Masked Summarization to Generate Factually Inconsistent Summaries for Improved Factual Consistency Checking

Hwanhee Lee[1*], **Kang Min Yoo**[2], **Joonsuk Park**[2,3], **Hwaran Lee**[2†] **and Kyomin Jung**[1†]
[1]Seoul National University, [2]NAVER AI Lab, [3]University of Richmond
{wanted1007,kjung}@snu.ac.kr
{kangmin.yoo, hwaran.lee}@navercorp.com
park@joonsuk.org

Abstract

Despite the recent advances in abstractive summarization systems, it is still difficult to determine whether a generated summary is factual consistent with the source text. To this end, the latest approach is to train a factual consistency classifier on factually consistent and inconsistent summaries. Luckily, the former is readily available as reference summaries in existing summarization datasets. However, generating the latter remains a challenge, as they need to be factually inconsistent, yet closely relevant to the source text to be effective. In this paper, we propose to generate factually inconsistent summaries using source texts and reference summaries with key information masked. Experiments on seven benchmark datasets demonstrate that factual consistency classifiers trained on summaries generated using our method generally outperform existing models and show a competitive correlation with human judgments. We also analyze the characteristics of the summaries generated using our method. We will release the pre-trained model and the code at https://github.com/hwanheelee1993/MFMA.

1 Introduction

As textual content available on- and offline explodes, automated text summarization is becoming increasingly crucial (El-Kassas et al., 2020); with the advances in neural text generation methods, abstractive summarization systems that generate paraphrases are quickly replacing extractive ones that simply select essential sentences from the source text (Nallapati et al., 2017). While abstractive summaries can be more coherent and informative (given the same length) than their extractive counterparts, they frequently contain information inconsistent with the source text. This is a critical issue, as it directly affects the reliability of the generated summaries. (Cao et al., 2018; Zhao et al., 2020; Maynez et al., 2020).

Article: Guus Hiddink, the Russia and Chelsea coach, has had much to smile about in his 22-year managerial career. ,. . . , Enjoying success around the world – at different levels with different players in different cultures – has made Guus Hiddink one of the most admired bosses around. ,. . . , Hiddink's resume includes stints in other high-pressure jobs such as Fenerbahce, Valencia and Real Madrid. ,. . . , But the straight-speaking Dutchman is loyal to the project he has in charge of the Russian national side and insists he will leave Chelsea at the end of the season regardless.

Reference Summary: Born in 1946, Hiddink has become one of the best managers in the world . Dutchman has enjoyed huge success at club and international level. He's currently coach of Russia and is in charge of Chelsea until end of the season.

Mask-and-fill Summary Without Article:
Born in 1946, *Dutchman* has become one of *the most respected politicians* in the world. Dutchman is enjoyed *success at the Olympics and World Cup*. He's currently the *President of Russia* and is in charge of *the country* until the end of the season.

Mask-and-fill Summary With Masked Article:
Born in 1946, *Hiddink* has become one of *the most admired managers* in the world. Dutchman has enjoyed *successful spells* at *Chelsea and Real Madrid*. He's currently *manager of Russia* and is in charge of *the country* until the end of the season.

Figure 1: An example of generated negative summary using masked article. Spans that are highlighted are masked when generating the negative summary. Note that red spans are factually inconsistent with the given article and blue spans are factually consistent.

Unfortunately, existing approaches to identify such factual inconsistency without constructing new resources have not been satisfactory. Directly measured similarity between the summary and its source text—using popular n-gram similarity metrics such as ROUGE (Lin, 2004) and BLEU (Papineni et al., 2002)—exhibits low correlation with

[*]Work done during an internship at NAVER AI Lab.
[†]Corresponding authors.

Findings of the Association for Computational Linguistics: NAACL 2022, pages 1019 - 1030

human judgments for factual consistency. Also, leveraging related tasks—such as natural language inference (NLI) (Bowman et al., 2015) and fact verification (Thorne et al., 2018)—is not ideal. This is because these tasks aim to identify relations between two sentences, whereas factual consistency checking involves a multi-sentence summary and an even longer source text (Bora-Kathariya and Haribhakta, 2018; Falke et al., 2019).

A remaining solution is to train a factual consistency classifier with a dataset specifically constructed for this purpose. Note that ***positive summaries*** are readily available. That is, the reference summaries from existing text summarization datasets can be assumed to be factually consistent with the respective source texts. Thus, the main challenge is in generating effective ***negative summaries***, i.e., summaries that are factually inconsistent with the source text. Recent works generate negative summaries by simply replacing keywords in the reference summaries or sentences extracted from the source texts (Kryscinski et al., 2020; Yin et al., 2021). This, however, results in negative summaries that significantly diverge from the source texts and positive summaries, which is not ideal for training factual consistency classifiers. For instance, Figure 1 shows that *coach* in the reference summary is changed to *President of Russia*, which is an inconsistency that is too obvious.

