On Faithfulness and Factuality in Abstractive Summarization

Joshua Maynez* **Shashi Narayan*** **Bernd Bohnet** **Ryan McDonald**
Google Research
{joshuahm,shashinarayan,bohnetbd,ryanmcd}@google.com

Abstract

It is well known that the standard likelihood training and approximate decoding objectives in neural text generation models lead to less human-like responses for open-ended tasks such as language modeling and story generation. In this paper we have analyzed limitations of these models for abstractive document summarization and found that these models are highly prone to hallucinate content that is unfaithful to the input document. We conducted a large scale human evaluation of several neural abstractive summarization systems to better understand the types of hallucinations they produce. Our human annotators found substantial amounts of hallucinated content in all model generated summaries. However, our analysis does show that pretrained models are better summarizers not only in terms of raw metrics, i.e., ROUGE, but also in generating faithful and factual summaries as evaluated by humans. Furthermore, we show that textual entailment measures better correlate with faithfulness than standard metrics, potentially leading the way to automatic evaluation metrics as well as training and decoding criteria.[1]

1 Introduction

Current state of the art conditional text generation models accomplish a high level of fluency and coherence, mostly thanks to advances in sequence-to-sequence architectures with attention and copy (Sutskever et al., 2014; Bahdanau et al., 2015; Gu et al., 2016), fully attention-based Transformer architectures (Vaswani et al., 2017; Dai et al., 2019) and more recently pretrained language modeling for natural language understanding (Devlin et al., 2019; Radford et al., 2018; Yang et al., 2019; Liu et al., 2019). There has been a growing interest in understanding how maximum likelihood training and approximate beam-search decoding in these models lead to less human-like text in *open-ended text generation* such as language modeling and story generation (Holtzman et al., 2020; Welleck et al., 2020; See et al., 2019). In this paper we investigate how these models are prone to generate hallucinated text in *conditional text generation*, specifically, extreme abstractive document summarization (Narayan et al., 2018a).

Document summarization — the task of producing a shorter version of a document while preserving its information content (Mani, 2001; Nenkova and McKeown, 2011) — requires models to generate text that is not only human-like but also faithful and/or factual given the document. The example in Figure 1 illustrates that the faithfulness and factuality are yet to be conquered by conditional text generators. The article describes an event of "*Conservative MP Zac Smith winning the primary for 2016 London mayoral election*", but summaries often forge entities (e.g., "Nigel Goldsmith" or "Zac Goldwin") or information (e.g., "UKIP leader Nigel Goldsmith", "Nigel Goldsmith winning the mayoral election", "Sadiq Khan being the former London mayor" or "Zac Goldwin being the Labour's candidate") that are not supported by the document or are factually wrong. Interestingly, all summaries are topical and fluent, and perform well in terms of ROUGE scores (Lin and Hovy, 2003).

We conducted a large-scale human evaluation of hallucinated content in systems that use Recurrent Neural Network (RNN) (See et al., 2017), Convolutional Neural Network (CNN) (Narayan et al., 2018a), and Transformers (Radford et al., 2019; Rothe et al., 2020), as well as human written summaries for the recently introduced eXtreme SUMmarization task (XSUM, Narayan et al., 2018a). We seek to answer the following questions: (i) How frequently do abstractive summarizers hallucinate content?; (ii) Do models hal-

* The first two authors contributed equally.

[1] Our human annotated summaries for faithfulness and factuality will be released at https://github.com/google-research-datasets/xsum_hallucination_annotations.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1906–1919
July 5 - 10, 2020.

GOLD	Zac Goldsmith will contest the 2016 London mayoral election for the Conservatives, it has been announced.	
DOCUMENT:	The Richmond Park and North Kingston MP said **he was "honoured" after winning** 70% of the 9,227 votes cast using an online primary system. He beat London Assembly Member Andrew Boff, MEP Syed Kamall and London's deputy mayor for crime and policing Stephen Greenhalgh. **Mr Goldsmith**'s main rival is likely to be **Labour's Sadiq Khan**. *(2 sentences with 59 words are abbreviated here.)* **Mr Goldsmith**, who was **the favourite for the Tory nomination**, balloted his constituents earlier this year to seek permission to stand. At the very point of **his entry into the race for London mayor**, **Zac Goldsmith**'s decision revealed two big characteristics. *(5 sentences with 108 words are abbreviated here.)* **Mr Goldsmith** - who first entered Parliament in 2010 - told the BBC's Daily Politics that he hoped his environmental record would appeal to Green and Lib Dem voters and he also hoped to "reach out" to **UKIP** supporters frustrated with politics as usual and the UK's relationship with the EU. **Zac Goldsmith** Born in 1975, educated at Eton and the Cambridge Centre for Sixth-form Studies *(5 sentences with 76 words are abbreviated here.)* **Mr Goldsmith**, who has confirmed he would stand down from Parliament if he became mayor, triggering a by-election, said he wanted to build on **current mayor Boris Johnson**'s achievements. *(3 sentences with 117 words are abbreviated here.)* Both **Mr Khan and Mr Goldsmith** oppose a new runway at Heathrow airport, a fact described by the British Chambers of Commerce as "depressing". *(1 sentences with 31 words is abbreviated here.)* **Current mayor Boris Johnson** will step down next year after two terms in office. He is also currently the MP for Uxbridge and South Ruislip, having been returned to Parliament in May. Some **Conservatives** have called for an inquiry into **the mayoral election process** after only 9,227 people voted - compared with a 87,884 turnout for the Labour contest. *(4 sentences with 121 words are abbreviated here.)*	
PTGEN	UKIP leader Nigel Goldsmith has been elected as the new mayor of London to elect a new Conservative MP.	[45.7, 6.1, 28.6]
TCONVS2S	Former London mayoral candidate Zac Goldsmith has been chosen to stand in the London mayoral election.	[50.0, 26.7, 37.5]
TRANS2S	Former London mayor Sadiq Khan has been chosen as the candidate to be the next mayor of London.	[35.3, 12.5, 23.5]
GPT-TUNED	Conservative MP Zac Goldwin's bid to become Labour's candidate in the 2016 London mayoral election.	[42.4, 25.8, 36.4]
BERTS2S	Zac Goldsmith has been chosen to contest the London mayoral election.	[66.7, 40.0, 51.9]

Figure 1: Hallucinations in extreme document summarization: the abbreviated article, its gold summary and the abstractive model generated summaries (PTGEN, See et al. 2017; TCONVS2S, Narayan et al. 2018a; and, GPT-TUNED, TRANS2S and BERTS2S, Rothe et al. 2020) for a news article from the extreme summarization dataset (Narayan et al., 2018a). The dataset and the abstractive models are described in Section 3 and 4. We also present the [ROUGE-1, ROUGE-2, ROUGE-L] F_1 scores relative to the reference gold summary. Words in red correspond to hallucinated information whilst words in blue correspond to faithful information.

lucinate by manipulating the information present in the input document (*intrinsic hallucinations*) or by adding information not directly inferable from the input document (*extrinsic hallucinations*)?; (iii) How much hallucinated content is *factual*, even when unfaithful?; and (iv) Are there automatic means of measuring these hallucinations?

