Nie et al., 2019) is a famous text entailment dataset. Natural language inference models trained on ANLI show good performance in factual detection. Therefore, we use ANLI as one of the factual consistency evaluation metrics.

FEQA: Durmus et al. (2020) proposes a factual consistency evaluation metric based on automatic question answering. It uses the summary to generate question-answer pairs, and then extracts answers from the document; mismatched answers indicate that the information in the summary is not true.

Q2: Honovich et al. (2021) proposes an automatic evaluation metric for evaluating factual consistency in dialogue using automatic question generation and question answering, and combines the NLI model to compare the answer range.

QUALS: QUALS(Nan et al., 2021) is also an evaluation metric based on automatic question answering, and proposes an efficient algorithm to speed up the calculation. Its evaluation results have been proven to have a strong correlation with FEQA.

4.4 Features of Data Filtered by Different Metrics

In order to select the optimal filtering metric, we calculated the factual consistency scores for each data point in the XSum training set using a variety of factual consistency evaluation metrics, then ranked and filtered out the data with higher factual consistency. We conducted an in-depth analysis of the characteristics of data filtered by different metrics from perspectives such as summary length, the number of entities in the summary, the length of the original text, the number of entities in the original text, and the maximum value obtained from calculating the ROUGE scores sentence by sentence between the summary and the original text. As shown in Figure 2, as the filtered data decreases from 100,000 to 10,000, that is, the factual consistency continues to increase, the characteristics of the filtred data by various metrics also change accordingly. By analyzing the curves in Figure 2 and the principles of calculating factual consistency with different metrics, we have summarized some characteristics of different metrics when filtering data:

(1) From the perspective of summaries, as the factual consistency of the selected summaries improves, the length of the summaries gradually decreases. In addition, the summaries filtered out by the FEQA and QUALS metrics based on the question-answering model and the ClozE metric based on the cloze test always contain more entities. We speculate that this is related to their design of question-answer pairs or cloze tests always revolving around the entities in the summary.

(2) From the perspective of the original text, the data filtered out by the SummaC metric has a far greater original text length and number of entity words than other metrics. This can be attributed to the SummaC metric dividing the document into sentence units and aggregating the scores between each sentence and the summary, solving the input granularity mismatch between the NLI dataset (sentence level) and inconsistency detection (document level). Other evaluation metrics usually truncate when the original text length exceeds the model encoding limit, so their evaluation capabilities for long texts are limited. Therefore, SummaC is suitable for evaluating texts of various lengths, is a very balanced metric, and is more suitable for filtering datasets. We will further verify this speculation through the results of fine-tuning the model.

(3) From the perspective of the maximum value obtained by calculating the ROUGE scores sentence by sentence between the summary and the original text, as the factual consistency of the selected data increases, the maximum ROUGE scores of the summary and the original text filtered out by other metrics is increasing, while the maximum ROUGE scores of the data filtered out by the FactCC metric has no significant change. We speculate that this is related to the FactCC metric's evaluation of factual consistency when extracting text fragments that support consistency or do not support consistency in the source document is relatively scattered.

4.5 Results of Fine-tuned Model on Filtered Data

To further analyze the effect of data filtering by different metrics, we sorted the data filtered by different metrics based on factual consistency. We selected the top 10,000 data to fine-tune the PEGASUS model, then predicted on the XSum test dataset, and evaluated the factual consistency of the generated summaries.

As shown in Table 2, when the SummaC metric is used to filter data and fine-tune the model, a significant performance improvement can be achieved, which is in stark contrast to the performance of other metrics under the same conditions. When we use the model fine-tuned with data filtered by SummaC to predict the XSum test dataset, the generated summaries not only have the highest ROUGE scores but also achieve the best results on all five factual consistency evaluation metrics. In addition, significant improvements have also been made on the other three factual consistency evaluation metrics. These results further validate our previous conclusion: SummaC is a very balanced metric, suitable for evaluating texts of various lengths, and therefore, it is more suitable for data filtering.

4.6 Threshold of Data Filtering

	Threshold	RL	SummaC	FactCC	ClozE	DAE	ANLI
SummaC	10000	33.19	**29.31**	**31.69**	**75.52**	**70.37**	**90.24**
	20000	34.53	26.58	28.75	73.55	68.66	89.60
	30000	35.13	25.99	28.39	72.96	67.98	89.26
	40000	34.98	26.06	28.29	73.35	67.83	89.19
	50000	**35.30**	25.70	26.84	73.23	67.46	88.87

Table 3: Results obtained by fine-tuning the model with different amounts of data selected according to the factual consistency score using the SummaC metric.

After filtering the data using the SummaC metric, in order to determine a reasonable threshold, we use the SummaC metric to filter out varying amounts of data for fine-tuning the PEGASUS model, followed by making predictions on the XSum test dataset.

