On Hallucination and Predictive Uncertainty in Conditional Language Generation

Yijun Xiao, William Yang Wang
University of California, Santa Barbara
{yijunxiao,william}@cs.ucsb.edu

Abstract

Despite improvements in performances on different natural language generation tasks, deep neural models are prone to hallucinating facts that are incorrect or nonexistent. Different hypotheses are proposed and examined separately for different tasks, but no systematic explanations are available across these tasks. In this study, we draw connections between hallucinations and predictive uncertainty in conditional language generation. We investigate their relationship in both image captioning and data-to-text generation and propose a simple extension to beam search to reduce hallucination. Our analysis shows that higher predictive uncertainty corresponds to a higher chance of hallucination. Epistemic uncertainty is more indicative of hallucination than aleatoric or total uncertainties. It helps to achieve better results of trading performance in standard metric for less hallucination with the proposed beam search variant.

1 Introduction

Modern deep neural network models have brought drastic improvements of generation quality measured by standard metrics on different natural language generation (NLG) tasks. However, along with these improvements, researchers find that neural models are more prone to a phenomenon called hallucination, where models generate description tokens that are not supported by the source inputs. This phenomenon seriously damages the applicability of neural language generation models in practice where information accuracy is vital.

Hallucination has been observed in various conditional NLG tasks such as image captioning (Rohrbach et al., 2018), data-to-text generation (Wiseman et al., 2017; Nie et al., 2019; Parikh et al., 2020), abstractive summarization (Cao et al., 2018; Durmus et al., 2020), and neural machine

translation (NMT) (Müller et al., 2019). These studies tackle hallucinations within a specific task and give possible explanations of why hallucinations occur. For example, Rohrbach et al. (2018) attributes object hallucination in image captioning to visual misclassification and over-reliance on language priors; Nie et al. (2019) believes hallucination in neural surface realization comes from the misalignment between meaning representations and their corresponding references in the dataset; Müller et al. (2019) claims that hallucinations in NMT are mainly due to domain shift.

We believe that there is a common theme across all the hallucination explanations in conditional NLG tasks: predictive uncertainty. In language generation, predictive uncertainty quantifies the entropy of the token probability distributions a model predicts. There are multiple sources of uncertainty. Two major ones frequently studied are aleatoric and epistemic uncertainties, where the former comes from the data or measurements, and the latter is concerned with the model. With recent progress in Bayesian neural networks (BNNs) (Hinton and Van Camp, 1993; Neal, 1995) and uncertainty quantification (Blundell et al., 2015; Gal and Ghahramani, 2016; Lakshminarayanan et al., 2017), we are able to quantify both parts of predictive uncertainty in neural NLG.

This study draws connections between hallucination and predictive uncertainty and empirically investigates their relationship in image captioning and data-to-text generation tasks. We propose an uncertainty-aware beam search algorithm to reduce the chance of hallucination by penalizing parts or the entirety of the predictive uncertainty during model decoding. We find that the choice of uncertainty matters, and penalizing epistemic uncertainty yields better results compared to penalizing aleatoric or total uncertainty. Our contributions are:

2734

- We draw connections between hallucination and predictive uncertainty across various conditional natural language generation tasks and empirically investigate their relationship.

- We propose an uncertainty-aware beam search approach for hallucination reduction to demonstrate that lowering uncertainty can lead to less hallucination.

- We show that uncertainty decomposition helps to achieve better trade-offs between hallucination and performance.

2 Hallucination and Predictive Uncertainty

2.1 Hallucination Probability

In general, hallucination refers to the phenomenon where the model generates false information not supported by the input. For example, in the context of image captioning, hallucination can be defined as generating captions that contain descriptions not present in the given image. Let (x, y) be the pair of variables at interest where x is some structured data containing facts and y is a natural language sentence based on the facts. The task is to learn the conditional distribution of $p(y|x)$ in order to generate sentence y given any new input x. Most neural approaches break the probability into a sequence of single token predictions:

$$p(y|x) = p(y_1|x) \prod_{i=2}^{k} p(y_i|x, y_1, \cdots, y_{i-1}) \quad (1)$$

where $\{y_1, \cdots, y_k\}$ is the collection of tokens in sentence y. We denote $c_i = \{x, y_1, \cdots, y_{i-1}\}$ as the context of the i-th prediction in the following sections for simplicity.