Our main conclusions are as follows: First, intrinsic and extrinsic hallucinations happen frequently – in more than 70% of single-sentence summaries. Second, the majority of hallucinations are extrinsic, which potentially could be valid abstractions that use background knowledge. However, our study found that over 90% of extrinsic hallucinations were erroneous. Thus, hallucinations happen in most summaries and the majority of these are neither faithful nor factual. Third, models initialized with pretrained parameters perform best both on automatic metrics and human judgments of faithfulness/factuality. Furthermore, they have the highest percentage of extrinsic hallucinations that are factual. This suggests that while some studies argue that large-scale pretrained models are merely better at learning data-specific regularities (Niven and Kao, 2019), at least on in-domain summarization the gains in automatic metrics are realized in observable differences by humans. Fourth, ROUGE (Lin and Hovy, 2003) and BERTScore (Zhang et al., 2020) correlates less with faithfulness/factuality than metrics derived from automatic semantic inference systems, specifically the degree to which a summary is entailed by the source document. This presents an opportunity for improved automatic evaluation measures as well as model training and decoding objectives. We show preliminary experiments in this direction.

2 Hallucinations in Summarization

Open-ended generation — the task of generating text that forms a natural continuation from the input text — requires the model to hallucinate text; hence the focus has been to ensure that the model learns to generate text that is more human-like (i.e., less repetitive or dull with more content-related words)

(Holtzman et al., 2020; Welleck et al., 2020; See et al., 2019). In contrast, tasks such as document summarization (Nenkova and McKeown, 2011; See et al., 2017; Paulus et al., 2018) and data-to-text generation (Lebret et al., 2016; Wiseman et al., 2017) which are not open-ended, require models to be factual and/or faithful to the source text.

Despite recent improvements in conditional text generation, most summarization systems are trained to maximize the log-likelihood of the reference summary at the word-level, which does not necessarily reward models for being faithful. Moreover, models are usually agnostic to the noises or artifacts of the training data, such as reference divergence, making them vulnerable to hallucinations (Kryscinski et al., 2019a; Wiseman et al., 2017; Dhingra et al., 2019). Thus, models can generate texts that are not consistent with the input, yet would likely have reasonable model log-likelihood.

2.1 Intrinsic and Extrinsic Hallucinations

Given a document D and its abstractive summary S, we try to identify all hallucinations in S with respect to the content of D, regardless of the quality of the summary. In this work, we define a summary as being hallucinated if it has a span(s) $w_i \dots w_{i+j}$, $j \geq i$, that is not supported by the input document. To distinguish hallucinations further in the context of a document and a summary, we categorize hallucinations by the information source as *intrinsic* and *extrinsic* hallucinations. Note, paraphrases or any information that can be inferred from the document are not categorized as hallucinations.

Intrinsic hallucinations are consequences of synthesizing content using the information present in the input document. For example, in Figure 1, "Former London mayoral candidate" in the TConvS2S abstract and "Former London mayor" in the Trans2S abstract are hallucinations of intrinsic nature; both use terms or concepts from the document but misrepresent information from the document, making them unfaithful to the document. The article does not confirm if "Zac Goldsmith" was a "Former London mayoral candidate" or if "Sadiq Khan" was a "Former London mayor". One may suspect that a model with poor input document representation will fail to do document level inference, often required for abstraction, and will be vulnerable to such errors.

Extrinsic hallucinations are model generations that ignore the source material altogether. For example, in Figure 1, "Nigel" in the PtGen abstract and "2016" in both Gold and GPT-Tuned are extrinsic hallucinations; these terms are not introduced in the document. A model with a poorly-informed decoder and that is agnostic to the divergence issue between the source and target texts (Wiseman et al., 2017; Dhingra et al., 2019), will function more as an open-ended language model and will be prone to extrinsic hallucinations.

2.2 Factual Hallucinations in Summarization

A summary S of a document D contains a *factual hallucination* if it contains information not found in D that is factually correct. Factual hallucinations may be composed of intrinsic hallucinations or extrinsic hallucinations.

By definition, abstractive summaries are written to preserve the salient information in the input document, but they are expressed in the words of the summary author as opposed to the input document author (Nenkova and McKeown, 2011). As such, it is natural to construct summaries that integrate with the author's background knowledge (van Dijk and Kintsch, 1978; Brown and Day, 1983). Such knowledge integration can also be desirable in real world applications. For instance, an engaging sports report will reflect an understanding of the game to provide color and context. Another example is audience-targeted summarization where a good summary will reflect understanding of both the article domain and the desired audience. Nonetheless, there is no consensus in the research community if the summary should be faithful (without any hallucinations) to the input document or if there is tolerance for factual hallucinations.

Recent deep learning approaches to abstractive summarization naturally learn to integrate knowledge from the training data while generating an abstractive summary for a document (See et al., 2017; Gehrmann et al., 2018). More advanced pretrained text generators (Radford et al., 2018, 2019; Dong et al., 2019; Song et al., 2019; Khandelwal et al., 2019; Rothe et al., 2020) are even better at capturing world knowledge as they are informed by a vast amount of background text. This can be observed in the example shown in Figure 1 as the input document does not mention that the discussed "London mayoral election" is from "2016"; but the abstract generated by the pretrained text generator GPT-Tuned correctly predicts this information similar to the human-authored abstract.[2]

[2]Despite the correct extrinsic hallucination ("2016 "), the GPT-Tuned abstract overall is still not factual due to the incorrect extrinsic hallucination in "Conservative MP Zac Goldwin." There is no Conservative MP named Zac Goldwin.