In this paper, we propose a novel method, Masked-and-Fill with Masked Article (MFMA), where parts of the source text and reference summary are masked and later inferred to generate a plausible but factually inconsistent summary. Experiments on seven benchmark datasets demonstrate that factual consistency classifiers trained on negative summaries generated with our method mostly outperform existing models and show a competitive correlation with human judgment. We also analyze the characteristics of the negative summaries generated. Our main contributions are as follows:

- We propose a novel negative summary generation method for training factual consistency classifiers for abstractive summaries.
- We show the efficacy of our method on seven benchmark datasets using classification performance and correlation with human judgment.
- We analyze the characteristics, such as affinity and diversity, of the negative summaries generated using our method.

2 Related Work

2.1 Factual Inconsistency in Summarization Systems

Previous works (Maynez et al., 2020; Zhao et al., 2020; Cao et al., 2018) have studied the factual inconsistency in abstractive summarization systems. Especially, (Cao et al., 2018) demonstrates that 30% of the model generated summaries have at least one factual error, and this obstacle the practical usage. (Maynez et al., 2020) specifies these factual errors in the abstractive summarization system into two types: *intrinsic errors* and *extrinsic errors*. Intrinsic errors occur using the contents present in the source article like "Switzerland" and "England" in the negative summary example in Figure 2. On the other hand, extrinsic errors are the errors generated by ignoring the source article when generating summaries. "in the second half" in Figure 2, which is not included in the source article, is an example of extrinsic errors.

In this work, we propose a system for detecting these various factual errors that are necessary for developing a summarization system. We propose a unified method for intentionally modeling both types of errors to build a dataset for training this system.

2.2 Measuring Factual Consistency

As a better way to evaluate the factual consistency, recent works such as QAGS (Wang et al., 2020) and QuestEval (Scialom et al., 2021) adopt question generation and question answering frameworks to evaluate the factual consistency. Both methods firstly generate questions using entities or noun phrases in the candidate summary and then compare the answers of these questions between the source and the summary. Although these methods do not require any reference summaries, they have a higher correlation with human judgments than previous metrics in consistency checking. Also, the generated questions and their answers are easily interpretable. But due to their complicated structure, the computational complexity of these methods is relatively heavy and the errors in each component can be cascaded.

Following the idea that all of the contents in the summaries should be entailed by source document, models from the related tasks such as Natural Language Inference(NLI) (Bowman et al., 2015; Williams et al., 2018; Falke et al., 2019) are also used to verify the factual consistency of the sum-

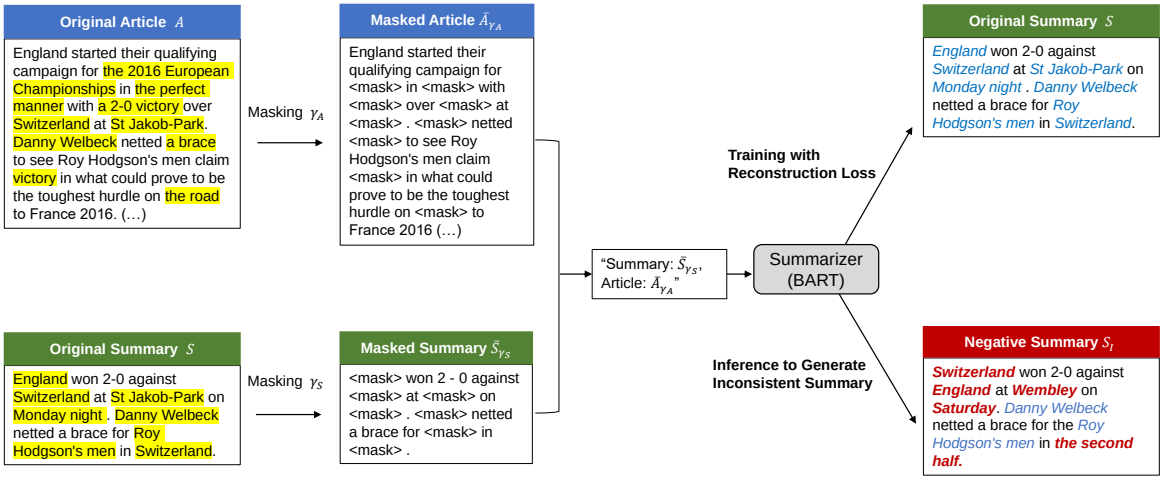

Figure 2: Overall flow of our proposed negative summary generation method Mask-and-Fill-with-Masked Article.

maries. These approaches are simpler and more intuitive than QA-based metrics. But the data pairs in these datasets are usually composed of single sentences, and this makes it difficult to be directly used for factual consistency checking in summarization where the task requires multi-sentence level reasoning. For this reason, two recent studies FactCC (Kryscinski et al., 2020) and DocNLI (Yin et al., 2021) have studied ways to make synthetic datasets for training factual consistency checking model. Both works create synthetic negative summaries using the pre-defined rules such as entity substitution or mask-and-fill. In this paper, we propose a more general negative summary generation method additionally using the masked source.
CoCo (Xie et al., 2021) compares the likelihood of the generated summaries using the original source and the masked source to estimate the counterfactual samples. Different from CoCo, our work directly augments the negative summaries and train the classifier using them.