Apparently, hallucination is context-dependent which means we need to look at a certain context c_i and determine whether the next token prediction y_i is hallucinated or not. Let $\mathcal{V}_h^{(c_i)}$ denote the set of tokens that are considered false information given the current context c_i and $\mathcal{V}$ the whole vocabulary. Consider a random sampling decoder where a token is generated based on the predicted categorical distribution. i.e. $\text{Cat}(|\mathcal{V}|, p(y_i|c_i))$. The probability of hallucination at the current step is simply:

$$P(y_i \in \mathcal{V}_h^{(c_i)}) = \sum_{v \in \mathcal{V}_h^{(c_i)}} p(y_i = v|c_i) \quad (2)$$

Practically, it is hard to automatically determine the context-dependent set $\mathcal{V}_h^{(c_i)}$. Task-specific heuristics are often used to determine which tokens are hallucinated. In specific restrictive applications, the context-dependent set can be relaxed to a context-independent one to reduce the complexity of determining hallucination.

2.2 Relationship with Predictive Uncertainty

We use entropy to measure the predictive uncertainty in this work. The total uncertainty of predicting token y_i is:

$$
\begin{aligned}
&H(y_i|c_i) \\
&= -\sum_{v \in \mathcal{V}} p(y_i = v|c_i) \log p(y_i = v|c_i) \\
&= -\sum_{v \in \mathcal{V} \setminus \mathcal{V}_h^{(c_i)}} p(y_i = v|c_i) \log p(y_i = v|c_i) \\
&\quad -\sum_{v \in \mathcal{V}_h^{(c_i)}} p(y_i = v|c_i) \log p(y_i = v|c_i) \quad (3)
\end{aligned}
$$

From Equation 3, we can see that there are two sources of uncertainty for the token predictions: one from the uncertainty of choosing suitable tokens to describe the input; another from some unsuitable tokens attaining considerable probability mass either by being confusing in the current context or due to an insufficiently trained system.

The second source of uncertainty is directly related to hallucination probability. Although no monotonic relationship can be derived, a near-zero hallucination probability requires a near-zero value of the second source of uncertainty. This observation prompts us to investigate the relationship between hallucination and predictive uncertainty in practice. Intuitively, the higher the predictive uncertainty is, the more probable some of the probability mass gets assigned to unsuitable tokens.

2.3 Uncertainty Decomposition

There are often two types of uncertainties frequently mentioned in uncertainty quantification literature: epistemic and aleatoric uncertainty (Der Kiureghian and Ditlevsen, 2009; Kendall and Gal, 2017; Depeweg et al., 2018). Epistemic uncertainty reflects the uncertainty on model weights, and aleatoric uncertainty concerns inherent uncertainty in the data or measurement. We are interested in whether the relationship with hallucination is the same for both types of uncertainties.

	Token 1	Token 2	Token 3
Prediction 1	0.33	0.33	0.33
Prediction 2	0.33	0.33	0.33
Prediction 3	0.33	0.33	0.33

(a)

	Token 1	Token 2	Token 3
Prediction 1	0.98	0.01	0.01
Prediction 2	0.01	0.98	0.01
Prediction 3	0.01	0.01	0.98

(b)

Figure 1: Examples of predictions with (a) high aleatoric but low epistemic uncertainty; and (b) high epistemic but low aleatoric uncertainty.

Bayesian deep learning approaches (Blundell et al., 2015; Gal and Ghahramani, 2016; Lakshminarayanan et al., 2017) are widely studied for uncertainty quantification with neural networks. Following the notations in Section 2.2, the predictive distribution of $p(y_i|c_i)$ can be written as:

$$p(y_i|c_i) = \int_w p(y_i|c_i, w) q(w) dw \qquad (4)$$

where w parameterizes the neural network that makes predictions and $q(w)$ denotes the approximate posterior distribution of the weights w given the training data. Notice that if we fix the weights w, $H(y_i|c_i, w)$ represents the entropy that is unrelated to the uncertainty of the model weights. Therefore the aleatoric part of the predictive uncertainty can be calculated with $\mathbb{E}_{q(w)}[H(y_i|c_i, w)]$. The epistemic part of the uncertainty is the difference between the total and the aleatoric uncertainty as shown below:

$$u_{al}(y_i|c_i) = \mathbb{E}_{q(w)}[H(y_i|c_i, w)] \qquad (5)$$
$$u_{ep}(y_i|c_i) = H(y_i|c_i) - \mathbb{E}_{q(w)}[H(y_i|c_i, w)] \qquad (6)$$

