Impact Analysis of Entailment Awareness in Text Summarization Models

Brijendra Kumar Asthana, Karan Joglekar, Shashakt Jha
New York University
{bka2022, kaj9196, ssj2580}@nyu.edu
Github Link: https://github.com/bka2022/entailment-awareness-text-summarization

Abstract

Text Summarization models are used to condense lengthy text documents into a shorter, compact version while retaining all important information and meaning from the source. Although large pre-trained models have significantly advanced the field of abstractive summarization, they can generate highly fluent outputs that contain content not supported by the source material. This content, referred to as hallucinations, can introduce factual inaccuracies or misleading information. The goal is to produce a summary that accurately represents the content of the original text in a concise form. Such models are prone to inaccuracies that either misrepresent or add additional incorrect content to the summary referred to, in literature, as hallucinations. To minimize hallucinated content, we suggest a new fine-tuning objective that combines traditional cross-entropy loss with an entailment reward. We investigate the optimal application of sentence-level entailment models to multi-sentence inputs and demonstrate that our entailment scores effectively differentiate between hallucinated and non-hallucinated summaries. We compare the performance of entailment-aware summarization models using metrics such as fluency, factuality, and faithfulness of the summary where entailment is a reward function, to determine the impact of entailment on existing models such as BART. We show that this fine-tuning approach reduces the occurrence of hallucinations in a small dataset and discuss future research directions in entailment-aware summarization.

Introduction

While highly fluent, current state-of-the-art neural models are prone to inaccuracies, either misrepresenting information from the source text or introducing additional content (Maynez et al., 2020). Addressing hallucination is difficult due to the challenge in directly measuring it; many summarization models use lexical-overlap metrics like ROUGE as a proxy for summary quality, but these metrics don't always correlate with factuality. Moreover, some instances of hallucination can be factual, as they are grounded in information learned during pre-training (Huang et al., 2020).

Summarization methods can be classified into two main categories: extractive and abstractive. Extractive summarization methods (e.g., Aho & Ullman 1972, Mallick et al. 2019, Liu 2019) select key phrases or sentences directly from the source text to create a condensed version. In contrast, abstractive methods (e.g., Lewis et al. 2019, Zhang et al. 2020, Zheng et al. 2021) generate new text for the summary. Although extractive models tend to score higher on factuality metrics, they perform poorly in certain settings, such as dialogue summarization (Gliwa et al., 2019) and extreme summarization (Narayan et al., 2018). Abstractive models, however, are more susceptible to factual errors or misrepresentations, known as hallucinations.

Hallucinations can be categorized into intrinsic and extrinsic types. Intrinsic hallucinations are misrepresentations of the information in the source document, while extrinsic hallucinations contain information not present in the source text, regardless of their factuality.

Textual entailment, or natural language inference (NLI), is a related task often discussed in the context of hallucination detection. Given a premise text (p) and a hypothesis text (h), the goal is to classify the relationship between the pair (h, p) as entailment, contradiction, or neutral. Entailment-based evaluation methods are expected to classify extrinsic hallucinations as neutral and intrinsic hallucinations as contradictions.

Motivation

The choice of entailment as a reward metric is motivated by several factors. Prior work shows that the factuality-aware evaluation of summarization has highlighted the effectiveness of entailment measures for detecting hallucinated content. Additionally, datasets for entailment are available for pre-training. However, it is also vital that the model does not give too many misclassifications, leading to a noisy reward. In all experiments for this project, we use ALBERT (Lan et al., 2019) fine-tuned on SNLI. This model achieves an F1 of 0.9060 on SNLI, near the state-of-the-art F1 of 0.9301 (Wang et al., 2021). Crucially, it is relatively small for a BERT-based model, at 18M parameters.

Literature Survey

Falke et al. (2019) were the pioneers in using entailment tasks to evaluate the factuality of summarization models.

The limited success in this area can be attributed to two factors. First, NLI objectives often depend on heuristics like lexical overlap (e.g., ROUGE, BLEU) to assess entailment; abstractive summaries might score well on these metrics but still contain factual inconsistencies, as these metrics may not adequately penalize quantitative inconsistencies. Second, a domain shift issue arises between NLI datasets and summarization datasets. Mishra et al. (2021) hypothesize that the key difference between NLI datasets and summarization datasets lies in the size of the premise. NLI datasets have premises and hypotheses that span a few sentences, while in summarization, the premise (the text to be summarized) is usually a long paragraph or a full article. To address this problem, they create a long-premise NLI dataset from existing QA datasets and observe accuracy improvements in factual inconsistency evaluation.

