# Fact-Enhanced Synthetic News Generation

**Kai Shu[‡*], Yichuan Li[♭*], Kaize Ding[†] and Huan Liu[†]**

[‡]Illinois Institute of Technology, Chicago, IL, USA
[♭]Worcester Polytechnic Institute, Worcester, MA, USA
[†]Arizona State University, Tempe, AZ, USA
[‡]kshu@iit.edu, [♭]yli29@wpi.edu, [†]{kaize.ding, huan.liu}@asu.edu

## Abstract

The advanced text generation methods have witnessed great success in text summarization, language translation, and synthetic news generation. However, these techniques can be abused to generate disinformation and fake news. To better understand the potential threats of synthetic news, we develop a novel generation method FACTGEN to generate high-quality news content. The majority of existing text generation methods either afford limited supplementary information or lose consistency between the input and output which makes the synthetic news less trustworthy. To address these issues, FACTGEN retrieves external facts to enrich the output and reconstructs the input claim from the generated content to improve the consistency among the input and the output. Experiment results on real-world datasets demonstrate that the generated news contents of FACTGEN are consistent and contain rich facts. We also discuss an effective defending technique to identify these synthetic news pieces if FACTGEN was used to generate fake news.

## Introduction

With the success of natural language processing, there has been a significant performance improvement in text generation applications, including document summarization (Gehrmann, Deng, and Rush 2018), machine translation (Johnson et al. 2016) and synthetic news generation (Leppänen et al. 2017a). For example, we can use the generative adversarial network (GAN) (Aghakhani et al. 2018) or sequence-to-sequence (seq2seq) model (Yang et al. 2019b) to generate human-like comments. More recently, one approach named Grover (Zellers et al. 2019a) has achieved promising result on synthetic news generation. It generates news pieces conditioned on multiple attributes such as headlines, authors, and website domains.

However, these methods could also be abused to generate and amplify disinformation and fake news. For example, the machine generated fake review threaten business reputations[1]and virtual characters sends generated story to spread propaganda[2]. The wide dissemination of synthetic disinformation and fake news will bring new challenges to the news ecosystem. Therefore, it becomes critical to understand synthetic fake news for further achieving accurate detection.

Table 1: Example claim and the beginning part of a news pieces from CNN/DailyMail dataset. The **black bold** sentence fragments are the consistent word and *Itatic red* fragment is the supplementary information.

| | |
|---|---|
| Claim | iran nuke framework agreement should be judged on merits, not disinformation. |
| Content | The united states and its negotiating partners reached **a very strong framework agreement with iran** in *lausanne , switzerland , on thursday* that limits iran 's nuclear program in such a way as to effectively **block it from building a nuclear weapon**. The debate that has already begun since the announcement of the new framework will likely result in *more heat than light*. It will not be helped by the **gathering swirl of dubious assumptions and doubtful assertions.** |

In the real-world scenario, fake news deliberately imitates the writing styles of real news, which makes it hard to be identified by human and computational detection methods (Shu et al. 2020). Both fake and real news usually contain additional facts[3] that are consistent and supplementary to the news claims. For example, in Table 1 , the news mainly focuses on *framework agreement with Iran*, and provide additional facts like the location and time of the agreement. To eventually identify synthetic disinformation, from an adversarial perspective, we attempt to build a powerful synthetic news generation model by closing the inherent *factual* discrepancies between human and machine-generated text. Existing methods on generating synthetic news may fall short with the following limitations: (1) *factual inconsistency*, indicating the generated news contradict or refute the news claims; and (2) *factual scarcity*, meaning the generated news content may miss essential details to supplement the claim. However, directly using or fine-tuning language models does not help as it is non-trivial to enhance factual consistency and richness on a language model directly. Therefore, in this study, we aim to address the following challenges in synthetic news generation: (1) how to generate news content re-

[*]Equal contributions.
[1]https://bit.ly/349tPW2
[2]https://bit.ly/36if2e9
[3]We follow the definition of `fact` as, according to Oxford Dictionary, the information used as evidence or as part of news article.

lated to a given claim/context; and (2) how to ensure that the generated content contains supplemental fact information.

Our solution to these challenges results in a novel framework FACTGEN[4] (Fact-Enhanced Synthetic News Generation). FACTGEN consists of three major components: (1) *Pseudo-Self-Attentive (PSA) Language Model* (Ziegler et al. 2019), where the customized encoder deceptively injects source information (claims and external facts) into pre-trained decoder for the generation. The adapted deceptive injection mechanism can resolve the mismatch between the untrained encoder and the well-trained decoder; (2) *Fact Retriever*, which heuristically retrieves the supplemental information from external fact corpus to provide more candidate facts during generation; and (3) *Claim Reconstructor*, a randomly initialized masked language model (Devlin et al. 2019) which enhances the output consistency by reconstructing the masked claim tokens from both the representation of the generated content and the unmasked claim tokens. During training, the *PSA Language Model* takes the news claim and the retrieved facts from the *Fact Retriever* as input, then generates highly consistent news content by incorporating the *Claim Reconstructor* into the generation process. In this way, the proposed framework can generate both fact-consistent and fact-enriched news content. To summarize, our main contributions are as follows:

- We study a novel problem of fact-enhanced synthetic news generation, which aims to generate consistent and fact-enriched news content.
- We propose a principled framework FACTGEN generates realistic synthetic news by retrieving external facts and reconstructing the input claim.
- We conduct experiments on real-world datasets using quantitative and qualitative metrics to demonstrate the effectiveness of FACTGEN for synthetic news generation and its defense.