In this paper we stand in favour of the assertion that abstractive systems may integrate with the background knowledge to generate rich and meaningful summaries. More concretely, "*hallucinations in summarization are acceptable if they lead to better summaries that are factual with respect to the document and the associated background knowledge*." This assumption also allows us to assess the capability of recent neural models to integrate with the background knowledge to generate factual abstracts (see Section 5.3).

3 Extreme Document Summarization

We focus on the recently introduced extreme summarization dataset (XSUM, Narayan et al., 2018a)[3] which comprises 226,711 British Broadcasting Corporation (BBC) articles paired with their single-sentence summaries, provided by the journalists writing the articles. The dataset is split into three subsets: training (90%, 204,045), validation (5%, 11,332), and test (5%, 11,334) sets. All models in §4 trained to generate abstractive summaries are trained and evaluated using this standard split, provided by the authors.

We choose to focus our study on extreme summarization for the following reasons: First, this task aims to create a single-sentence summary of a news article; these shorter summaries are relatively easier to annotate and analyze than longer summaries such as story highlights from the CNN/Dailymail dataset (Hermann et al., 2015) or abstracts from the NY Times (Sandhaus, 2008) or the WikiSum (Liu et al., 2018) dataset. Secondly, the gold summary in the extreme summarization dataset is an introductory sentence prefacing each article. By virtue of this property, the extreme summarization task is not amenable to extractive strategies and requires an abstractive modeling approach. Hence, it provides us a better benchmark to assess abstractive models' abilities to produce abstractions which are faithful and factual. Finally, since we conclude that hallucination is a problem on this dataset, then we can safely conclude it is a problem for summarization datasets with longer summaries, as modeling longer-distance dependencies and discourse structures make the task harder.

4 Abstractive Summaries

We evaluate summaries from RNN, CNN and Transformer-based state-of-the-art abstractive summarization methods and the reference human written summaries. See the Appendix for hyperparameter and decoding details for all models.

[3]https://github.com/EdinburghNLP/XSum

Human Written Reference Summaries. The single-sentence summaries contained in the extreme summarization dataset (XSUM) are also evaluated as part of this study. These summaries were written by journalists as introductions to the news articles they precede. These summaries, therefore, often have true additional information not found in the document. Such divergence issue between source and target is not uncommon in conditional text generation (Kryscinski et al., 2019a; Wiseman et al., 2017; Dhingra et al., 2019).

RNN-based Seq2Seq. We use the Pointer-Generator model (**PTGEN**) introduced by See et al. (2017), an RNN-based attention-based sequence to sequence model which not only generates from the target vocabulary but can copy words from the source text.

Topic-aware Convolutional Seq2Seq. The Topic-aware Convolutional Sequence to Sequence model (**TCONVS2S**) introduced by Narayan et al. (2018a) is an abstractive system which is conditioned on the article's topics and based entirely on Convolutional Neural Networks (Gehring et al., 2017). TCONVS2S is better suited for extreme summarization, as convolution layers capture long-range dependencies between words in the document more effectively than RNNs. Simultaneously, the convolutional encoder associates each word with a topic vector, capturing whether it is representative of the document's content.

Transformer-based Abstractive Methods. We experiment with three Transformer-based model variants, all of which have 12 layers, a hidden size of 768, filter size of 3072, and 12 attention heads.
GPT-TUNED: Radford et al. (2019) proposed Transformer-based Generative Pre-Trained (GPT) language models that can generate high quality text in open-ended generation setups. The proposed decoder-only architecture for language modeling can be easily adapted to abstractive summarization where the model first sees the document and, given a prompt, such as TL;DR;, generates its summary. Our GPT-TUNED is warm-started with a publicly available GPT checkpoint (Radford et al., 2019), but fine-tuned with supervised training on the extreme summarization dataset.
TRANS2S and **BERTS2S**: TRANS2S and BERTS2S are sequence to sequence models

Models	Human Eval Test Set			
	R1	R2	RL	BERTScore
PtGen	30.01	9.38	23.76	74.30
TConvS2S	30.89	11.47	25.80	75.23
TranS2S	32.28	11.66	24.65	75.69
GPT-Tuned	21.82	4.72	16.28	70.35
BertS2S	**38.42**	**16.96**	**31.27**	**78.85**

Table 1: ROUGE and BERTScore F_1 scores for non-pretrained (the top block) and pretrained (the bottom block) models reported on the XSum dataset. These results are on the sampled human evaluation (500 items) dataset. The best results are **boldfaced**.

where both encoder and decoder are composed of Transformer layers (Vaswani et al., 2017; Rothe et al., 2020). All weights in TranS2S are randomly initialized, but in BertS2S, both encoder and decoder are initialized with the BERT-Base checkpoints (Devlin et al., 2019), with parameter sharing between the encoder and decoder, following Rothe et al. (2020). The only variable that is initialized randomly is the encoder-decoder attention in BertS2S. Both models are then trained on the extreme summarization dataset.

5 Experiments and Results

The main focus of this work is not to propose a solution to hallucination related issues, but to achieve a better understanding of hallucinations in abstractive summarization through their human assessment. We randomly sampled 500 articles from the test set to facilitate our study. Using the full test set was unfeasible given its size and the cost of human judgments. We have trained annotators (fluent in English) specifically for our assessment. Our annotators went through two pilot studies to have a better understanding of intrinsic and extrinsic hallucinations, and factuality of summaries. Documents used in the pilot studies were not used in the final annotation. We also report on ROUGE (Lin and Hovy, 2003) scores, BERTScore (Zhang et al., 2020) and semantic inference metric such as textual entailment (Pasunuru and Bansal, 2018; Welleck et al., 2019; Falke et al., 2019; Kryscinski et al., 2019b) and question answering (Arumae and Liu, 2019; Wang et al., 2020).

5.1 Automatic Evaluation of Summaries

ROUGE (Lin and Hovy, 2003) provides a means to quickly assess a model's ability to generate summaries closer to human authored summaries. We report on ROUGE-1 and ROUGE-2 for informativeness and ROUGE-L, for fluency. Like ROUGE, BERTScore (Zhang et al., 2020) computes a similarity score for each token in the candidate summary with each token in the reference summary. However, instead of exact matches, it computes token similarity using contextual embeddings. Results are presented in Table 1.