Entailment knowledge can be further incorporated into the summarization training process. Li et al. (2018) suggest a multi-task learning framework where both the encoder and decoder are made entailment-aware. They achieve this through a novel objective function that includes an entailment-based reward for the summarization task. However, this model doesn't benefit from any pretraining. In contrast, we directly use pretrained BART models and incorporate entailment-aware fine-tuning to reap the advantages of both entailment awareness and large pretrained seq2seq models.

Technical Details

Implementation of REINFORCE

To enhance the model's performance, we introduce an additional loss term using REINFORCE with policy gradients. This loss term, referred to as a self-loss, measures the likelihood of generating the model's proposed summary sequence $\hat{Y}$ based on the input document. The self-loss is then adjusted by the entailment reward, which is explained in detail below.

Consequently, the fine-tuning loss becomes a weighted combination of the label-smoothed cross-entropy loss and the policy gradient loss:

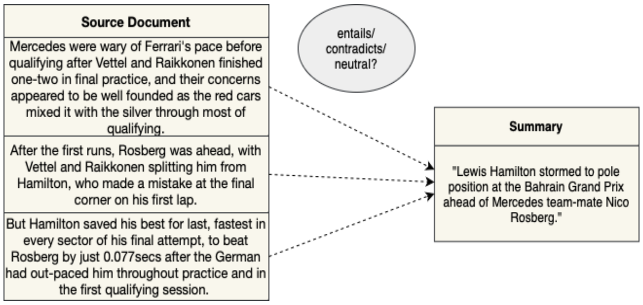
$$L = \alpha L_{CE} + \beta R L_{self}$$

In this context, α and β represent hyperparameters that govern the weights assigned to each term. One notable benefit of this approach is its compatibility with other entailment models. The cross-entropy loss and self-loss remain unaffected by the specific choice of entailment model. Consequently, the entailment reward calculation can be substituted with alternative models to strike a balance between accuracy and computational efficiency, without necessitating any modifications to the remaining components of the training process.

Entailment

The selection of entailment as a metric for rewarding the model is driven by several factors. Previous research on evaluating summarization with a focus on factuality has demonstrated the effectiveness of entailment measures in identifying generated content (Falke et al., 2020). In essence, a faithful summary should be completely entailed by the source document. Moreover, datasets specifically designed for entailment tasks are readily accessible, enabling pre-training on this aspect. One such dataset is the Stanford Natural Language Inference dataset (SNLI), comprising 570K sentence pairs with corresponding entailment labels (Bowman et al., 2015).

During the fine-tuning process, the chosen entailment model needs to be executed efficiently during the forward pass. However, it is equally important that the model does not produce an excessive number of misclassifications, which would introduce excessive noise into the reward signal. In all our experiments, we employ ALBERT (Lan et al., 2019) that has been fine-tuned on SNLI. This model achieves an F1 score of 0.9060 on SNLI, approaching the state-of-the-art F1 score of 0.9301 (Wang et al., 2021). Importantly, ALBERT is relatively compact for a BERT-based model, with 18M parameters.



Mercedes were wary of Ferrari's pace before qualifying after Vettel and Raikkonen finished one-two in final practice, and their concerns appeared to be well founded as the red cars mixed it with the silver through most of qualifying.

After the first runs, Rosberg was ahead, with Vettel and Raikkonen splitting him from Hamilton, who made a mistake at the final corner on his first lap.

But Hamilton saved his best for last, fastest in every sector of his final attempt, to beat Rosberg by just 0.077secs after the German had out-paced him throughout practice and in the first qualifying session.

"Lewis Hamilton stormed to pole position at the Bahrain Grand Prix ahead of Mercedes team-mate Nico Rosberg."

Figure 1: The calculation of the entailment vector involves a many-to-one relationship.

Sentence-level entailment

Both Mishra et al. (2021) and Maynez et al. (2020) acknowledge the discrepancy between Natural Language Inference (NLI) datasets, where premises and hypotheses consist of a single sentence, and summarization datasets, where both can be significantly longer. To effectively utilize the SNLI-pretrained model, we calculate entailment scores individually for each sentence in the input and the output. It's worth noting that for XSum, where the output is a single sentence, this entails a many-to-one computation. However, the same approach can be extended to datasets where summaries may comprise multiple sentences. Figure 1 illustrates this process. The outcome is a matrix of entailment scores, denoted as $S \in \mathbb{R}^{n \times 3}$, where n represents the number of sentences in the input document. To transform this matrix into a single reward value, we explore various functions applied to S. In order to determine the most effective function for computing the reward, we treat each potential function as a distinct evaluation measure and evaluate their relative effectiveness using a dataset annotated for hallucinated content.