3 Methods

For a given article A and a summary S, we aim to develop a factual consistency checking system that can evaluate whether S is factual consistent with A. In other words, the system is required to discriminate a factual consistent summary S_C with the factual inconsistent summary S_I that consists of at least one factual error. We consider this problem as a classification task between S_C and S_I. However, large-scale human-annotated training datasets for this task have not been constructed yet, especially for the inconsistent summaries S_I.

In this paper, we focus on effective augmentation methods of the inconsistent summaries. In order for that, there are two crucial conditions: 1) guarantee of inconsistency; the generated summaries should be indeed inconsistent with the source article, 2) relevance to the source article; the generated summaries should include contents related to the article. These two factors are in trade-off relations, which means that when the generated summaries are strongly inconsistent they might not be related to the article and vice versa. Therefore appropriate negative summary augmentation is required to improve the factual consistency classifier.

To generate confusing and hard negative summaries, we propose a summary generation using a masked article and a masked reference summary where some salient information is hidden. By doing so, we let the summarizer model infer hidden information through the masked article to generate plausible negative summaries. Note that, previous works such as FactCC and DocNLI generate negative summaries S_I by changing positive summaries S_C through entity replacements or mask-and-fill methods without referring to the source article. We observe that previous methods can easily guarantee negativeness, but they often generate summaries that are very irrelevant to the source article or unnatural as shown in Figure 1.

3.1 Mask-and-Fill with Masked Article

To model inconsistent summaries but related to the article, we propose a method, **M**ask-and-**F**ill with **M**asked **A**rticle (MFMA), which generates negative summaries with masked articles and masked reference summaries, as shown in Figure 2.

Specifically, we assumed *noun phrases* and *entities* in the articles are salient information, and mask

them with the ratio of γ_A, resulting in masked article $\overline{A}_{\gamma_A}$. Similarly, we also mask the salient spans in the positive summary, i.e., reference summary, with the ratio of γ_S to form a masked summary $\overline{S}_{\gamma_S}$. Then, we concatenate $\overline{A}_{\gamma_A}$ and $\overline{S}_{\gamma_S}$ by prepending prefix token for each input text (i.e., "Summary: $\overline{S}_{\gamma_S}$, Article: $\overline{A}_{\gamma_A}$") as shown in Figure 2. Next, we train a summarizer based on an encoder-decoder model, BART (Lewis et al., 2020), to reconstruct the original summary S with the following loss:

$$\mathcal{L} = \sum_t -\log P(S_t|S_{<t}, [\overline{S}_{\gamma_S}; \overline{A}_{\gamma_A}]). \quad (1)$$

After training, we generate negative summaries of unseen and masked article-summary pairs through inference. Obviously, if the mask ratio is high enough, the model is hard to correctly fill the masked contents from the erased article and reference summary. However, we assume the trained reconstruction model is able to fill the masks with plausible contents by inferring the related contents with the masked article.

3.2 Masked Summarization

As a variant of MFMA, we also study another negative summary generation model, **M**asked **SuM**marization(MSM). The model aims to generate summaries using masked articles $\overline{A}_{\gamma_A}$ but without masked reference summaries as follows:

$$\mathcal{L} = \sum_t -\log P(S_t|S_{<t}, \overline{A}_{\gamma_A}). \quad (2)$$

The MSM model is trained to generate the entire summaries without the information guidance of masked reference summaries, so MSM has merits in generating more diverse summaries than MFMA.

3.3 Training Factual Consistency Checking Model

Finally, for the factual consistency checking model, we train a binary classifier of consistent summaries and inconsistent generated summaries. The pair of summary and the corresponding article are concatenated and then fed into the classification model as an input. We fine-tuned the pre-trained ELECTRA (Clark et al., 2019) by adding a classifier head with binary cross-entropy loss.

4 Experiments

4.1 Implementation Details

Negative Summary Generation We randomly split the training set of CNN/DM dataset (Nallapati et al., 2016) in half and use half for training negative summarizer and the other half for generating negative summary after training. We use *spaCy* for finding entities and noun phrases in both summaries and articles. We train *bart-base*[1] for five epochs to train MFMA, and use *bart-base* model without fine-tuning for MF. We use *t5-small* (Raffel et al., 2020)[2] for MSM, which shows better results than *bart-base* for this task. We attach the further details in Appendix.

Training Classifier We train *google/electra-base-discriminator*[3] for five epochs with learning rate 2e-5, batch size of 96 using adam optimizer (Kingma and Ba, 2015) with the dataset we generate using MF, MFMA and MSM. For DocNLI and FactCC, we get the original training dataset that each author release, and we train a model with the same setting as our method except for the training datasets for a fair comparison. We choose model using the balanced accuracy on validation set of FactCC (Kryscinski et al., 2020) which consists of 1k human annotated summaries.