For both cases, the pretrained encoder-decoder architecture BertS2S performed far superior to any other randomly initialized models, such as PtGen, TConvS2S and TranS2S, and the decoder-only architecture GPT-Tuned. The differences between PtGen, TConvS2S and TranS2S are not significant; all other differences are significant.[4]

ROUGE and BERTScore are indicators of informativeness of summaries but they are not sufficient metrics to assess the overall quality of summaries. This becomes evident from our human assessments in the following sections where we employ human annotators to evaluate summaries generated with PtGen, TConvS2S, TranS2S and BertS2S, and the human authored summaries. We excluded GPT-Tuned abstracts from our study after their poor performance on the automatic measures.

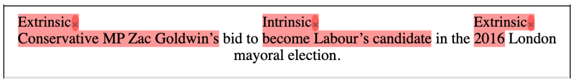

Figure 2: Human assessment of a system generated summary for the article in Figure 1. The annotation user interface is shown as it was shown to raters.

5.2 Assessment of Hallucinations

In this assessment, human annotators were presented an article and a single-sentence summary for the article. They were stringently told to only assess the hallucinations in the summary and to not confuse their assessment with the quality of the summary. For summaries containing hallucinations, annotators were tasked with (i) identifying those text spans that were unfaithful to the article and (ii) for each text span, annotating whether the hallucination was intrinsic or extrinsic. We elicited judgments from three different annotators for each of 2500 (500x5) document-summary pairs. Figure 2 shows an example assessment of a summary of an article from Figure 1. Results from the full assessment are shown in Table 2, which shows the percentage of documents per system that were annotated as faithful or hallucinated (faithful = 100 - hallucinated). The Appendix provides inter-annotator agreement of hallucinations as well as hallucinated span characteristics.

Extrinsic Hallucination due to Divergence between Source and Target. Our results con-

[4] Pairwise comparisons between all models using a one-way ANOVA with post-hoc Tukey HSD tests; $p < 0.01$.

Models	Hallucinated			Faith.	+Fact.
	I	E	I ∪ E		
PTGEN	19.9	**63.3**	75.3	24.7	27.3
TCONVS2S	17.7	71.5	78.5	21.5	26.9
TRANS2S	19.1	68.1	79.3	20.7	25.3
BERTS2S	16.9	64.1	**73.1**	**26.9**	**34.7**
GOLD	**7.4**	73.1	76.9	23.1	—

Table 2: Intrinsic vs. Extrinsic Hallucinations. The numbers in "Hallucinated" columns show the percentage of summaries where at least one word was annotated by all three annotators as an intrinsic (I) or extrinsic (E) hallucination. When a summary is not marked with any hallucination, it is "faithful" (100 - I∪E), column "Faith.". The final "+Fact." column shows the total percentage of faithful and/or factual summaries, which includes all faithful summaries plus the percentage of non-faithful summaries annotated by all three annotators as factual. Higher numbers for faithful/factual and lower numbers for hallucinations are **boldfaced**.

firmed that the BBC gold summaries often have extrinsic hallucinations due to the dataset artifact that gold summaries are introductory sentences prefacing each article. It was not surprising that most models also had significant extrinsic hallucinations.

Intrinsic Hallucination is Also Common in Abstractive Summaries. Gold summaries can also display intrinsic hallucinations. For example, a news article could describe an event related to "Barack Obama" and "the office of the President of the United States" without inferring that "Obama is the President of the United States." A journalist with the knowledge of the event in the article could write a summary stating "President Obama."

However, the percentage of system summaries with intrinsic hallucination was much higher than in gold summaries (7.4% vs others). This phenomenon particularly revealed the models' tendency to misrepresent information in the document due to the lack of document-level understanding and inference. The copy mechanism in PTGEN is good at copying from the source (showing the least percentage of extrinsic hallucination of 63.3%), but the mechanism lacks the inference capability and is prone to generate a summary that is not supported by the document (19.9% intrinsic hallucination). TRANS2S showed similar performance to PTGEN and ranked second worst. The BERTS2S showed the least number of intrinsic hallucination (16.9%) among all four abstractive systems.

Pretraining Improves Faithfulness. Hallucinations do not result from the artifacts in the training data only, but also due to model shortcomings. The PTGEN model with the copy mechanism (Gu et al., 2016; See et al., 2017) had the lowest extrinsic hallucination (63.3%), but BERTS2S reported the highest number of faithful summaries. It appears that BERTS2S is overall most conservative among all four abstractive systems while getting closer to reference summaries in terms of ROUGE. The pretraining prepares BertS2S to be more aware of the domain of the document and less prone to language model vulnerabilities. Consequently, BertS2S is more confident in predicting tokens from the document than TranS2S, hence, improving faithfulness.

5.3 Assessment of Factual Hallucinations.

Hallucinations are not necessarily erroneous. In our second human assessment, we measured to what extent this is the case. Our annotators were presented a single-sentence summary with hallucinations and were asked to assess whether it is true or false. To better explain the context of the summary, annotators were made available the source document as well as the external resources such as Wikipedia or Google Search. The source document can be particularly important for generic summaries to better understand context. External resources assisted the evaluators to validate grounded facts in public knowledge bases.

Annotators were expected to validate the summary by looking for supporting evidence for the information found on the summary. If information in the summary contradicts the document, then the summary is not factual. If supporting evidence is found for all the information, then the summary is factual. The document is not useful when the summary has information that is neither supported nor contradicted in the article. For example, the summary in Figure 2 mentions "Conservative MP Zac Goldwin" which can not be verified from the article in Figure 1. Here, annotators could use Wikipedia or Google Search to check that there had not been a Conservative MP named Zac Goldwin who tried to change their party and become a Labour's candidate in the 2016 London mayoral election.

We dropped the human authored gold summaries from this evaluation; they were presumably factual. We also dropped the abstracts that were faithful to their input documents from the previous study. Finally, there were 1869 document-summary pairs where the summaries were marked with at least one intrinsic or extrinsic hallucination. We elicited judgments from three different annotators for each of them. Results from this assessment are also presented in Table 2 (see the column labelled "+Fact.") along with the hallucination assessment.