Evaluation Metrics

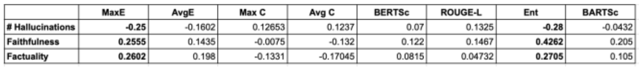
We examine two aggregation functions, namely average and maximum, for consolidating the individual scores of all (premise, hypothesis) pairs. In this context, the hypothesis represents the summary (a single sentence), while the premise is selected from the collection of source sentences. The individual score itself can be either the entailment score or the contradiction score. Thus, we have a total of 4 potential metrics. It is evident that the minimum function does not yield satisfactory results as an aggregation measure. There is no inherent justification for minimum entailment to exhibit a stronger negative correlation with hallucination compared to maximum entailment. This observation is supported by the lower magnitude of Pearson correlation scores.

We assess the performance of all proposed functions by examining the correlation between their scores and human annotations for hallucination. The dataset used for evaluation consists of 500 summaries compiled by Maynez et al. (2020). In this dataset, each span of hallucinated content in the 500 candidate summaries was labeled by three human annotators. Maynez et al. shared their annotations for four different models, and we focus on evaluating our metrics on the two stronger models, namely BertS2S and PtGen.

To quantify hallucination, we consider the total number of labeled spans as the hallucination score for each summary. We calculate Pearson's correlation between our entailment reward values and the hallucination scores. Additionally, we compute Pearson's correlation scores between our entailment reward values and the binary faithfulness and factuality scores provided by Maynez et al.

We compare the performance of our aggregated entailment evaluation metrics (Max E, Avg E, Max C, Avg C) with other commonly used metrics such as ROUGE-L, BERTScore, BARTScore, and the entailment score reported by Maynez et al., which we refer to as Ent. The entailment score reported by Maynez et al. was calculated using a BERTlarge model trained on the MNLI dataset.

We find that the Ent score performs the best across all metrics. It exhibits the highest negative correlation with the number of hallucinations and the highest positive correlation with faithfulness and factuality scores. Our proposed metric, the maximum entailment score (Max E), is the second-best measure. It demonstrates comparable performance to the entailment measure reported by Maynez et al., but with the advantage of being computationally more tractable. This is because Max E is computed using a smaller model with significantly fewer parameters, around 20 times fewer than the BERTlarge model used for the Ent score.



	MaxE	AvgE	Max C	Avg C	BERTSc	ROUGE-L	Ent	BARTSc
# Hallucinations	**-0.25**	-0.1602	0.12653	0.1237	0.07	0.1325	**-0.28**	-0.0432
Faithfulness	**0.2555**	0.1435	-0.0075	-0.132	0.122	0.1467	**0.4262**	0.205
Factuality	**0.2602**	0.198	-0.1331	-0.17045	0.0815	0.04732	**0.2705**	0.105

Table 1:Pearson correlation scores of various metrics with number of hallucinations, faithfulness and factuality

Experiments

BART (Lewis et al., 2019) is a sequence-to-sequence model designed for denoising tasks. It consists of a bidirectional encoder and a left-to-right decoder. During training, BART learns to reconstruct original text from corrupted input.

To fine-tune BART for our purposes, we used our modified loss function. To reduce computation time, we trained each model on the first 25,000 examples in the XSum dataset's train split, which accounts for approximately 10% of the dataset. We then evaluated the models on the first 10,000 examples in the XSum validation split. Our training setup involved using a batch size of 8, 500 warmup steps, a weight decay of 0.1, a label smoothing factor of 0.1, and an initial learning rate of 3e-05. Each model was trained for 5 epochs.

Based on the results obtained from the comparison of evaluation metrics, we selected the maximum entailment score as our reward for each example. Since a higher reward corresponds to a larger loss, we computed the reward as $R = 1 - max\_ent(X, \hat{Y})$, where $max\_ent(X, \hat{Y})$ represents the maximum entailment score between the input (X) and the generated summary ($\hat{Y}$). The reward term falls within the range of [0, 1] since all entailment scores are within this range. We trained a model using this reward directly.

Additionally, we implemented a model with a baseline entailment score**.** In this case, we calculated the reward for each gold summary and subtracted it from the reward of the generated summary. Mathematically, this can be expressed as:

$$R = max\_ent(X, Y) - max\_ent(X, \hat{Y})$$

When the entailment score of the generated summary is higher than that of the gold summary, a negative reward is assigned, which reduces the overall loss. On the other hand, if the entailment score of the generated summary is lower than that of the gold summary, the model is still penalized but to a lesser extent. This helps stabilize the training process, especially when using a low batch size. If the model encounters a batch of examples where even the gold summary has relatively weak entailment, the penalty received during that update is reduced.