4.2 Benchmark Datasets

For evaluating the performance of factual consistency checking system, it is necessary to compare the human judgments of the consistency for the summary with the system. And these human judgment exist in two forms, binary level(*consistent*, *inconsistent*) or numerical levels such as likert scale. In general, in the case of binary level data, performance is measured through accuracy with human judgments. For the case of numerical levels, correlation with human judgments is measured. In addition to using the results for the existing benchmark dataset in this way, we also report the accuracy by casting these numerical level datasets to the binary level dataset since we develop classifier based system. We report the results on the following datasets.

FC-Test (Kryscinski et al., 2020) release a human-annotated factual consistency for the model generated summaries for CNN/DM Dataset in

[1]https://huggingface.co/facebook/bart-base
[2]https://huggingface.co/t5-small
[3]https://huggingface.co/google/electra-base-discriminator

Table 1: Macro F1-score(F1) and class-balanced accuracy(BA) of the human annotated factual consistency for the benchmark datasets based on CNN/DM.

Dataset	FactCC-Test		SummEval		QAGS-CNN/DM		FRANK-CNN/DM		Average	
Metric	**F1**	**BA**	**F1**	**BA**	**F1**	**BA**	**F1**	**BA**	**F1**	**BA**
Baselines										
FactCC	71.0	71.3	65.1	68.2	69.3	69.6	64.1	63.9	67.4	68.2
DocNLI	67.2	71.0	**71.5**	**71.3**	62.4	66.2	66.0	66.0	66.8	68.6
MNLI	55.0	56.0	51.7	51.7	48.6	53.4	50.4	53.3	51.4	53.6
FEVER	57.9	56.2	52.6	53.6	39.4	53.3	49.8	55.6	49.9	54.7
MF	59.9	64.1	68.2	67.5	47.6	56.9	62.4	62.7	59.5	62.8
Ours										
MFMA	**79.7**	**84.5**	71.3	69.6	**70.5**	**72.3**	**69.5**	69.2	**72.8**	**73.9**
MSM	70.6	72.7	66.8	68.2	67.6	68.7	69.6	**69.3**	68.6	69.7

Table 2: Macro F1-score(F1) and class-balanced accuracy(BA) of the human annotated factual consistency for the benchmark datasets based on XSum.

Dataset	XSumHall		QAGS-XSum		FRANK-XSum		Average	
Metric	**F1**	**BA**	**F1**	**BA**	**F1**	**BA**	**F1**	**BA**
Baselines								
FactCC	52.1	**61.8**	63.6	63.7	50.7	58.0	55.5	61.2
DocNLI	55.1	**56.4**	65.3	66.0	**60.3**	**63.4**	60.2	**61.9**
MNLI	33.3	52.1	45.2	51.1	28.8	50.6	35.8	51.3
FEVER	53.1	55.5	62.2	63.7	54.9	63.5	56.7	60.9
MF	53.6	53.3	54.6	54.9	55.7	55.3	54.6	54.5
Ours								
MFMA	**55.5**	56.0	**66.6**	**67.0**	59.6	59.6	**60.6**	60.9
MSM	52.6	53.9	50.8	55.5	50.8	51.3	51.4	53.6

binary-level to test the performance of FactCC. There are 513 instances in this dataset.

XSumHall (Maynez et al., 2020) study the types of hallucination in the generated summaries and collect the annotation on the errors in the 2K model generated summary for BBC XSum dataset (Narayan et al., 2018). We use the datasets as binary level benchmark for XSum dataset as in (Kryscinski et al., 2020).

SummEval (Fabbri et al., 2021) collect the likert scale human judgments for the 1600 summaries generated from sixteen abstactive summarizer on CNN/DM testset. This dataset provides human judgments scores in terms of *"coherence" , "consistency", "fluency", and "relevance"* by three expert annotators in likert scale. We only use *"consistency"* score of three annotators, for evaluating our proposed metric. For casting this score to binary level, we let the cases where at least one annotators give less than 5 points for *"consistency"* as *inconsistent*, otherwise *consistent*.

QAGS-CNN/DM & XSum (Wang et al., 2020) release a human judgments for factual consistency on the model generated summaries for 235 summaries on CNN/DM testset and 239 summaries on XSum testset. Each summary is annotated by three annotators. We also cast the dataset to binary level by assigning *inconsistent* if at least one annotators give *inconsistent* label, otherwise *consistent*.

FRANK-CNN/DM & XSum (Pagnoni et al., 2021) releases a benchmark dataset FRANK for summarization factual metrics which consists of 2246 summaries on the model generated summaries for 1250 summaries in CNN/DM and 996 summaries XSum. Three annotators evaluated factual consistency of the generated summaries in this dataset. We also convert this dataset to binary level as same as QAGS-CNN/DM and QAGS-XSum.

4.3 Baseline Metrics

We compare our methods with the following metrics. For all of the baseline metrics, we manually compute the score using the official repository which each author provided or reproducing the model for a fair comparison.