## Methodology

Our goal is to incorporate external facts into news generation that are consistent with the news claim. Given a sequence of tokens from the claim $X = \{x_1, x_2 \ldots, x_N\}$, the fact retriever retrieves related fact information $F = \{f_1, f_2, \ldots, f_K\}$ by semantic similarity, then the language model generates the news content $Y = \{y_1, y_2 \ldots, y_M\}$ based on claim $X$ and $F$. It should be noticed that the length of $Y$ is much larger than $X$, which is $M >> N$, and $x_i, y_i, f_i$ are words. Figure 1 illustrates the architecture of the proposed model and the objective functions. The causal language loss $L_{CLL}$ depicts the loss of generating news content based on the input claim and fact. The masked language loss $L_{MLL}$ is to reconstruct the masked input claim based on the language model output and the unmasked claims. This has a twofold benefit. Initially, the pre-trained decoder and the retrieved facts will bring unrelated information. This technique encourages the generated content to cover the input claim and provides a regularization effect. In addition,

[4]The code is available at https://github.com/bigheiniu/FactGen

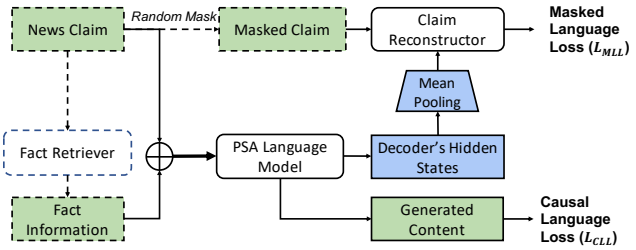


Figure 1: The proposed model, FACTGEN. The black dashed line indicates no differential dependency and the black bold line otherwise. $\oplus$ is the text concatenation.

it is fully differentiable so we can minimize the objective function end-to-end. Overall, we minimize:

$$L = L_{CLL} + \lambda\, L_{MLL} \tag{1}$$

where $\lambda$ is the hyperparameter to control the contribution of claim reconstruction. The formulas of $L_{CLL}$ and $L_{MLL}$ are in Eq. 5 and Eq. 6 respectively.

### Preliminary

Self-attentive (SA) language models (Devlin et al. 2018; Radford et al. 2019; Song et al. 2019) have achieved impressive performance gains in various language generation tasks. These models are stacks of several SA blocks which encode the input $X = \{x_1, \ldots, x_i, \ldots, x_N\}$ into key-value pairs $(K, V) = \{(k_1, v_1), \ldots, (k_i, v_i), \ldots, (k_N, v_N)\}$ and query $Q = \{q_1, \ldots, q_i, \ldots, q_N\}$. The next output is produced by taking the weighted sum of values $v_i$, where the weight assigned toward each value is the dot-product of the query $Q$ with all the keys $K$. The formula of SA is:

$$K = H_X W_k,\ \ V = H_X W_v,\ \ Q = H_X W_q \tag{2}$$

$$SA(X) = \text{softmax}\left(QK^T\right) V \tag{3}$$

where $H_X \in \mathbb{R}^{N \times D}$ is the hidden representation of the input $X$, $D$ is the hidden dimension, and $W_k, W_v, W_q \in \mathbb{R}^{D \times D}$ are the parameters to map the hidden representation of tokens $H_X$ into key, value and query space, respectively.

### Proposed Method

**Pseudo-Self-Attentive Language Model:** Although the fine-tuned self-attentive language models like GPT-2 (Radford et al. 2019) have been applied to many text generation tasks, the application of using GPT-2 for the synthetic news generation may not be satisfactory. Since the GPT-2 is an autoregressive model, only encoding the forward information, it will lose backward information from the input. Besides, without a specific encoder, GPT-2 cannot capture the dependent relationship between the news claim and the retrieved facts, which will hurt the performance of the decoder (Edunov, Baevski, and Auli 2019). Therefore, we need a new encoder to capture bi-directional information and dependency among the input.

Following (Ziegler et al. 2019)'s setting, we employ a pseudo-self-attentive(PSA) language model, where the

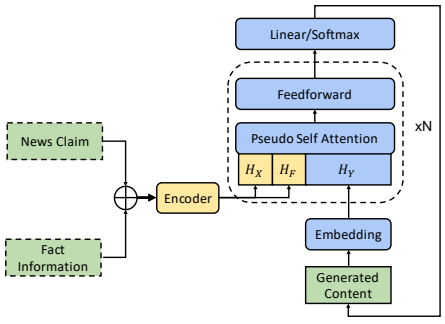


Figure 2: Our pseudo-self-attentive language model. Best visualized in color. The blue indicates decoder's pre-trained parameters. The yellow indicates the randomly initialized parameters of the encoder. $N$ is the number of PSA blocks.

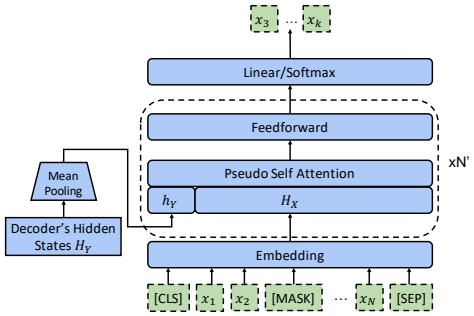


Figure 3: The overview of the claim reconstructor. It is also the PSA structure where we inject the mean pooling of decoder's hidden states to its key-value pairs. $N'$ is the number of PSA blocks.