Gold summaries in XSum can have low maximum entailment scores when the summary sentence is highly abstract and combines small details from multiple sentences in the document. Furthermore, gold summaries in XSum may contain content that is not present in the rest of the article, which further lowers the entailment score.

Hyperparameters

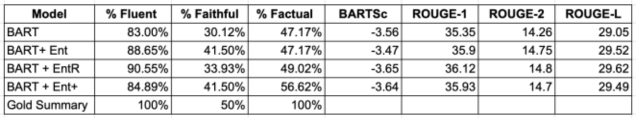
We trained a model using the regularizing baseline mentioned above. In all previous experiments, we used hyperparameters $\alpha = \beta = 1$ for the combination of losses. However, in the last experiment, we increased β significantly and decreased α to amplify the effect of the entailment reward on the training process. Specifically, we set $\alpha = 0.2$ and $\beta = 1.8$ for this experiment and ran the model without the regularizing baseline.

Results

We conduct evaluations on four models: BART-base, BART-base fine-tuned using an unregularized entailment reward, BART-base fine-tuned using our baseline regularized entailment reward, and BART-base fine-tuned with a stronger entailment reward where the reward-based loss is given more weight.

For our human evaluation, we assign binary labels to each generated summary based on three metrics: fluency, faithfulness, and factuality. Fluency refers to whether the sentence is syntactically complete and semantically consistent. Faithfulness assesses whether all statements made in the summary, including entity names and quotes, can be found in the corresponding article. Factuality determines if the information in the summary aligns with common-sense reasoning and is not directly contradicted by the article. If the information is too specific to be easily verified and does not appear in the article, we label the summary as non-factual. It's important to note that gold label summaries are considered fluent and factual, and we only annotate them for faithfulness. Therefore, our factuality scores are a relaxed version of the faithfulness scores, allowing the model to make substitutions for improved fluency. For example, if the article describes events in a city and the summary mentions that the events occur in a general region, the faithfulness score would be 0 if the region name does not appear in the article. However, the factuality score would be 1 if the city mentioned is within that region. Overall, we annotated 53 documents, each with 5 summaries, resulting in a total of 265 summary-document pairs.

The results of our evaluation are presented in Table 2. We observe that models incorporating entailment-aware techniques achieve higher scores in terms of both faithfulness and factuality, with significant differences compared to the baseline. The model with the strongest entailment reward exhibits the highest factuality scores. However, we do not observe a noticeable trend in relation to any ROUGE-based metric. In terms of fluency, all entailment models outperform the baseline in our manual annotation. It is possible that this improvement is influenced by the entailment model itself, as non-fluent generated summaries are less likely to receive favorable scores from the entailment model.



Model	% Fluent	% Faithful	% Factual	BARTSc	ROUGE-1	ROUGE-2	ROUGE-L
BART	83.00%	30.12%	47.17%	-3.56	35.35	14.26	29.05
BART+ Ent	88.65%	41.50%	47.17%	-3.47	35.9	14.75	29.52
BART + EntR	90.55%	33.93%	49.02%	-3.65	36.12	14.8	29.62
BART + Ent+	84.89%	41.50%	56.62%	-3.64	35.93	14.7	29.49
Gold Summary	100%	50%	100%				

Table 2: The baseline and entailment-aware models were evaluated using both human and automatic methods. The fine-tuned BART model was tested without entailment awareness, while BART+Ent represents an entailment-aware version of BART. BART+EntR incorporates a regularized baseline in addition to the policy gradient, and BART+Ent+ has an enhanced impact of the entailment reward. Human evaluation scores were included to compare against the gold summaries.

Interestingly, the addition of the regularization baseline leads to lower faithfulness scores. This outcome may be attributed to the regularization process weakening the rewards and placing greater emphasis on the traditional cross-entropy loss. Furthermore, simply increasing the contribution of the entailment reward to the loss function does not proportionally enhance faithfulness. Although this approach yields gains in factuality, it does not translate into improved faithfulness, suggesting that the entailment-boosted model is generating more accurate extrinsic hallucinations. Additionally, fluency decreases when the entailment reward is significantly boosted, indicating that the model starts prioritizing extreme matching to a single sentence at the expense of other considerations.

Qualitative notes

During our manual error analysis, we identified several additional insights. One notable observation is that a significant number of hallucinations (four cases in our set of 53 documents) occur when the model tries to guess the first name of a person who is only referred to by their last name in the article. This issue stems from the construction of the dataset itself. The input text removes the first sentence of the article to use it as the gold label summary. In the news domain, it is common practice to mention a person's full name only once, usually when they are first introduced, and subsequently refer to them by their last name throughout the rest of the article. As a result, the model's tendency to generate the first name in such cases leads to these specific types of hallucinations.