Pretraining Helps Generating Factual Summaries. In total, 34.7% of the BERTS2S abstracts were faithful (26.9%) and/or factual (+7.8%). This is 7.4% absolute better than the next-best model (PTGEN). The number of unfaithful yet factual summaries for BERTS2S, 7.8%, was also the highest. In fact, for extrinsic hallucinations, even though PTGEN hallucinates less than BERTS2S (63.3% vs. 64.1%), 6.6% of BERTS2S hallucinations were factual, compared to 2.2% of PTGEN.[5] Thus, if we consider factual hallucinations to be valid, this means that even for extrinsic cases, BERTS2S hallucinates the least.

The superior performance of BERTS2S is most likely due to its exposure to vast amount of text through pretraining, allowing it to integrate background knowledge with generation. Even so, over 90% of BERTS2S hallucinations are erroneous.

Finally, we carried out pairwise comparisons between all models (using a one-way ANOVA with post-hoc Tukey HSD tests; $p < 0.01$). For intrinsic hallucinations (the second column in Table 2), GOLD is significantly different from all other systems. For extrinsic hallucinations (the third column in Table 2), there were significant differences between PTGEN and TCONVS2S, PTGEN and GOLD, and, BERTS2S and GOLD. For factuality, the differences between PTGEN, TCONVS2S, and TRANS2S were insignificant.

5.4 Automatic Measures for Hallucinations

Summaries are a proxy for their source documents under the assumption that they highlight the most important content. With this assumption, we further studied the extent to which the hallucinated content can be measured by semantic inference related measures, such as textual entailment and question answering.

Textual Entailment. We trained an entailment classifier by finetuning a BERT-Large pretrained model (Devlin et al., 2019) on the Multi-NLI dataset (Williams et al., 2018). We calculated the entailment probability score between the document and its abstractive summaries. Note that this entailment classifier is not optimal for the BBC article-summary pairs; the Multi-NLI dataset contains sentence-sentence pairs.

Ideally a summary should entail the document or perhaps be neutral to the document, but never contradict the document. As can be seen in Table 3, the BERTS2S abstracts showed the least number of contradictions compared to other system-generated abstracts and was at par with the GOLD summaries. Similar to the performance on extrinsic hallucination in Table 2, the TCONVS2S abstracts also displayed the highest number of contradictions. Interestingly, the GOLD summaries are more neutral to their documents, whereas the BERTS2S summaries are more entailed by their documents. This is probably due to the nature of the data and that journalists tend to add color and have a high number of extrinsic (but valid) hallucinations.

Models	Textual Entailment			QA
	entail.	**neut.**	**cont.**	
PTGEN	38.4	34.4	27.2	20.2
TCONVS2S	29.6	37.4	33.0	19.9
TRANS2S	34.6	39.8	25.6	22.4
BERTS2S	**41.8**	37.8	20.4	**23.0**
GOLD	32.8	**47.2**	**20.0**	19.3

Table 3: Textual entailment and question answering (QA) based measures for summary evaluation. For entailment, we show the percentage of times a summary entails (entail.) the document, is neutral (neut.) to the document and contradicts (cont.) the document. For QA, we report the percentage of questions that were correctly answered by a system. The highest numbers for entail., neut. and QA, and the lowest number for cont. are boldfaced.

Question Answering. QA frameworks have been used to assess or promote summary informativeness (Narayan et al., 2018b; Arumae and Liu, 2019). We adapted the QA framework to assess hallucination in model generated summaries; a faithful model will generate a summary that only has information that is supported by its document. Under this assumption, any question answerable by the summary should also be answerable by the source.

Given an abstractive summary, we used the round-trip consistency method of Alberti et al. (2019), which combines question generation and answer extraction models to generate synthetic question-answer pairs. For the 500 document-summary pairs, we generated 731, 708, 720, 725 and 820 question-answer pairs for PTGEN, TCONVS2S, TRANS2S, BERTS2S and GOLD, respectively. Finally, we used a machine reading comprehension model to answer these questions using the document as context. As in Alberti et al. (2019), we trained all models: question generation, answer extraction and reading comprehension models; using a BERT-Base pretrained model (Devlin et al., 2019) finetuned on the Natural Questions dataset (Kwiatkowski et al., 2019).

Similar to textual entailment results, the

[5]See Appendix for full results.

PTGEN	Leeds United fought back from 2-0 down to beat Huddersfield town in the first round of the EFL cup. (**Q:** *What team did Leeds United beat in the first round of the EFL cup?*, **A:** *Huddersfield town*)
TCONVS2S	A coal mine in South Yorkshire has collapsed as a result of the loss of a coal mine. (**Q:** *What type of mine has collapsed?*, **A:** *Coal*)
TRANS2S	Star Wars actor James Davis said he was "locked in a caravan" and had his caravan stolen during a break-in. (**Q:** *Who said he was locked in a caravan?*, **A:** *Davis*)

Figure 3: Sample of question-answer pairs generated from hallucinated summaries that are correctly answered by their source articles. Highlighted spans in the summaries are marked as extrinsic or intrinsic hallucinations by our annotators.

Metric	Faithful	Factual
ROUGE-1	0.197	0.125
ROUGE-2	0.162	0.095
ROUGE-L	0.162	0.113
BERTScore	0.190	0.116
QA	0.044	0.027
Entailment	**0.431**	**0.264**

Table 4: Spearman's correlation coefficient ($|r_s|$) of different metrics with faithful and factual annotations.

BERTS2S abstracts were the most faithful to their source documents in terms of question answering. The GOLD abstracts were the least accurate due to a high number of extrinsic hallucination in them.

Spearman's Correlation. We estimate Spearman's correlation coefficients of different metrics with the faithful and factual human scores (see Table 4). We found that the textual entailment scores are best correlated with both faithful (moderate, $0.40 \leq |r_s| \leq 0.59$) and factual (weak, $0.20 \leq |r_s| \leq 0.39$) human scores. Comparatively, ROUGE-based metrics and BERTScore have very weak correlation, our findings are consistent with the recent studies (Goodrich et al., 2019; Kryscinski et al., 2019a; Wang et al., 2020). Surprisingly, the question answering scores showed a very weak correlation ($0.0 \leq |r_s| \leq 0.19$) with faithful and factual human scores. We hypothesize that this is due to a compounding of errors where (i) the question generator is used to generate questions from the systems' generated abstracts, instead of human-written text on which they were trained, (ii) the question generator is susceptible to generate questions with hallucinated content when fed in with hallucinated summaries, and (iii) our assumption that a summary is faithful if the answers from the source and the summary match, is rather poor for extreme summarization. We demonstrate these issues in Figure 3. Irrespective of questions with hallucinated content, our reading comprehension model can fortuitously answer them correctly from their source articles. Better ways of generating questions (Narayan et al., 2020) and measuring factual consistency may alleviate some of these issues (Wang et al., 2020).