"pseudo" is that the encoder deceptively extends the decoder's key-value pairs by the encoder's pairs, and the decoder predicts the next token not only based on previous output tokens but also from the input. To model the dependency between claim and retrieved facts, we wrap them with "[Claim]" and "[Fact]" separately, and specifically all the retrieved facts are contacted together without any special separation token. The architecture of the language model is shown in Figure 2 and the formula of PSA is:

$$\text{PSA}(Y, X, F) = \text{softmax}\left(Q_Y \begin{bmatrix} K_Y \\ K_X \\ K_F \end{bmatrix}^\top\right) \begin{bmatrix} V_Y \\ V_X \\ V_F \end{bmatrix} \quad (4)$$

Note that $K_X, K_F, V_X$, and $V_F$ are using different projection matrix $W_*$ and are randomly initialized. The objective function of the language model is:

$$L_{CLL} = -\sum_{i=1}^{M} (log\ P(y_i|y_1, \ldots, y_{i-1}; X, F)) \quad (5)$$

**Fact Retriever:** Directly training a sequence to sequence model on $(X, Y)$ often results in fact scarcity. One main reason is that facts from the input are extremely insufficient compared to the output. Thus, the language model is more likely to generate repeated sentences. Our solution towards the facts imbalance between the input and output is to increase the facts in the source side by retrieving related facts and considers them as part of the input. Our fact retriever (FR) heuristically retrieves external facts in two steps. Firstly, to omit the computation limitation, we retrieve the related document based on the tf-idf vectors' cosine similarity between the claim and the document. Here we only keep the $top{-}k_1$ similar documents. Secondly, to accurately identify related sentences in the document, we utilize the pre-trained BERT (Devlin et al. 2018) to encode all sentences presented in the picked documents and choose the $top - k_2$ most similar sentences based on the cosine similarity.

**Claim Reconstructor:** Since the aforementioned modules FR and PSA language model will bring inconsistency during the generation, there still needs an extra mechanism to guarantee the consistency between the input news claims and the generated news content. We propose to reconstruct the claim from the generated content through masked language model.

The existing reconstruction approaches for the consistent generation require the prior knowledge of the input, such as utilizing the topic label to learn a topic consistent reward function (Yang et al. 2019a), or key-entities for multi-classification on the hidden states to entail the key information (Wiseman, Shieber, and Rush 2017). Our claim reconstructor (CR) does not require any prior knowledge about the input. It reconstructs the masked claim $X_{[Masked]}$ based on the mean pooling of output hidden representation $h_Y$ and unmasked sentence fragments $X_{[Unmasked]}$ . We mask claim's tokens with probability $P_{mask}$ and we follow the pseudo-self-attention (Ziegler et al. 2019) projecting $h_Y$ into CR's key-value pairs to predict the masked sentence fragment, $X_{[Masked]}$. The objective function of CR is:

$$L_{MLL} = \sum_{x \in X_{[Masked]}} -log\ P(x|X_{[Unmasked]},\ h_Y)) \quad (6)$$

## Training Schedule

Since FACTGEN needs to guarantee that there is no contradiction between the factual consistency and factual richness, we cannot directly train the model via minimizing eq 1. We then train FACTGEN in two-stages. The overview of the training procedure is summarized in Algorithm 1. The two stages of training bring several advantages: firstly, it allows to start the PSA language model and CR warmly, omitting the gradient explosion problem during training; secondly, because the claim is the main idea of the generated text and the retrieved facts are the auxiliary information during the generation, this order can help the decoder understand the importance of different input sources. The joint training can align the latent space of these two modules.

# Experiments

In this section, we conduct experiments on two real-world news dataset to demonstrate the effectiveness of FACTGEN for news generation. Specifically, we aim to evaluate the quality of the generated news pieces in terms of *fluency*, *Consistency*, *richness*, and *trustworthiness*.

**Algorithm 1** Training Procedure of FACTGEN

**Input:** The source claims, relevant facts and target news pieces corpus $S = \{(X, F, Y)\}$; the masked and unmasked claims $D = \{(X_{[Masked]}, X_{[Unmasked]})\}$; first and second stage epoch number $epochs_1$ and $epochs_2$.

**Output:** $PSA$ language model and claim reconstructor $CR$;

```
Initialize PSA_encoder and CR with random weights
Pre-train the PSA via minimizing eq.5 on {(X, Y)};
Pre-train the CR via minimizing eq.6 on D.
for epoch = 1 to epochs_1 do
    Jointly training PSA and CR via minimizing eq.1 on
    {(X, Y)} and D;                          ▷First Stage
end for
for epoch = 1 to epochs_2 do
    Jointly training PSA and CR via minimizing eq.1 on
    S and D;                                 ▷Second Stage
end for
```

Table 2: The statistical information of the datasets.

| Dataset | # of train | # of val | # of test |
|---|---|---|---|
| GossipCop | 7,331 | 1,459 | 974 |
| CNN/DailyMail | 278,408 | 11,490 | 13,368 |

## Dataset

We utilize two news dataset in our experiment. The first dataset is a widely used fake news detection dataset collected from a fact-checking website, GossipCop (Shu et al. 2018). Each sample contains the news' claim, content, metadata, label, and social engagements. The average lengths of the claim and content are 30 words and 250 words respectively. The second dataset is the CNN/DailyMail news highlight dataset (Hermann et al. 2015) which contains the news content and selected highlight. In contrary to the text summarization, we use the highlight sentence as the source claim and the news content as the target text. On average, the claim has 56 tokens and the content has 790 tokens. As for prepossessing, we truncate the news claim longer than 100 words and content longer than 300 words in both datasets. For the dataset splitting, we randomly sample 75% training set, 15% validation set, and 10% test set in the GossipCop dataset and follow the same splitting setting in (See, Liu, and Manning 2017). Datasets' statistical information is listed in Table 2.