Table 3: Summary level Pearson Correlation(r) and Spearman's Correlation(ρ) between various automatic metrics and human judgments of factual consistency for the model generated summaries. Note that we use the confidence of consistency label for entailment based metrics.

Dataset	**SummEval**		**QAGS-CNN/DM**		**QAGS-XSum**		**FRANK-CNN/DM**		**FRANK-XSum**	
Metric	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ
Baselines										
ROUGE-L	0.16	0.14	0.29	0.24	0.13	0.13	0.16	0.13	0.16	0.13
BLEU-4	0.11	0.12	0.18	0.23	0.03	0.03	0.16	0.17	0.11	0.14
METEOR	0.18	0.16	0.26	0.25	0.11	0.12	0.29	0.28	0.18	0.16
BERTScore	0.16	0.14	0.37	0.36	0.11	0.13	0.33	0.30	0.19	0.17
QuestEval	0.35	0.30	0.42	0.36	0.20	0.20	0.46	0.41	0.19	0.18
CoCo	0.42	0.36	**0.67**	0.57	0.20	0.18	0.50	0.45	0.14	0.12
FactCC	0.38	0.36	0.45	0.48	0.30	0.30	0.32	0.36	0.09	0.08
DocNLI	0.51	**0.41**	0.60	0.59	0.36	0.35	0.49	**0.49**	**0.25**	**0.21**
MNLI	0.11	0.13	0.19	0.22	0.08	0.10	0.15	0.16	0.02	0.03
FEVER	0.33	0.32	0.40	0.34	**0.38**	**0.41**	0.38	0.43	0.20	0.19
MF	0.44	0.35	0.43	0.30	0.10	0.10	0.40	0.39	0.10	0.13
Ours										
MFMA	**0.52**	0.38	0.62	**0.65**	0.37	0.38	**0.52**	0.45	0.16	0.17
MSM	0.43	0.36	0.50	0.48	0.20	0.22	0.51	0.48	0.05	0.09

Entailment Based Metrics We adopt the model trained on MNLI (Bowman et al., 2015) and FEVER (Thorne et al., 2018) for factual consistency checking as in (Kryscinski et al., 2020). FactCC (Kryscinski et al., 2020) and DocNLI (Yin et al., 2021) are also entailment based models trained on synthetic dataset as in our work.

QA-Based Metrics QuestEval (Scialom et al., 2021) uses the question generation and answering framework for evaluating the factual consistency of the summaries. QuestEval generates the question both the generated summaries and the source article, and then compare the answers of them with both summaries and the article to compute the factuality score of the summary.

N-gram Similarity Metrics BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005) are widely used for evaluating the summaries. Among them, ROUGE-L, which uses F-measure based on the longest common subsequence between a candidate summary an the reference is the most widely used.

Other Metrics BERTScore (Zhang et al., 2020) utilizes cosine similarity of BERT (Devlin et al., 2019) embeddings between the reference and the generated summary. CoCo (Xie et al., 2021) computes the difference of likelihood of the summarizer between the summary with the original source and the summary with the masked source.

4.4 Results

Classification Accuracy Due to the imbalance in each dataset, we report the macro-F1 and class balanced accuracy in Table 1 and Table 2. We observe that macro-F1 score of our proposed methods MFMA outperforms baseline entailment metrics in five of seven benchmark datasets. MFMA shows better performances than other methods in especially for CNN/DM benchmarks, and shows similar performance to other baseline in XSum datasets. We explain that this is because we only use training set of CNN/DM to construct training set. On the other hand, DocNLI additionally uses the human annotated datasets from related tasks such as ANLI (Nie et al., 2020) and SQuAD (Rajpurkar et al., 2016) except for synthetic negative summaries. Another proposed method MSM also shows competitive performance for CNN/DM benchmarks, but relatively lower performance in XSum based benchmark datasets. We explain the performance gap between MSM and MFMA is due to the properties that directly generates summaries, resulting in many noisy samples that are relatively easy to be distinguished.

Correlation with Human Judgments To compare with general metrics that are not classification level, we also report the correlation with human judgments for five datasets in Table 3. We demonstrate that our proposed method has higher pearson correlation coefficient with human judgments in three of five benchmark datasets and competitive

with the best results results in the spearman correlation coefficient. Especially, entailment based methods, which are relatively easy to compute, including our proposed methods show better results than QA-based QuestEval or likelihood based CoCo. Also, reference based methods such as ROUGE-L show very lower performance than other methods that do not require any references.

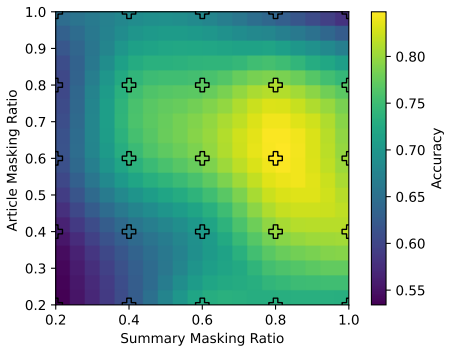

Figure 3: Validation Performance among Masked Ratio for Mask-and-Fill with Masked Article. We experiment with each of the five combinations of article mask ratio and summary mask ratio, and then plot the interpolated results.