Table 3 illustrates the impact of adjusting the threshold of data filtering on the fine-tuning model. Here, the threshold represents the amount of data taken after sorting according to the factual consistency score of the data. The lower the threshold, the higher the factual consistency score of the filtered data, and the model will use these factually consistent articles and summaries for training, thereby improving the factual consistency score. However, the lower the threshold, the fewer the data filtered out, that is, the training samples are reduced, which may lower the ROUGE-L score. In the several sets of threshold experiments we are currently conducting, when the threshold is 10,000, the ROUGE score is within an acceptable decline range, while achieving the best factual consistency.

4.7 Construction of the Dataset for Model Self-Improvement Contrastive Learning

The key to Model Self-Improvement Contrastive Learning lies in the construction of the dataset. For the XSum training set, we constructed a dataset that includes the original text, factual summaries, and hallucinatory summaries using the positive and negative sample construction methods introduced in Section 3.2 of this paper. We then introduced contrastive learning to fine-tune the PEGASUS model to enhance the model's factual consistency. Table 4 shows an example of the self-improvement contrastive learning dataset we constructed. As can be seen from Table 4, the dataset we constructed takes into account the factual consistency of both positive and negative samples, and has specifically improved the quality of positive and negative samples.

Document	[...]Last week it was revealed Ofsted reported that Market Rasen School in Lincolnshire could not be rated as outstanding because pupils lack experience of "the diverse make-up of modern British society".[...]
Factual Summary	A school has been told it cannot be rated as "outstanding" by Ofsted because pupils need to have more awareness of other British cultures.
Hallucinated Summary	A school has been rated "inadequate" by Ofsted because pupils do not have a strong awareness of other cultures.

Table 4: Examples from the dataset for Model Self-Improvement Contrastive Learning

5 Result

5.1 Fine-tuned result

We utilized multiple factual consistency metrics described in Section 4.3 of this paper to calculate the factual scores of the model on the XSum test dataset. As shown in Table 5, compared to other baseline models, the model fine-tuned with the Model Self-Improvement Contrastive Learning method achieved comprehensive and significant improvements on multiple factual consistency evaluation metrics, reaching the state-of-the-art level. However, the score of our model on the ROUGE metric has decreased. We speculate that this may be because in the XSum dataset, more than 70% of the gold summaries contain hallucinations (Maynez et al., 2020), including its test dataset. Therefore, improving the factuality of the model may lead to the generated summary not completely matching the gold summary, thereby leading to a decrease in the ROUGE score. Therefore, this is a reasonable phenomenon.

Method	Lexical overlap			Factual Consistency							
	R1	R2	RL	SummaC	FactCC	ClozE	DAE	ANLI	FEQA	Q2	QUALS
PEGASUS	**42.88**	**21.73**	**37.06**	25.12	24.71	70.64	64.22	86.27	14.51	28.40	-0.6786
DAE	24.46	7.23	20.44	25.87	34.11	51.85	55.64	71.10	10.58	26.94	-0.7501
CLIFF	41.81	20.54	35.86	24.84	25.14	73.54	67.49	86.53	16.20	30.83	-0.6627
FactPEGASUS	25.74	8.23	21.53	27.69	**40.96**	70.99	70.28	58.99	16.95	35.48	-0.5862
MSCL	36.86	16.71	31.31	**31.83**	38.43	**76.25**	**73.16**	**90.25**	**20.82**	**39.84**	**-0.5433**

Table 5: Prediction results of different models or methods on the XSum test dataset. In the table, 'MSCL' stands for Model Self-Improvement Contrastive Learning, and this applies to the subsequent tables as well.

Method	RL	SummaC	FactCC	ClozE	DAE	ANLI	FEQA	Q2	QUALS
PEGASUS	**37.06**	25.12	24.71	70.64	64.22	86.27	14.51	28.40	-0.6786
Data Filtering	33.19	29.31	31.69	75.52	70.37	90.24	18.68	36.40	-0.5822
Contrastive Learning	34.55	26.45	27.43	74.16	68.54	87.99	18.05	34.31	-0.6500
MSCL	31.31	**31.83**	**38.43**	**76.25**	**73.16**	**90.25**	**20.82**	**39.84**	**-0.5433**

Table 6: Results of ablation study on Data Filtering and Contrastive Learning

5.2 Ablation Studies

Table 6 presents the results of the ablation study of our proposed method, which includes the model's scores on multiple evaluation metrics. The results indicate that compared to the baseline model, either the individual application of data filtering method or contrastive learning method can improve the model's factual consistency. However, our proposed method of Modle Self-Improvement Contrastive Learning can combine the advantages of data filtering and contrastive learning to further enhance the model's factual consistency. It is worth noting that the model fine-tuned with our proposed method of Model Self-Improvement Contrastive Learning achieved the best performance on multiple evaluation metrics without a significant drop in ROUGE scores.