In this paper, we consider the factual sentences in the training dataset as our external fact corpus. This brings two advantages: firstly, utilizing several sentences instead of whole news pieces can avoid the model learning from copy the information from the source to the target side; secondly, the fact sentences from the training dataset can omit the data leakage problem during testing. Here we focus on external facts as the format of the text, though it can be extended to tabular data or knowledge graph.

## Experiment Settings

We implement FACTGEN on OpenNMT (Klein et al. 2017). We tune the hyper-parameter $\lambda$ on the validation set. The encoder of FACTGEN is 4 blocks of SA block with 12 attention heads and 3072 hidden units. The weight of the decoder is initialized with the median pre-trained GPT-2 (Radford et al. 2019) model. The claim reconstruction module is 3 blocks of SA block with 4 attention heads and 256 hidden sizes and $P_{mask}$ is set as 0.5. The optimizer is Adam (Kingma and Ba 2014) with $\beta_1 = 0.9$ and $\beta_2 = 0.998$. It should be noticed that the learning rate for the encoder is $1e-3$, for the decoder is $1e-5$, and $5e-5$ for the claim reconstruction. The number of retrieved documents $k_1$ and sentences $k_2$ is set to 10 and 5 respectively. The $epochs_1$ and $epochs_2$ in the training schedule are set to 4 and 2 respectively. During decoding we used Nucleus Sampling (top-*p*) with $p = 0.9$.

## Evaluation Metrics

**Automatic Evaluation** The traditional text generation metrics like BLEU (Papineni et al. 2002) and ROUGE (Lin 2004) which are focus on the overlap between the generated content and the reference text which is not enough to reflect the claim-content consistency and the richness of the generated content. To remedy this, we develop two new evaluation metrics to measure the quality from different perspectives.

- **Fluency:** we report the BLEU-4 score for the text fluency.
- **Consistency:** The ideal news content should support its claim. Therefore, we propose a stance detection model to detect whether the content is in favor of the claim or against it. Given the claim and the generated news content $\{X, Y\}$, the stance detection model will output the relation of the text pair in *(Agrees, Disagrees, Discusses, Unrelated)*.We utilize the Fake News Challenge dataset[5] to fine-tune RoBERTa (Liu et al. 2019). This approach achieves a 0.93 accuracy score on the test dataset of the Fake News Challenge. We report the ratio of the "agrees":

$$Consistency = \frac{\#\ of\ agree\ samples}{\#\ of\ all\ samples} \tag{7}$$

- **Richness:** The richness of the output can be evaluated by the number of unique name entities in the generated text (Fan, Lewis, and Dauphin 2019). We utilize spaCy[6] to extract the named entity from the output.

**Human Evaluation** We distribute the 100 generated samples in CNN/DailyMail dataset to 2 annotators with a linguistic background. They have no advanced knowledge about the source of the generated content. They are asked to evaluate the generated content from fluency, richness, consistency, and Trustworthiness, 4 different perspectives[7]. So totally, there are 7,200 evaluation questions in our human evaluation. The annotator should answer each question from a score of 1 to 3 (3 being the best, 1 being the worst).

## Baseline Methods

To demonstrate the quality of the generated text, we compare our proposed model on content quality with the following text generation models:

[5] http://www.fakenewschallenge.org/

[6] https://spacy.io/

[7] The details of human evaluation questions are in Appendix.

- CopyTransformer (See, Liu, and Manning 2017): a sequence-to-sequence transformer with a pointer network that can copy the word from the source to the target.
- Conv Seq2Seq (Fan, Lewis, and Dauphin 2018): a seq2seq convolution neural network to generate claim consistent stories.
- PPLM (Dathathri et al. 2019): a topic and content controlled language model. It can directly control the pre-trained language model without fine-tuning.
- GPT-2 (Radford et al. 2019): a large pre-trained language model which is the decoder part of the transformer. For a fair comparison, we utilize the median size of the model.
- Grover (Zellers et al. 2019a): generating news text conditioned on the news title, authors, and website domains. It uses the same architecture as GPT-2.

## Experimental Results

The automatic and human evaluation results are shown in Table 3 and 4, respectively. We evaluate the quality of the text generation through the following perspectives:

- **Fluency:** From the human evaluations on fluency in CNN/DailyMail dataset and fluency score in two datasets, we can find that our model achieves the best performance. In the meantime, we find that the pre-trained language model achieves better human evaluation results than the model trained from scratch (PPLM, GPT-2, Grover $>$ CopyTransformer, Conv Seq2Seq). This indicates the importance of incorporating the large pre-trained language model in the synthetic news generation. Besides, FACTGEN's performance indicates the pseudo-self attention properly connecting the randomly initialized encoder and the pre-trained decoder.
- **Consistency:** The consistent result in both human evaluation and automatic evaluation demonstrates the effectiveness of our approach. Especially, in the GossipCop dataset, our approach achieves 42% performance improvement over the best baseline in automatic metric, and in CNN/DailyMail, the human evaluation also shows that our approach achieves 6% performance improvement compared with the best baseline method. The main reason for the increase is reconstructing the claim increases the coverage of the output on the input information.
- **Richness:** Our approach achieves the best performance in CNN/DailyMail dataset and the second performance in the GossipCop dataset. The reason for the ordinary performance in GossipCop is that the size of the candidate documents in GossipCop is much smaller than the CNN/DailyMail (7,331 $<$ 278,408). The FR cannot retrieve enough related facts from the external corpus and CR will reject the inconsistent facts during generation. This indicates that FR can bring rich facts in generation.
- **Trustworthiness:** Human evaluation of the Trustworthiness of synthetic news content indicates that overall, FACTGEN can generate high-quality text content. This helps us to understand the difference between machine-generated news content and true news in the future.