Models	R1	R2	RL	Faith.	+Fact.
BERTS2S	**38.42**	**16.96**	**31.27**	26.9	34.7
ENTAIL	35.93	14.02	28.87	31.5	38.6
→FAITH	37.31	15.21	30.12	**31.7**	**38.8**

Table 5: ROUGE and faithfulness/factuality scores for BERTS2S plus systems that use textual entailment as a criteria or fine-tuned on faithful annotations.

5.5 Model Selection with Entailment

Our study suggests that entailment could be used as an automatic measure for faithfulness. However, we should point out that this measure is *reference-less*. Thus, it can easily be gamed, i.e., the first sentence of any source document is always entailed by the whole document. Because of this, entailment-based measures for evaluation need to be coupled with *reference-based* measures like ROUGE.

However, one major advantage of the measure being reference-less is that we can use it as a model selection objective or during decoding. We tested the former. Specifically, we used the probability that a summary is entailed by a document as a selection criteria to select a summary between four candidates generated by systems evaluated: PTGEN, TCONVS2S, TRANS2S, and BERTS2S. Results are shown in the ENTAIL row of Table 5. We can see that indeed this is a strong metric to optimize towards if we want faithful summaries - almost 5% absolute better. There is a trade-off in terms of ROUGE, but this model must select amongst 4 systems, 3 of which have significantly lower ROUGE than the best model.

A further experiment is to train a model explicitly to predict faithfulness. In order to do this, we further fine-tuned the entailment model using the 'faithful' annotations generated during our evaluation. For all summary-document pairs marked as 'faithful', we set the associated class to 'entailment', otherwise we set it to 'neutral'. This allowed for us to also fine-tune the last classification layers taking advantage of the correlation between 'entailment' and 'faithfulness'. Results using 5-fold cross validation are shown in the ENTAIL→FAITH row of Table 5. We see here that indeed this does improve the ability to select faithful summaries from a set of candidates, though slightly. We would expect to see larger gains with more training data. However, this model is significantly better than ENTAIL on ROUGE-based metrics and seems like a good balance between ROUGE and better faithfulness.

6 Related Work

Following the Document Understanding Conference (DUC; Dang, 2005), a majority of work has focused on evaluating the content and the linguistic quality of summaries (Nenkova, 2005). Most popular among them is the automatic metric ROUGE (Lin and Hovy, 2003) that measures the unigram and bigram overlap (ROUGE-1 and ROUGE-2) as a proxy for assessing informativeness and the longest common subsequence (ROUGE-L), for fluency. ROUGE, however, can be misleading when used as the only means to assess the informativeness of summaries (Schluter, 2017). Hence, the ROUGE score is often complemented with subjective human assessment of summaries. More objective measures have been proposed to improve agreement among human annotators. Pyramid method (Nenkova and Passonneau, 2004) requires summaries to be annotated by experts for salient information. Narayan et al. (2018a,b) used a question-answering based approach where a summary is used as context to answer questions which were written based on its reference summary. Hardy et al. (2019) proposed a reference-less approach where a summary is assessed against the source document, highlighted with its pertinent content.

There has not been much work on evaluating faithfulness and truthfulness of abstractive summaries. The automatic evaluation such as ROUGE and the human evaluation of saliency and linguistic quality of summaries are not sufficient due to the complex nature of the task. Recently, Chen and Bansal (2018) asked human annotators to assess the summary relevance measuring both the saliency and the presence of contradictory/unrelated information. Dhingra et al. (2019) proposed a new automatic metric, PARENT, for data-to-text generation (Lebret et al., 2016; Wiseman et al., 2017) which aligns n-grams from the reference and generated texts to the source table to measure the accuracy of n-grams that are entailed from the source table. Goodrich et al. (2019) proposed a model-based automatic metric to assess the faithfulness of Wikipedia summaries; they trained an end-to-end model to extract a complete set of OpenIE-style (Banko et al., 2007) facts from both the source text and the generated summary. The summary is faithful if it is precise in generating facts from the source text. In our experiments with OpenIE-based measures, we found that they are not suited for evaluating extreme summarization models; all models perform poorly on these metrics without any significant differences. Like ours, few recent works (some in parallel) have explored natural language inference and question answering models to detect factual consistency in generated text (Welleck et al., 2019; Falke et al., 2019; Kryscinski et al., 2019b; Wang et al., 2020). In line with our findings, Falke et al. (2019) observed that the BERT-based NLI models substantially improved summaries reranking in terms of their correctness. Kryscinski et al. (2019b) proposed an NLI-based fact checking model that is trained on a dataset tailored for detecting factual inconsistencies in generated text. Wang et al. (2020) proposed a question answering and generation based automatic evaluation protocol that is designed to identify factual inconsistencies in a generated summary. Future work will likely investigate better ways of generating questions and measuring factual consistency to address poor correlation with faithfulness and factuality annotations.

Finally, others have used reinforcement learning to improve informativeness and reduce contradictory information in abstractive summaries, e.g., Pasunuru and Bansal (2018) used a textual entailment-based reward and Arumae and Liu (2019), a question-answering based reward. However, these approaches don't evaluate if these rewards improve faithfulness of summaries.

7 Conclusion

We conducted a large-scale study of hallucinations in abstractive document summarization. We found that (i) tackling hallucination is a critical challenge for abstractive summarization, perhaps the most critical, (ii) NLU-driven pretraining in neural text generators is key to generate informative, coherent, faithful and factual abstracts, but it is still far from solving the problem; and (iii) measures such as ROUGE or BERTScore will not be sufficient when studying the problem; semantic inference-based automatic measures are better representations of true summarization quality.

Acknowledgments

We thank Ratish Puduppully, Yova Kementchedjhieva, Ankur Parikh, Peter Liu, Slav Petrov, the reviewers and the action editor for invaluable feedback. The hard work of Muqthar Mohammad, Mohd Majeed and Ashwin Kakarla made our human annotation possible.