In cases where the person's first name only appears in the gold summary and not in the article, all four models attempt to guess the first name, but they fail to do so accurately. To address this issue in future work, it may be beneficial to penalize the models for such guessing behavior in these common scenarios.

Another notable failure point is articles with a more informal tone, which may begin with a question. We encountered two instances of this among our 53 examples. While the models can generate questions as output, the quality of these questions as summaries is questionable, and their factuality is difficult to evaluate. In our evaluation, we assess these results based on the consistency between the premise of the question and the premise of the article.

Conclusion

We present evidence of the effective application of reinforcement learning in reducing the occurrence of false information in abstractive summarization. Moreover, we highlight that adjusting the balance between traditional and entailment-based losses can influence the model's performance. Our most successful model demonstrates enhanced faithfulness, fluency, and BARTScore compared to a baseline model trained on the same data and for the same duration. To achieve even greater improvements, future research could explore the utilization of additional entailment data for pretraining, employ larger language models, or enable more detailed exploration of hyperparameters. This study underscores the value of entailment as a valuable tool for generating more faithful summaries while recognizing the need for further investigation. Our analysis of gold-label summaries suggests that prioritizing factuality as a metric for optimization in the future may require more sophisticated common-sense reasoning than faithfulness alone. By incorporating entailment into the model training process, both faithfulness and factuality can be enhanced.

References

Alfred V. Aho and Jeffrey D. Ullman. 1972. The Theory of Parsing, Translation and Compiling, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

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

Tobias Falke, Markus Boese, Daniil Sorokin, Caglar Tirkaz, and Patrick Lehnen. 2020. Leveraging user paraphrasing behavior in dialog systems to automatically collect annotations for long-tail utterances. In Proceedings of the 28th International Conference on Computational Linguistics: Industry Track, pages 21–32, Online. International Committee on Computational Linguistics.

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. Association for Computational Linguistics.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. CoRR, abs/1911.12237.

Dandan Huang, Leyang Cui, Sen Yang, Guangsheng Bao, Kun Wang, Jun Xie, and Yue Zhang. 2020. What have we achieved on text summarization? In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 446–469, Online. Association for Computational Linguistics.

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

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A lite BERT for self- supervised learning of language representations. CoRR, abs/1909.11942.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. arXiv:1910.13461 [cs, stat]. ArXiv: 1910.13461.

Haoran Li, Junnan Zhu, Jiajun Zhang, and Chengqing Zong. 2018. Ensure the correctness of the summary: Incorporate entailment knowledge into abstractive sentence summarization. In Proceedings of the 27th International Conference on Computational Linguistics, pages 1430–1441, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Yang Liu. 2019. Fine-tune BERT for Extractive Summarization. arXiv:1903.10318 [cs]. ArXiv: 1903.10318.

Chirantana Mallick, Ajit Kumar Das, Madhurima Dutta, Asit Kumar Das, and Apurba Sarkar. 2019. Graph- Based Text Summarization Using Modified TextRank. In Soft Computing in Data Analytics, pages 137–146, Singapore. Springer Singapore.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On Faithfulness and Factuality in Abstractive Summarization. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1906–1919, On- line. Association for Computational Linguistics.

Anshuman Mishra, Dhruvesh Patel, Aparna Vijayakumar, Xiang Lorraine Li, Pavan Kapanipathi, and Kartik Talamadupula. 2021. Looking beyond sentence-level natural language inference for question answering and text summarization. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational

Linguistics: Human Language Technologies, pages 1322–1336, On- line. Association for Computational Linguistics.

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

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! Topic-aware convolutional neural networks for extreme summarization. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium.

Sinong Wang, Han Fang, Madian Khabsa, Hanzi Mao, and Hao Ma. 2021. Entailment as few-shot learner. CoRR, abs/2104.14690.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Pe- ter J Liu. 2020. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. page 12.

Chujie Zheng, Kunpeng Zhang, Harry Jiannan Wang, Ling Fan, and Zhe Wang. 2021. Topic-Guided Abstractive Text Summarization: a Joint Learning Approach. arXiv:2010.10323 [cs]. ArXiv: 2010.10323.

Chunting Zhou, Jiatao Gu, Mona T. Diab, Paco Guzman, Luke Zettlemoyer, and Marjan Ghazvininejad. 2020. Detecting hallucinated content in conditional neural sequence generation. CoRR, abs/2011.02593.