## Case Study

One case study of the generated samples is listed in Table 5. We only reveal the output from the model with pre-trained language models and we have several observations: *(i)* Our model mainly talks about the *agreement of nuclear weapons in Iran* and includes the supplemental information about *Iraq and UK's* action toward *nuclear weapons*. This brings more context information about the news claims and makes the generated news more convincing. *(ii)* Although Grover mentions much additional factual information, it is unrelated to the *nuclear agreement* with *Iran*. *(iii)* The outputs of GPT-2 and PPLM mainly discuss the *nuclear agreement* without supplemental information about the agreement.

## Ablation Study

**Impact of $\lambda$:** To learn the impact of the hyper-parameter $\lambda$ in our objective function in Eq. 1, we change $\lambda$ from $\{0.001, 0.01, 0.1, 1, 10\}$ and calculating all the automatic evaluation metrics. From Figure 4 we can find that $\lambda = 0.001$ achieves the best performance across all the automatic evaluations and with the increase of $\lambda$, the fact richness has been greatly decreased. This is because the CR will constrain the coverage of the generated content and cause the language model to only generated content around the input, which will reduce the richness of the generated content.

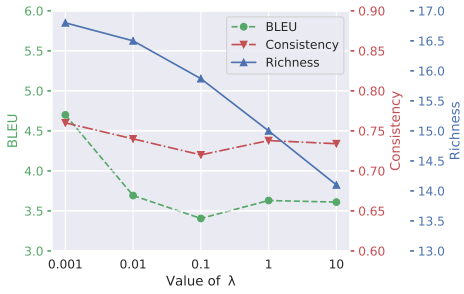


Figure 4: Impact of hyper-parameter $\lambda$ in CNN/DailyMail.

**Impact of Model Components:** To evaluate the importance of each key components, we set up three different ablation studies of FACTGEN: without claim reconstruction(*w/o CR*), without Fact Retriever (*w/o FR*) and without these two components (*w/o CR and FR*). It should be noticed that all versions of the model have been pre-trained on $\{X, Y\}$. The automatic and human evaluation in Table 6 and Table 7 show that the performance decrease in all ablation study. However, an interesting finding is that there seems to have a contradiction between the CR and FR. From Table 6, we find that *w/o CR* contains the richest fact information but has the lowest consistency score; *w/o FR* achieves the best fluency score and compatible consistency score but the worst richness score. The impact of CR matches the observation of hyperparameter analysis, which improves the consistency of the generated content while decreases the fact richness. These results indicate the effectiveness of CR and FR in improving the richness and consistency in the generation.

Table 3: The performance comparison for the quality of the generated news pieces.

| Models | GossipCop | | | CNN/DailyMail | | |
|---|---|---|---|---|---|---|
| | Fluency | Richness | Consistency | Fluency | Richness | Consistency |
| CopyTransformer | 0.2 | 11.0 | 0.04 | 0.5 | 9.5 | 0.66 |
| ConvSeq2seq | 0.5 | 5.9 | 0.09 | 3.3 | 9.5 | 0.44 |
| PPLM | 0.7 | 12.5 | 0.67 | 0.8 | 13.1 | 0.68 |
| GPT-2 | 0.8 | 13.4 | 0.35 | 1.65 | 13.5 | 0.70 |
| Grover | 1.2 | **15.7** | 0.56 | 0.3 | 15.3 | 0.72 |
| **FACTGEN** | **2.1** | 14.5 | **0.80** | **4.6** | **16.6** | **0.76** |

Table 4: The human evaluation result of generated samples in the CNN/DailyMail dataset. We calculate the Pearson correlation to show the inter-annotator agreement.

| Methods | Fluency | Richness | Consistency | Trustworthiness |
|---|---|---|---|---|
| CopyTransformer | 1.68 | 1.65 | 1.89 | 1.62 |
| ConvSeq2seq | 1.95 | 2.12 | 2.00 | 1.94 |
| PPLM | 1.96 | 1.77 | 1.96 | 1.92 |
| GPT-2 | 2.03 | **2.32** | 1.95 | 2.08 |
| Grover | 2.08 | 2.15 | 1.78 | 1.97 |
| **FACTGEN** | **2.17** | 2.28 | **2.12** | **2.18** |
| Correlation | 0.14 | 0.26 | 0.21 | 0.21 |

**Impact of Training Schedule:** To understand the effectiveness of our two-stage training schedule, we compare it with single-stage training where the model directly takes the claims and external fact information in the first stage. From the automatic and human evaluation result in Table 6 and 7, we can find that two stages training schedule achieve better performance in all categories compared with single-stage. This stipulates the effectiveness of our training schedule.

## Further Analysis

### Difficulty of Defending Synthetic Fake News

To understand the difficulty in synthetic fake news detection, we test the fake news detection methods and synthetic generation detection method on generated fake news and human-written real text. To guarantee the veracity of the test content, we select the fake generated content which is conditioned on fake claim and human-written real text is from real news pieces in GossipCop. To understand the difficulty in synthetic fake news detection, we test the fake news detection methods and synthetic generation detection method on generated fake news and human-written real text. To guarantee the veracity of the test content, we select the fake generated content which is conditioned on fake claims and human-written real text is from real news pieces in GossipCop. The reason for different training datasets for these approaches is to test whether the fake news detection model can transfer the knowledge in human written fake news into machine-generated fake news. To give limited access to generated content, the training dataset for both approaches will include extra 100 fake synthetic news pieces. We test the classification accuracy in 300 fake generated news contents and the same amount of human-written real news content. To omit the data leakage problem for evaluation, the test dataset is also the test data for synthetic generation evaluation. From the result in Table 8, we observe that fake news detection methods achieve worse performance than neural text classification (RoBERTAa $>$ EANN, MWSS-CNN) which indicates the difficulty of the current fake news detection method in detecting fake synthetic news.