5.3 Human Evaluation

Method	Factual Consistency
PEGASUS	31
DAE	33
CLIFF	29
FactPEGASUS	40
MSCL	**44**

Table 7: Manual evaluation results of summaries generated by different methods.

We conducted a manual evaluation of the summaries generated by different models, randomly selecting one hundred pieces of data to assess their factual consistency. As shown in Table 7, consistent with the results of our evaluations using various evaluation metrics, our model generates significantly more factual summaries compared to models such as PEGASUS, DAE, CLIFF, and FactPEGASUS.

6 Conclusion

In this research, we propose a Model Self-Improvement Contrastive Learning method that takes into account the factual consistency of both positive and negative samples. We purposefully construct negative samples, enhance positive samples, and then reinforce model-related knowledge through contrastive learning to improve the model's factual consistency. The evaluation results of multiple factual consistency metrics, along with ablation studies and human results, demonstrate that our method effectively enhances the factual consistency of the model, achieving significant improvements in experiments on the XSum dataset. These results further validate the effectiveness and practicality of the method we proposed.

7 Limitations

Due to the limitations of computational resources, we did not directly conduct experimental verification on large models.Therefore, one of our future research directions is to conduct more in-depth experimental verification on large models. We look forward to these experiments further confirming the effectiveness and scalability of our proposed method on more complex models and larger datasets.

References

Shuyang Cao and Lu Wang. 2021. Cliff: Contrastive learning for improving faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2109.09209*.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. Faithful to the original: Fact aware neural abstractive summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Subhajit Chaudhury, Sarathkrishna Swaminathan, Chulaka Gunasekara, Maxwell Crouse, Srinivas Ravishankar, Daiki Kimura, Keerthiram Murugesan, Ramón Fernandez Astudillo, Tahira Naseem, Pavan Kapanipathi, et al. 2022. X-factor: A cross-metric evaluation of factual correctness in abstractive summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7100–7110.

Sihao Chen, Fan Zhang, Kazoo Sone, and Dan Roth. 2021. Improving faithfulness in abstractive summarization with contrast candidate generation and selection. *arXiv preprint arXiv:2104.09061*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Esin Durmus, He He, and Mona Diab. 2020. Feqa: A question answering evaluation framework for faithfulness assessment in abstractive summarization. *arXiv preprint arXiv:2005.03754*.

Hongchao Fang, Sicheng Wang, Meng Zhou, Jiayuan Ding, and Pengtao Xie. 2020. Cert: Contrastive self-supervised learning for language understanding. *arXiv preprint arXiv:2005.12766*.

Tanya Goyal and Greg Durrett. 2021. Annotating and modeling fine-grained factuality in summarization. *arXiv preprint arXiv:2104.04302*.

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. Q^2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. *arXiv preprint arXiv:2104.08202*.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Evaluating the factual consistency of abstractive text summarization. *arXiv preprint arXiv:1910.12840*.

Philippe Laban, Tobias Schnabel, Paul N Bennett, and Marti A Hearst. 2022. Summac: Re-visiting nli-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Yafu Li, Leyang Cui, Jianhao Yan, Yongjng Yin, Wei Bi, Shuming Shi, and Yue Zhang. 2023. Explicit syntactic guidance for neural text generation. *arXiv preprint arXiv:2306.11485*.

Yiyang Li, Lei Li, Qing Yang, Marina Litvak, Natalia Vanetik, Dingxin Hu, Yuze Li, Yanquan Zhou, Dongliang Xu, and Xuanyu Zhang. 2022. Just cloze! a fast and simple method for evaluating the factual consistency in abstractive summarization. *arXiv preprint arXiv:2210.02804*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Zhou LIN and Qi-feng ZHOU. 2023. A text summarization model guided by key information. *Journal of Northeastern University (Natural Science)*, 44(9):1251.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*.

Feng Nan, Cicero Nogueira dos Santos, Henghui Zhu, Patrick Ng, Kathleen McKeown, Ramesh Nallapati, Dejiao Zhang, Zhiguo Wang, Andrew O Arnold, and Bing Xiang. 2021. Improving factual consistency of abstractive summarization via question answering. *arXiv preprint arXiv:2105.04623*.

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. *arXiv preprint arXiv:1808.08745*.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2019. Adversarial nli: A new benchmark for natural language understanding. *arXiv preprint arXiv:1910.14599*.

David Wan and Mohit Bansal. 2022. Factpegasus: Factuality-aware pre-training and fine-tuning for abstractive summarization. *arXiv preprint arXiv:2205.07830*.

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.