4.5 Analysis and Discussion

Performance among Masked Ratio We analyze the effects of the mask ratio for both source article and summary in our proposed method MFMA and present results using the validation set in Figure 3. Through this experiment, we investigate the tradeoff in adjusting both the article masking ratio and summary masking ratio for generating negative summaries. As shown in Figure 3, we find that too high masking ratio decreases performance by sacrificing affinity. On the other hand, if the masking ratio is insufficient, the generated negative sample is often not really negative. In other words, too lower masking ratio leads to generate positive samples that are almost same as the original summary, and this degrades the performance of factual consistency checking model. Also, we can infer that there is an optimal masking ratio combination where the performance of factual consistency checking model is maximized.

Generated Samples among Masking Ratio We visualize the generated negative summaries through our proposed method MFMA and MSM using CNN/DM in Figure 4. We also visualize the example through MF, which simply fills in the mask

Article: Tropical Storm Andrea formed in the Gulf of Mexico on Wednesday, marking the first storm of the 2013 Atlantic hurricane season. The storm was about 270 miles west-southwest of Tampa, Florida, and moving north at 6 mph, said the National Hurricane Center. It had maximum sustained winds of 40 mph. ,..., Thursday afternoon and move over southeastern Georgia and eastern South Carolina Thursday night and Friday. Little change in strength is expected over the next 48 hours.

Reference Summary: Storm moving north at faster rate (6 mph) A storm warning is issued for parts of the Florida coast . Little change in strength is forecast for the next 48 hours . Andrea the first storm of the 2013 Atlantic hurricane season .

MF: Storm moving north at 6 mph (11 kph) A storm warning is issued for parts of the Carolinas. Little change in strength is forecast for the Carolines. Andrea the first storm of the season.

MFMA $\gamma_A = 0.4$**:** Andea moving north at faster rate (6 mph) A storm warning is issued for parts of the Florida coast. Little change in strength is forecast for the next 48 hours. Andea the first storm of the 2013 Atlantic hurricane season.

MFMA $\gamma_A = 0.6$**:** Storm moving north at faster rate (6 mph) Tropical storm watch is issued for west coast of Florida. Little change in strength is forecast for next 48 hours. Storm the first storm of the 2013 Atlantic hurricane season.

MFMA $\gamma_A = 0.8$**:** Andrea moving north at 12 mph (6 mph) A tropical storm warning is issued for parts of the East Coast. A significant increase in strength is forecast for Thursday. Andrea the first storm of the Atlantic hurricane season.

MFMA $\gamma_A = 1.0$**:** Storm moving north at 6 mph (10 kph) A tropical storm watch is issued for the west coast of Florida. A hurricane watch in effect is forecast for the East Coast. The center of the storm forms.

MSM $\gamma_A = 0.4$**:** The storm is about 270 miles west-southwest of Fort Myers, Florida. NEW: The storm is in effect for the west coast of Florida. The storm is the first of the 2013 Atlantic hurricane season.

Figure 4: Generated negative summaries among various masking ratio in CNN/DM dataset. For MFMA and MF, we fix the summary masking $\gamma_S = 0.6$:

without the article. We observe that if the article masking ratio γ_A is too low, the generated summaries become almost similar to the original summary since there are enough information to fill the mask. However, if the γ_A is too high, the generated examples are too far from the article, resulting in too negative summary similar to filling the mask without article.

Table 4: Balanced accuracy of the human annotated factual consistency among masking unit. NP/Ent denotes *noun phrases* and *entities*.

Dataset	Avg-CNN/DM	Avg-XSum
NP/Ent	**73.9**	**60.9**
Token	58.6	53.9
Sentence	53.5	53.4

Performance among Masking Unit We basically perform masking operation in the *noun phrases* and *entities* units for both summary and article. In order to see the effect of the masking

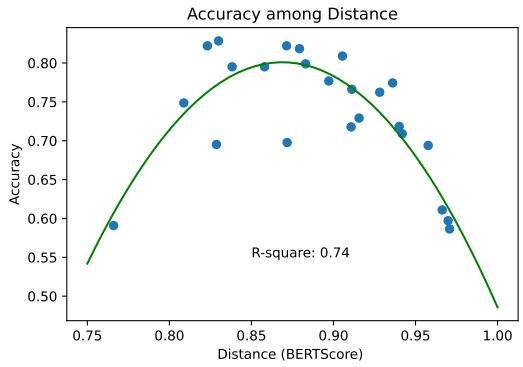

Figure 5: Validation Set Performance among BERTScore between the original reference summaries and the negative summaries we generate using the various combinations of article and summary masking ratios.