In this study, the aleatoric and epistemic parts of predictive uncertainty are estimated using deep ensembles (Lakshminarayanan et al., 2017). More concretely, denote the model predictions as $\{p_m(y_i|c_i)\}_{m=1}^M$ and the aggregated prediction as $p(y_i|c_i) = \frac{1}{M}\sum_{m=1}^M p_m(y_i = v|c_i)$, aleatoric and

epistemic uncertainties are calculated as:

$$u_{al}(y_i|c_i) = \frac{1}{M}\sum_{m=1}^M H_m(y_i|c_i) \qquad (7)$$
$$u_{ep}(y_i|c_i) = H(y_i|c_i) - u_{al}(y_i|c_i) \qquad (8)$$

where $H_m(y_i|c_i)$ and $H(y_i|c_i)$ are the entropy of $p_m(y_i|c_i)$ and $p(y_i|c_i)$ respectively.

Intuitively, in the case of deep ensembles, aleatoric uncertainty measures the average spread of all model predictions, while epistemic uncertainty measures the agreement among all model predictions. Examples with three possible tokens are illustrated in Figure 1.

3 Case Study: Image Captioning

In this section, we analyze image captioning models trained on MSCOCO (Chen et al., 2015) data set.

3.1 Hallucination Probability at Different Uncertainty Levels

The first question we want to investigate is whether hallucination probabilities change at different predictive uncertainty levels. Some experimental settings are listed below.

Model architecture We consider four different image captioning models: **FC** model (Rennie et al., 2017) where image features are used to initialize the RNN decoder; **Att2In** model from (Rennie et al., 2017) applies attention on image features and feeds it into the decoder LSTM (Hochreiter and Schmidhuber, 1997) cell gate; **BUTD** model from (Anderson et al., 2018) uses bottom-up attention which operates at the level of objects and other salient image regions; **Transformer** model where transformers (Vaswani et al., 2017) are used in the encoder-decoder structure for generation. All models are implemented in the open source framework by Luo et al. (2018)[1].

Training We consider the same data split from (Karpathy and Fei-Fei, 2015). All models are trained with batch size 50 for 30 epochs with Adam optimizer (Kingma and Ba, 2014). Evaluations are done on the Karpathy test set.

Hallucination and uncertainty evaluation As in (Rohrbach et al., 2018), synonyms for all possible MSCOCO objects are used to determine

---

[1]https://github.com/ruotianluo/self-critical.pytorch

Model	Action hallucination % at uncertainty level					
	≤ 0.8	0.8 - 1.6	1.6 - 2.4	2.4 - 3.2	3.2 - 4.0	> 4.0
FC	0.00	0.00	2.27	12.86	15.71	31.03
Att2In	0.00	0.00	3.39	6.58	12.07	22.03
BUTD	0.00	2.94	1.92	12.77	17.24	25.53
Transformer	2.99	5.48	6.58	8.82	12.00	43.75

Table 1: Action hallucination percentages at different levels of predictive uncertainty. Action predictions with higher uncertainty are more prone to hallucination.
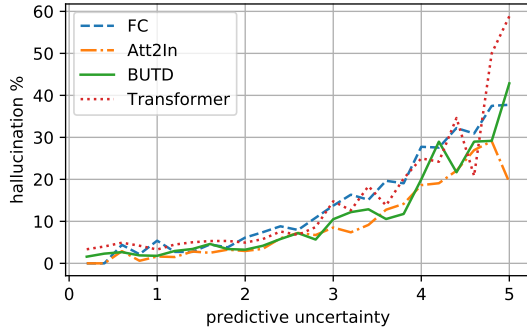

Figure 2: Object hallucination chance at different predictive uncertainty levels. Higher predictive uncertainty corresponds to a higher level of hallucination percentage across all models.

whether an object generated by the captioning model is hallucinated. ==Hallucination probabilities are calculated by binning all object token prediction entropy and counting the percentage of hallucinated objects in each bin.==

3.2 Results and Discussions

Figure 2 shows the object hallucination percentages at different predictive uncertainty levels. At higher uncertainty levels, the generated objects are more likely to be hallucinated. The results are consistent across four different models. The transformer model seems to have a higher hallucination chance at high uncertainty levels than the other three models. However, this does not indicate Transformer models hallucinate more. In fact, the transformer model has an overall lowest hallucination percentage among all four models.