References

Chris Alberti, Daniel Andor, Emily Pitler, Jacob Devlin, and Michael Collins. 2019. Synthetic QA corpora generation with roundtrip consistency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6168–6173, Florence, Italy.

Kristjan Arumae and Fei Liu. 2019. Guiding extractive summarization with question-answering rewards. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2566–2577, Minneapolis, Minnesota.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations*, San Diego, CA, USA.

Michele Banko, Michael J. Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni. 2007. Open information extraction from the web. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, pages 2670–2676, Hyderabad, India.

Ann L. Brown and Jeanne D. Day. 1983. Macrorules for summarizing texts: The development of expertise. *Journal of Verbal Learning and Verbal Behaviour*, 22(1):1–14.

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 675–686, Melbourne, Australia.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy.

Hoa Trang Dang. 2005. Overview of DUC 2005. In *Proceedings of the Document Understanding Conference*, pages 1–12.

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*, pages 4171–4186, Minneapolis, Minnesota.

Bhuwan Dhingra, Manaal Faruqui, Ankur Parikh, Ming-Wei Chang, Dipanjan Das, and William Cohen. 2019. Handling divergent reference texts when evaluating table-to-text generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4884–4895, Florence, Italy.

Teun A. van Dijk and Walter Kintsch. 1978. Cognitive psychology and discourse: Recalling and summarizing stories. In Wolfgang U. Dressler, editor, *Current Trends in Textlinguistics*, pages 61–80.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems 32*, pages 13063–13075. Curran Associates, Inc.

Tobias Falke, Leonardo F. R. 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, Florence, Italy.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 1243—-1252, Sydney, NSW, Australia.

Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. 2018. Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4098–4109, Brussels, Belgium.

Ben Goodrich, Vinay Rao, Peter J. Liu, and Mohammad Saleh. 2019. Assessing the factual accuracy of generated text. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 166–175, New York, NY, USA.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O.K. Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1631–1640, Berlin, Germany.

Hardy, Shashi Narayan, and Andreas Vlachos. 2019. HighRES: Highlight-based reference-less evaluation of summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3381–3392, Florence, Italy.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems 28*, pages 1693–1701. Curran Associates, Inc.

Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *Proceedings of the 8th International Conference on Learning Representations*, Virtual Conference, Formerly Addis Ababa Ethiopia.

Urvashi Khandelwal, Kevin Clark, Dan Jurafsky, and Lukasz Kaiser. 2019. Sample efficient text summarization using a single pre-trained transformer. *CoRR*, abs/1905.08836.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019a. Neural text summarization: A critical evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 540–551, Hong Kong, China.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2019b. Evaluating the factual consistency of abstractive text summarization. *CoRR*, abs/1910.12840.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *CoRR*, abs/1808.06226.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Rémi Lebret, David Grangier, and Michael Auli. 2016. Neural text generation from structured data with application to the biography domain. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1203–1213, Austin, Texas.

Chin Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 150–157.

Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences. In *Proceedings of the 6th International Conference on Learning Representations*, Vancouver Canada.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Inderjeet Mani. 2001. *Automatic summarization*, volume 3. John Benjamins Publishing.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018a. 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.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018b. Ranking sentences for extractive summarization with reinforcement learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1747–1759, New Orleans, Louisiana.

Shashi Narayan, Gonçalo Simoes, Ji Ma, Hannah Craighead, and Ryan T. McDonald. 2020. QURIOUS: Question generation pretraining for text generation. *CoRR*, abs/2004.11026.

Ani Nenkova. 2005. Automatic Text Summarization of Newswire: Lessons Learned from the Document Understanding Conference. In *Proceedings of the 20th National Conference on Artificial Intelligence - Volume 3*, pages 1436–1441.

Ani Nenkova and Kathleen McKeown. 2011. Automatic summarization. *Foundations and Trends in Information Retrieval*, 5(2–3):103–233.

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating content selection in summarization: The Pyramid method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 145–152, Boston, Massachusetts, USA.

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy.

Ramakanth Pasunuru and Mohit Bansal. 2018. Multi-reward reinforced summarization with saliency and entailment. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 646–653, New Orleans, Louisiana.

Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. In *Proceedings of the 6th International Conference on Learning Representations*, Vancouver, BC, Canada.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. Technical report.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Technical report.

Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2020. Leveraging pre-trained checkpoints for sequence generation tasks. *To appear in Transactions of the Association for Computational Linguistics*, abs/1907.12461.

Evan Sandhaus. 2008. The New York Times Annotated Corpus. *Linguistic Data Consortium, Philadelphia*, 6(12).

Natalie Schluter. 2017. The limits of automatic summarisation according to rouge. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 41–45, Valencia, Spain.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1073–1083, Vancouver, Canada.

Abigail See, Aneesh Pappu, Rohun Saxena, Akhila Yerukola, and Christopher D. Manning. 2019. Do massively pretrained language models make better storytellers? In *Proceedings of the 23rd Conference on Computational Natural Language Learning*, pages 843–861, Hong Kong, China.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. MASS: Masked sequence to sequence pre-training for language generation. In *Proceedings of the 36th International Conference on Machine Learning*, Long Beach, California.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Alex Wang, Kyunghyun Cho, and Michael 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*, Virtual Conference, Formerly Seattle, USA.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2020. Neural text generation with unlikelihood training. In *Proceedings of the 8th International Conference on Learning Representations*, Virtual Conference, Formerly Addis Ababa Ethiopia.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. Dialogue natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3731–3741, Florence, Italy.

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*, pages 1112–1122, New Orleans, Louisiana.

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263, Copenhagen, Denmark.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *Proceedings of the 8th International Conference on Learning Representations*, Virtual Conference, Formerly Addis Ababa Ethiopia.

A Model Hyperparameters and Predictions

PTGEN and TCONVS2S model predictions are provided by Narayan et al. (2018a) and Transformer model predictions from GPT-TUNED, TRANS2S and BERTS2S, by Rothe et al. (2020). Both PTGEN and TCONVS2S use a Stanford tokenized vocabulary size of 50k. TRANS2S and BERTS2S use a vocabulary size of around ∼30k WordPieces (Wu et al., 2016) to match BERT pretrained vocabulary and, GPT-TUNED, a vocabulary size of around ∼50k SentencePieces (Kudo and Richardson, 2018) to match the GPT-2 pretrained vocabulary. All models use the same uncased vocabulary on both source and target sides. Both PTGEN and TCONVS2S summaries were generated using beam search with beam size 10, the Transformer models use beam size of 4. See Narayan et al. (2018a) and Rothe et al. (2020) for more details on these models.