### Defending Against Synthetic Fake News

To detect the new synthetic fake news, we follow (Zellers et al. 2019b) develop a defending method FACTGEN$_{def}$ based on the checkpoint of FACTGEN at iteration 20k. This setting can reduce the parameters overlap between the generator and the discriminator. We also use $h_Y$ as the final representation of the input, synthetic fake news or human written real news, and add a full connection layer to classify whether the input is fake or real. We utilize 100 synthetic fake news content and the same amount of human written real news to fine-tune FACTGEN$_{def}$. The result in Table 8 shows FACTGEN$_{def}$ achieves the best accuracy score. This is because FACTGEN$_{def}$ can learn a better representation of the input. We thus conclude that while the synthetic content is hard to be identified by existing methods, it still can be detected by FACTGEN$_{def}$.

## Related Work

**Synthetic News Generation** Most synthetic news generation systems used in the newsroom are heavily rule-based and template-based (Leppänen et al. 2017a). The neural synthetic news generation like Grover (Zellers et al. 2019a) utilizes an autoregressive language model to learns the dependency among news metadata fields include the domain, date, authors, title, and body. Sam and et. al (Wiseman, Shieber, and Rush 2017) propose a structured data to text challenge which is to generate a sport a news piece of sports games from the associated box- or line-score data. To better capture the input data, (Wiseman, Shieber, and Rush 2017) employs copy-mechanism and source reconstruction as their seq2seq model extensions and (Puduppully, Dong, and Lapata 2018) generate text in recording plan and realization two stages.

**Synthetic/Fake News Detection** Content-based fake news detection methods often leverage features from the feature engineering or latent features extracted by deep neural network (Pérez-Rosas et al. 2017). The deep learning models utilize the linguistic representation of news content to detect fake news. Qian *et al.* (Qian et al. 2018) proposes a method

Table 5: Samples of generated fake news with different fake news generation methods. The **black bold** sentence fragment is the consistent word and *Italic red* fragment is the supplementary information.

| | |
|---|---|
| Claim | iran framework agreement on nukes . . . opponents cast doubts. . . . uranium stockpile, centrifuges, implement rigorous inspections; . . . judged on merits, not disinformation. |
| PPLM | i think the agreement has been agreed on the whole of *American and nauru*,and is a good first step towards an **agreement** on other issues. . . . cut its **stockpiles** and implement some strict inspections. i think this is a good first step towards an **agreement**.i think *the deal is good but its not done yet and it still needs to be signed and accepted by america*,and it will have to meet certain conditions for american to accept its deal on **nuclear** |
| GPT-2 | The **iran nuclear** program does not have a single day's notice of a nuclear crisis, and it does not have a single day's notice of a **nuclear weapon** program. . . . Last week, the government of **iran** signed a framework agreement that *allows* the *permanent establishment of the nuclear program* at all levels . . . close its existing **enrichment plant**, . . . cooperation and the reduction of the **stockpile**. |
| Grover | Unlike (allegedly) other *traitors who've betrayed the established foreign policy approach* to the Iranian nuclear development , in addition to *erasing himself from IMDB*, with *a stigmata, an air cannon*, . . . published his own book for reasons which, honestly, I'm not sure are entirely clear. The main key player in the **NPT negotiations** |
| FACTGEN | billionaire businessman **fichard klass** has joined the *british establishment* to try to prevent the sale of **nukes** after **failing to address its facts** on how the **iran regime** used them to try to arm its **nuclear programme**. *the businessman, who once ran a world bank and was governor of iraq*, said the nukes would cut **uranium stockpile**, **implement a rigorous inspections** . . . and try to defend its nuclear program. while **opponents** said a *strong nuclear test would help persuade the uk to develop nuclear weapons*, **klass** said he hoped most sanctions would be **judged on merits**, **rather than disinformation** . |

Table 6: Results of automatic evaluation of model components ablation study in CNN/DailyMail dataset.

| Methods | Fluency | Richness | Consistency |
|---|---|---|---|
| Full Model | 4.7 | 16.8 | **0.76** |
| *single-stage* | 4.6 | 16.6 | 0.74 |
| *w/o CR* | 3.4 | **17.7** | 0.73 |
| *w/o FR* | **5.1** | 11.7 | 0.75 |
| *w/o CR and FR* | 4.0 | 12.9 | 0.73 |

Table 7: Results of human evaluation of model components ablation study in CNN/DailyMail dataset.

| Methods | Fluency | Richness | Consistency | Trustworthiness |
|---|---|---|---|---|
| Full Model | **2.17** | 2.28 | **2.12** | **2.18** |
| *single-stage* | 2.01 | 2.22 | 2.10 | 2.14 |
| *w/o CR* | 2.15 | **2.31** | 2.09 | 2.15 |
| *w/o FR* | 1.93 | 2.28 | 2.03 | 1.98 |
| *w/o CR and FR* | 2.09 | 2.19 | 2.03 | 2.10 |

Table 8: Results of synthetic fake news content detection.

| | EANN | MWSS-CNN | RoBERTa | FACTGEN$_{def}$ |
|---|---|---|---|---|
| Accuracy | 0.64 | 0.58 | 0.74 | **0.82** |

learning the representation of news content and reconstructing the users comment during training, and in inference, this model makes a classification based on the representation of news content and the generated news comment for early fake news detection. Tal Schuster (Schuster et al. 2020) stipulates that current synthetic disinformation detection methods are mainly based on the stylometry which is limited against machine-generated misinformation. Gehrmann *et al.* (Gehrmann, Strobelt, and Rush 2019) visualize the distribution of words that help non-expert users recognize generated text. (Zellers et al. 2019a) and (Solaiman et al. 2019) propose neural generation detectors that fine-tune classifiers on the generator's previous checkpoint. (Uchendu et al. 2020) try to identifying the NLP method of the generated text.