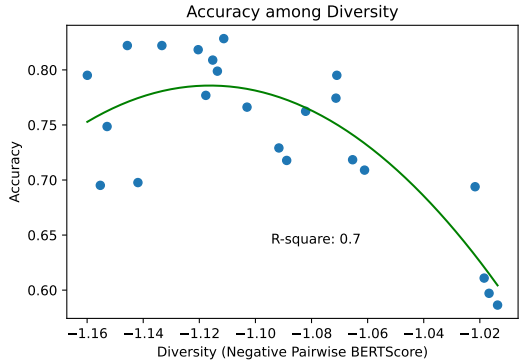

Figure 6: Validation Set Performance among diversity among various combinations of article masking ratio and summary masking ratio. Diversity is computed as negative of the pairwise BERTScore between four negative samples generated by each masking ratio.

unit, we also conduct an experiment on word level masking and sentence level masking, and present the classification level results in Table 4. We observe that *noun phrases* level masking shows the best results following the work (Goyal and Durrett, 2021) where many errors in summarization system are related to *noun phrases* and *entities*.

Distance from Original Reference Summary Using the results on various combinations of article masking ratio and summary masking ratio for MFMA as presented in Figure 3, we also investigate the relation between the average distance from the reference summary on each mask ratio combination and the performance. We compute BERTScore between original reference summary and the negative summary generated using the reference summary to get the distance. Interestingly, as shown in Figure 5, we observe the distribution in

Article: Nkaissery told reporters the university will be able to confirm Saturday if everyone has been accounted for. Thursday's attack by al-Shabaab militants killed 147 people, three security officers and two university security personnel. The attack left 104 people injured, including 19 who are in critical condition, Nkaissery said.,...,

Candidate Summary: 147 people, including 142 students, are in critical condition.

Ground Truth: *INCONSISTENT*
MFMA: *INCONSISTENT*
MSM: *INCONSISTENT*
DocNLI: *INCONSISTENT*
FactCC: *CONSISTENT*

Article: Media playback is not supported on this device United remain 15 points clear at the top of the table with eight games left after a 1-0 win at Sunderland. "We are not concerned with what we have left behind us, we are only focusing on what is in front of us," said Ferguson. ",...,

Candidate Summary: Manchester United manager Sir Alex Ferguson says he is not concerned about his side's unbeaten start to the season as they attempt to win the Premier League title.

Ground Truth: *CONSISTENT*
MFMA: *INCONSISTENT*
MSM: *INCONSISTENT*
DocNLI: *INCONSISTENT*
FactCC: *CONSISTENT*

Figure 7: Case study on entailment based models. First example comes from and FactCC-Test and second example comes from XSumHall.

which performance is maximized within the appropriate distance around 0.8 as the two-dimensional distribution with an R^2 of 0.74. This result shows how far the synthetic negative summaries must be from the reference summaries to help training the factual consistency checking model.

Diversity among Masked Ratio Our proposed method can generate various samples depending on the location of the mask for the same summary-article pair with the fixed mask ratio. Hence, we analyze the diversity of the generated negative summaries among the combinations of mask ratio for MFMA and present the result using validation set in Figure 6. We define the diversity of each mask ratio combination as the negation of pairwise similarity score for each sample following (Tevet and Berant, 2021). We sample four negative summaries using the given article for each method and then compute the pairwise similarity scores for all of the combinations. We also use BERTScore as a similarity measure. Similar to the distance, we observe that diversity has also similar to a two-dimensional form with an R^2 of 0.7, in which the accuracy is maximized at an appropriate point.

Case Study To understand the pros and cons of our proposed factual consistency checking system, we conduct a case study and illustrate the repre-

sentative success and failure cases in Figure 7. We observe that our system is good at judging the facts themselves in the summary like the first example, but still not perfect in examples that require high-level reasoning like the second example. We expect the system can be improved by adopting MFMA and MSM to the datasets that have more abstractive summaries which require more reasoning to check the factual consistency.

5 Conclusion

In this paper, we proposed an effective generation method of factually inconsistent summaries, called MFMA. In this method, some proportion of the source text and corresponding reference summaries is hidden, then a summarization model generates plausible but factually inconsistent summaries by inferring the masked contents. Experiments on seven benchmark datasets demonstrate that factual consistency classifiers trained using our method generally outperform existing models and show a competitive correlation with human judgment.

Ethical Considerations

Our approach creates a synthetic dataset using a public dataset to train a factual consistency checking model. Therefore, in the process of generating such samples, ethically problematic datasets can be generated due to the bias of the pre-trained models, similar to other text generation tasks. For this reason, once the training process is completed, we remove the generated sample. And, we will not release the synthetic dataset itself, and will release only the trained factual consistency checking model.

Acknowledgements

K. Jung is with ASRI, Seoul National University, Korea. This research was supported by SNU-NAVER Hyperscale AI Center.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Rajeshree Bora-Kathariya and Yashodhara Haribhakta. 2018. Natural language inference as an evaluation measure for abstractive summarization. In 2018 4th International Conference for Convergence in Technology (I2CT), pages 1–4. IEEE.

Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 632–642.

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.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2019. Electra: Pre-training text encoders as discriminators rather than generators. In International Conference on Learning Representations.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Wafaa S El-Kassas, Cherif R Salama, Ahmed A Rafea, and Hoda K Mohamed. 2020. Automatic text summarization: A comprehensive survey. Expert Systems with Applications, page 113679.

Alexander R Fabbri, Wojciech Kryscinski, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. Transactions of the Association for Computational Linguistics, 9:391–409.

Tobias Falke, Leonardo FR Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 2214–2220.

Tanya Goyal and Greg Durrett. 2021. Annotating and modeling fine-grained factuality in summarization. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1449–1462.

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

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 9332–9346, Online. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. 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. In Text Summarization Branches Out, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1906–1919.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In Thirty-First AAAI Conference on Artificial Intelligence.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

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.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial nli: A new benchmark for natural language understanding. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4885–4901.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with frank: A benchmark for factuality metrics. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4812–4829.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140):1–67.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. QuestEval: Summarization asks for fact-based evaluation. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 6594–6604, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Guy Tevet and Jonathan Berant. 2021. Evaluating the evaluation of diversity in natural language generation. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 326–346, Online. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018. The fact extraction and VERification (FEVER) shared task. In Proceedings of the First Workshop on Fact Extraction and VERification (FEVER), pages 1–9, Brussels, Belgium. Association for Computational Linguistics.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5008–5020, Online. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1112–1122.

Yuexiang Xie, Fei Sun, Yang Deng, Yaliang Li, and Bolin Ding. 2021. Factual consistency evaluation

for text summarization via counterfactual estimation. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 100–110.

Wenpeng Yin, Dragomir Radev, and Caiming Xiong. 2021. DocNLI: A large-scale dataset for document-level natural language inference. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 4913–4922, Online. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net.

Zheng Zhao, Shay B. Cohen, and Bonnie Webber. 2020. Reducing quantity hallucinations in abstractive summarization. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 2237–2249, Online. Association for Computational Linguistics.

A Experimental Details

A.1 Reproducibility Checklist

Source Code We attach the source in the submission and we will release the pre-trained factual consistency checking model.

Computing Infrastructure We use Intel(R) Xeon(R) Silver 4210R CPU (2.40 GHz) with NVIDIA RTX A5000 24GB for the experiments. The software environments are Python 3.8.8 and PyTorch 1.10.1.

Dataset Statistics We use the training of CNN/DM dataset that consists of 287113 examples. We divide it in half randomly and use one for MSM or MFMA training and the other for generating negative summaries. Then, we merge the generated article-negative summaries pairs and the article-positive summaries we used for training MFMA and MSM to construct the training set for factual consistency checking model.

Average runtime for each approach For training MFMA and MSM, it takes 10 hours to train the whole model. And it takes 3 hours to generate whole negative summaries that is to be used for training factual consistency checking. For training factual consistency checking model, it takes 7 hours using a single GPU.

Hyperparameters We train five epochs for MFMA and MSM using *bart-base* for MFMA and *t5-small* for MSM respectively. We train the model with batch size of 48, max input sequence size of 1024, and max target sequence size of 140. We conduct experiment with various article masking γ_A ratio-summary masking ratio γ_S combinations, at 0.2 intervals from (0.2, 0.2) to (1.0, 1.0). For the case of training classifier, we train *google/electra-base-discriminator* for five epochs with learning rate 2e-5 and batch size of 96. We choose the best parameters using the validation set provided by the (Kryscinski et al., 2020). The best mask ratio combination is γ_A = 0.6 and γ_S = 0.8.

Number of Model Parameters The number of parameters for negative summary generation model is 139M for MFMA, is 0.6M (t5-small) and the factual consistency classifier is 109M.

A.2 Computing Baseline Metrics

Even with the same dataset, the results may be different due to some factors such as type of tokenizer or case, so we calculate baseline ourselves as follows. For n-gram similarity metrics BLEU-4, ROUGE-L and METEOR, we compute the scores using the package *language evaluation*[4] which is based on COCOeval[5]. For BERTScore[6], QuestEval[7] and CoCO[8], we use the official repository with the default setting. For MNLI, we use *roberta-large-mnli*[9] and use tals/albert-base-vitaminc-fever[10] for FEVER.

A.3 Significance Test

We adopt standard way to test the significance of the correlation coefficient for all of the reported related correlation coefficients in Table 3. We compute the p-value for each coefficient with a t-test that uses a null hypothesis, which is an absence of association.

[4] https://github.com/bckim92/language-evaluation
[5] https://github.com/tylin/coco-caption
[6] https://github.com/Tiiiger/bert_score
[7] https://github.com/ThomasScialom/QuestEval
[8] https://github.com/xieyxclack/factual_coco
[9] https://huggingface.co/roberta-large-mnli
[10] https://huggingface.co/tals/albert-base-vitaminc-fever