Models	**Fleiss' Kappa**			
	Hall.	**Fact.**	**Rept.**	**Inco.**
PTGEN	0.70	0.91	0.89	0.84
TCONVS2S	0.73	0.91	0.93	0.90
TRANS2S	0.67	0.91	0.92	0.90
BERTS2S	0.67	0.88	0.94	0.93
GOLD	0.71	—	1.00	0.98

Table 6: Fleiss's Kappa scores measuring word-level agreements among annotators for different annotation tasks: hallucination (Hall.), factuality (Fact.), repetition (Rept.) and incoherence (Inco.) assessments.

B Inter annotator agreement

We estimated Fleiss's Kappa (k) to assess the agreement among our raters when categorizing a word in the summary as one of faithful, intrinsically hallucinated and extrinsically hallucinated. The results are shown in Table 6. All models showed substantial agreement ($0.61 \leq k \leq 0.80$; Landis and Koch, 1977) among their annotations.

Table 6 also shows Fleiss's Kappa (k) to assess the agreement among our raters for factuality. All models showed almost perfect agreement ($0.81 \leq k \leq 1.0$; Landis and Koch, 1977) among their annotations.

C Highlighted Span Characteristics

Results in Table 7 shed some light on the characteristics of hallucinated spans observed in different abstracts. GOLD abstracts showed the least number of intrinsically hallucinated spans (0.55 per document), whereas, PTGEN abstracts showed the least number of extrinsically hallucinated spans (2.85 per document). Interestingly, the average span length for PTGEN summaries was 8.48 words, much higher than 6.12 words for BERTS2S summaries. Our result demonstrates that (i) the effect of hallucination in BERTS2S is more local than what we observe in PTGEN and (ii) despite a lower number of extrinsically hallucinated spans or documents in PTGEN compared to that in BERTS2S (2.85 vs 3.04 spans per document, 63.3% vs 64.1% documents), the total number of words that were annotated as extrinsic hallucination is much higher in PTGEN than in BERTS2S (12075 vs 9302 words).

Models	**Intrinsic total (avg.)**	**Extrinsic total (avg.)**	**avg. length**
PTGEN	625 (1.35)	**1424 (2.85)**	8.48
TCONVS2S	518 (1.04)	1556 (3.11)	8.44
TRANS2S	589 (1.18)	1556 (3.11)	7.39
BERTS2S	530 (1.06)	1520 (3.04)	**6.12**
GOLD	**276 (0.55)**	1807 (3.61)	7.11

Table 7: Total number of spans and the average number of spans per document, annotated as intrinsic or extrinsic hallucinations for all 500 document-summary pairs by three annotators. We also show the average span length for each system.

Models	**Repetition**	**Incoherence**
PTGEN	17.5	20.3
TCONVS2S	16.7	17.7
TRANS2S	8.9	11.5
BERTS2S	8.7	9.5
GOLD	**0.0**	**0.8**

Table 8: Repetition and Incoherence Evaluation. The numbers show the the percentage of 500 summaries where at least one word in a summary was annotated by all three annotators with the "Repetition" or "Incoherence" related issue. The lowest numbers are boldfaced.

Metric	**Faithful**	**Factual**
ROUGE-1	0.197	0.125
ROUGE-2	0.162	0.095
ROUGE-L	0.162	0.113
BERTScore	0.190	0.116
Repetition	0.064	0.075
Incoherence	0.067	0.082
QA	0.044	0.027
Entailment	**0.431**	**0.264**

Table 9: Spearman's correlation coefficient ($|r_s|$) of different metrics with faithful and factual annotations.

D Assessment of Linguistic Irregularities.

Following standard practice in summarization, all 2500 document-summary pairs were annotated for repetition and incoherence related linguistic irregularities. Annotators were presented only a single-sentence summary and were asked to identify all

Models	Faithful	Hallucinated						Factual
		I		E		I ∪ E		
		total	factual	total	factual	total	factual	
PTGEN	24.7	19.9	0.4	**63.3**	2.2	75.3	2.6	27.3
TCONVS2S	21.5	17.7	0.8	71.5	5.0	78.5	5.4	26.9
TRANS2S	20.7	19.1	1.4	68.1	3.4	79.3	4.6	25.3
BERTS2S	**26.9**	16.9	**1.8**	64.1	**6.6**	**73.1**	**7.8**	**34.7**
GOLD	23.1	**7.4**	—	73.1	—	76.9	—	—

Table 10: Intrinsic vs Extrinsic Hallucinations and their factuality. The numbers in "Hallucinated" columns show the percentage of summaries out of 500 where at least one word was annotated by all three annotators as an intrinsic (I) or extrinsic (E) hallucination. When a summary is not marked with any hallucination, it is "faithful" (1- I∪E). The "factual" columns within the "Hallucinated" column show for each type (I, E and I∪E), the percentage of summaries out of 500 annotated by all three annotators as factual. The final "Factual" column shows the total percentage of factual summaries (Faithful + $\text{I}\cup\text{E}_{\text{factual}}$). The highest numbers for faithful and factual, and the lowest numbers for hallucinations are boldfaced.

spans of text in the summary that were either repeated or made the summary incoherent. We again elicited judgments from three different annotators for each document-summary pair. Results are shown in Table 8.

Overall, all neural text generation systems are getting better in generating repetition-free and coherent single-sentence summaries of news articles. Transformer-based models, TRANS2S and BERTS2S in particular, perform superior to RNN-based PTGEN and CNN-based TCONVS2S models. Nonetheless, Table 9 shows that these metrics fail to correlate with faithful, hallucinated and factual assessments of summaries. Fleiss's Kappa (k) values for repetition and incoherence assessments showed almost a perfect agreement ($0.81 \leq k \leq 1.0$; Landis and Koch, 1977) among our raters (see Table 6).

E Full Hallucination Results

Table 10 has the full results from our human study of hallucinations.