## Conclusion and Future Work

We propose a synthetic news generation method FACTGEN to ensure fact-consistency and fact-richness. From the automatic and human evaluation of the content quality, FACTGEN is more effective than existing methods. Simultaneously, we discuss the difficulty of detecting synthetic fake news content by current SOTA fake news and human-machine detection methods. For social good, we propose a defending method FACTGEN$_{def}$ that achieves outstanding performance in detecting synthetic fake news content. In the future, we would like to include other formats of facts like tabular or knowledge graph. This can help us retrieve up-to-date fact information during generation. Since fake news often contains propaganda and more likely to widely spread on the social network, we would like to explore the style control of the generated content to make it prone to be spread.

## Acknowledgments

This work is, in part, supported with funding from the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001120C0123, and the John S. and James L. Knight Foundation through a grant to the Institute for Data, Democracy Politics at The George Washington University. The views, opinions and/or findings expressed are those of the author and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

## References


[Adair et al. ] Adair, B.; Li, C.; Yang, J.; and Yu, C. Automated pop-up fact-checking : Challenges & progress.

[Aghakhani et al. 2018] Aghakhani, H.; Machiry, A.; Nilizadeh, S.; Kruegel, C.; and Vigna, G. 2018. Detecting deceptive reviews using generative adversarial networks.

[Dathathri et al. 2019] Dathathri, S.; Madotto, A.; Lan, J.; Hung, J.; Frank, E.; Molino, P.; Yosinski, J.; and Liu, R. 2019. Plug and play language models: a simple approach to controlled text generation.

[Devlin et al. 2018] Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

[Devlin et al. 2019] Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

[Edunov, Baevski, and Auli 2019] Edunov, S.; Baevski, A.; and Auli, M. 2019. Pre-trained language model representations for language generation.

[Fan, Lewis, and Dauphin 2018] Fan, A.; Lewis, M.; and Dauphin, Y. 2018. Hierarchical neural story generation.

[Fan, Lewis, and Dauphin 2019] Fan, A.; Lewis, M.; and Dauphin, Y. 2019. Strategies for structuring story generation.

[Gehrmann, Deng, and Rush 2018] Gehrmann, S.; Deng, Y.; and Rush, A. M. 2018. Bottom-up abstractive summarization.

[Gehrmann, Strobelt, and Rush 2019] Gehrmann, S.; Strobelt, H.; and Rush, A. M. 2019. Gltr: Statistical detection and visualization of generated text.

[Hermann et al. 2015] Hermann, K. M.; Kočiský, T.; Grefenstette, E.; Espeholt, L.; Kay, W.; Suleyman, M.; and Blunsom, P. 2015. Teaching machines to read and comprehend.

[Johnson et al. 2016] Johnson, M.; Schuster, M.; Le, Q. V.; Krikun, M.; Wu, Y.; Chen, Z.; Thorat, N.; Viégas, F.; Wattenberg, M.; Corrado, G.; Hughes, M.; and Dean, J. 2016. Google's multilingual neural machine translation system: Enabling zero-shot translation.

[Kingma and Ba 2014] Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization.

[Klein et al. 2017] Klein, G.; Kim, Y.; Deng, Y.; Senellart, J.; and Rush, A. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *ACL*.

[Leppänen et al. 2017a] Leppänen, L.; Munezero, M.; Granroth-Wilding, M.; and Toivonen, H. 2017a. Data-driven news generation for automated journalism. In *PINLG*.

[Leppänen et al. 2017b] Leppänen, L.; Munezero, M.; Granroth-Wilding, M.; and Toivonen, H. 2017b. Data-driven news generation for automated journalism. In *Proceedings of the 10th International Conference on Natural Language Generation*, 188–197. Santiago de Compostela, Spain: Association for Computational Linguistics.

[Lin 2004] Lin, C.-Y. 2004. ROUGE: A package for automatic evaluation of summaries. In *ACL*.

[Liu et al. 2019] Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach.

[Papineni et al. 2002] Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: A method for automatic evaluation of machine translation. In *ACL*.

[Puduppully, Dong, and Lapata 2018] Puduppully, R.; Dong, L.; and Lapata, M. 2018. Data-to-text generation with content selection and planning.

[Pérez-Rosas et al. 2017] Pérez-Rosas, V.; Kleinberg, B.; Lefevre, A.; and Mihalcea, R. 2017. Automatic detection of fake news.

[Qian et al. 2018] Qian, F.; Gong, C.; Sharma, K.; and Liu, Y. 2018. Neural user response generator: Fake news detection with collective user intelligence. In *IJCAI*.

[Radford et al. 2019] Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language models are unsupervised multitask learners.

[Schuster et al. 2019] Schuster, T.; Schuster, R.; Shah, D. J.; and Barzilay, R. 2019. The limitations of stylometry for detecting machine-generated fake news.

[Schuster et al. 2020] Schuster, T.; Schuster, R.; Shah, D. J.; and Barzilay, R. 2020. The limitations of stylometry for detecting machine-generated fake news.

[See, Liu, and Manning 2017] See, A.; Liu, P. J.; and Manning, C. D. 2017. Get to the point: Summarization with pointer-generator networks.

[Shu et al. 2018] Shu, K.; Mahudeswaran, D.; Wang, S.; Lee, D.; and Liu, H. 2018. Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. *arXiv preprint arXiv:1809.01286*.

[Shu et al. 2020] Shu, K.; Bhattacharjee, A.; Alatawi, F.; Nazer, T. H.; Ding, K.; Karami, M.; and Liu, H. 2020. Combating disinformation in a social media age. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10(6):e1385.

[Solaiman et al. 2019] Solaiman, I.; Brundage, M.; Clark, J.; Askell, A.; Herbert-Voss, A.; Wu, J.; Radford, A.; Krueger, G.; Kim, J. W.; Kreps, S.; McCain, M.; Newhouse, A.; Blazakis, J.; McGuffie, K.; and Wang, J. 2019. Release strategies and the social impacts of language models.

[Song et al. 2019] Song, K.; Tan, X.; Qin, T.; Lu, J.; and Liu, T.-Y. 2019. Mass: Masked sequence to sequence pre-training for language generation.

[Uchendu et al. 2020] Uchendu, A.; Le, T.; Shu, K.; and Lee, D. 2020. Authorship attribution for neural text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 8384–8395. Online: Association for Computational Linguistics.

[Wiseman, Shieber, and Rush 2017] Wiseman, S.; Shieber, S. M.; and Rush, A. M. 2017. Challenges in data-to-document generation.

[Yang et al. 2019a] Yang, P.; Li, L.; Luo, F.; Liu, T.; and Sun,

X. 2019a. Enhancing topic-to-essay generation with external commonsense knowledge. In *ACL*.

[Yang et al. 2019b] Yang, Z.; Xu, C.; Wu, W.; and Li, Z. 2019b. Read, attend and comment: A deep architecture for automatic news comment generation.

[Zellers et al. 2019a] Zellers, R.; Holtzman, A.; Rashkin, H.; Bisk, Y.; Farhadi, A.; Roesner, F.; and Choi, Y. 2019a. Defending against neural fake news.

[Zellers et al. 2019b] Zellers, R.; Holtzman, A.; Rashkin, H.; Bisk, Y.; Farhadi, A.; Roesner, F.; and Choi, Y. 2019b. Defending against neural fake news. *arXiv preprint arXiv:1905.12616*.

[Ziegler et al. 2019] Ziegler, Z. M.; Melas-Kyriazi, L.; Gehrmann, S.; and Rush, A. M. 2019. Encoder-agnostic adaptation for conditional language generation.

## Appendices on Reproducibility

In this section, we provide more details about the human evaluation questions, experimental settings and hyperparameter configuration to enable the reputability of our work.

### Human Evaluation Question

To evaluate the quality of FACTGEN, we ask human workers to answer four different questions. For each question, human worker need to give a score from 1 to 3 (1 means low quality and 3 is high quality).

- (Fluency) Is the output article written by human?
- (Richness) Does the text provide extra information not listed in the input claim?
- (Consistency) Is the output consistent with the input?
- (Trustworthiness) Do you trust the output content?

### Synthetic News Generation

In Section , we compare FACTGEN with 5 baseline methods, including Conv Seq2Seq, CopyTransformer, PPLM, fine-tuned GPT-2 and Grover.

For the dataset, GossipCop is available in the dataset section of the submission and CNN/DailyMail is available at [8]. The description of the Fact Retriever as follows:

- $top - k_1$: the number of documents we retrieved based on the tf-idf cosine similarity, we rank it from the biggest to the smallest. We set $top - k_1$ to 10 for both datasets.
- $top - k_2$: the number of sentences we retrieved from the $top - k_1$ Documents. We set it 5 for both datasets.

The parameters for evaluating the semantic similarly are the same as RoBERTa.

### Synthetic News Detection

The two fake news detection models EANN, MWSS can be obtained online. *EANN*: it is publicly available at https://github.com/yaqingwang/EANN-KDD18. *MWSS-CNN*: it is avaliable at https://github.com/microsoft/MWSS.

The hyper parameters for human written and machine generated content detection RoBERTa are as follows:

- Epochs: fine-tuning Epochs the RoBERTa model, 10.
- Patience: the number of epochs to wait before early stop if no progress on the validation set, 3.
- Batch size: number of samples in one iteration, 10.
- Learning Rate: model fine-tuning learning rate, 5e-5.
- Max Length: we pad the input sentence into 300 tokens.

## Appendices on Ethics Statement

To better understand the characteristics of synthetic fake news, we propose a fact-enriched synthetic news generation method to generate high quality news pieces. From the automatic and human evaluation results, we find that FACTGEN can generate human-like and convincing news pieces. In this paper, we also discuss a possible solution to defend this attack, which is to use the checkpoint of FACTGEN. We are discussing the further usage of FACTGEN and ethical concerns as follows:

**Journalism Assistants:** Since our method retrieves the external fact information and generate fact-consistent and fact-enriched news pieces, the journalists can utilize FACTGEN to automatically generate news by providing additional factual information and the claim. However, it still needs manually checking (Leppänen et al. 2017b).

**Synthetic Disinformation Detection:** In this paper, we shortly discuss the defending method, FACTGEN$_def$, and prove the effectiveness of it. However, like the Grover (Zellers et al. 2019a), this method mainly relies on semantic information rather than the veracity of the information (Schuster et al. 2019). Future work should verify the factual correctness of the text in the following pipeline: check-worthy sentence extraction, the verified claim matching, and prediction (Adair et al. ).

**Release Policy:** Since FACTGEN can generate human-like and convincing news content, we need to critically release the code and the model parameters. We propose to publicly release the code including generator and discriminator. However, as for the checkpoints of both models, we will only share for academic usage.

[8] https://github.com/harvardnlp/sent-summary