Beyond object hallucination Aside from object hallucination, we also analyze verbs generated by the models to see whether a similar relationship holds for other types of token generations. The same models and training procedures are adopted. We extract all present continuous tense verbs from the generated captions using spaCy part-of-speech

Model	Correlation coefficient	
	epistemic	aleatoric
FC	0.313	0.299
BUTD	0.334	0.228
Att2In	0.360	0.268
Transformer	0.269	0.131

Table 2: Pearson correlation coefficients between hallucination and epistemic/aleatoric uncertainty in image captioning task. Epistemic uncertainty is more indicative of hallucination across four models.

tagger[2] and manually label whether they are suitable to describe the corresponding images. There are approximately 3500 generated captions containing verbs, and 400 are annotated for each model. We refer to unsuitable verbs generated in the captions as action hallucinations.

Action predictions are binned according to their uncertainty values, and the results are shown in Table 1. We can observe that action tokens with higher predictive uncertainty are also more likely to be hallucinated. Noticeably, the transformer model also has a higher action hallucination rate at high uncertainty levels.

Examples of predictions with high and low uncertainty Figure 3 shows some example images and their captions generated from a BUTD model on the test set. The token predictions of interests and the corresponding uncertainty values are highlighted in bold and italic, respectively. We observe that highly uncertain predictions often correspond to unusual textures, features resembling the predicted tokens, or blurred images. For example, Figure 3(b) shows a motorcycle covered in vines; Figure 3(d) shows candles in the background which resemble cakes; Figure 3(f) is blurred.

Epistemic and aleatoric uncertainties As we could decompose the total uncertainty into two parts, we are interested in which part is more indicative of hallucination. Table 2 shows the Pear-

---

[2]https://spacy.io

(a) a red and black **motor-cycle** (*0.58*) parked in a parking lot

(b) a **motorcycle** (*4.80*) is parked on a dock with a bird perched on top of it

(c) a bride and groom cutting their wedding **cake** (*0.09*)

(d) a woman holding a cup and a **cake** (*5.29*)

(e) a man standing on a tennis court **holding** (*0.81*) a racquet

(f) a young man is **holding** (*4.76*) a skateboard in his hand

(g) a group of children sitting at a table **eating** (*1.00*) pizza

(h) a man is **eating** (*4.01*) a hot dog at a restaurant

Figure 3: Examples of token predictions generated with the BUTD model with high and low uncertainty values for objects (top) and actions (bottom). Numbers in italic are predictive uncertainty values for the token predictions preceding them. The examples are cherry-picked.

son correlation coefficients between hallucination (binary) and epistemic/aleatoric uncertainty for all four models. We can see that both parts of uncertainty are weakly correlated with hallucination, while epistemic uncertainty is more indicative of hallucination across all four models compared to aleatoric uncertainty.

4 Case Study: Data-to-text Generation

Data-to-text generation (Kukich, 1983; McKeown, 1992) is a task to generate textual content conditioned on input content in the form of structured data such as tables. Neural models are prone to hallucination in data-to-text generation tasks compared to traditional template-based systems, and methods are proposed to improve faithfulness (Wiseman et al., 2017; Nie et al., 2019; Tian et al., 2019). In this section, we discuss the relationship between predictive uncertainty and hallucination in data-to-text generation with ToTTo dataset (Parikh et al., 2020).

4.1 Generation Quality and Average Uncertainty

We conduct token-level analysis in Section 3. Now we take a different route and analyze sentence-level quality with different average predictive uncertainty values. Experiment settings are described below.

Dataset   ToTTo dataset consists of tables from English Wikipedia articles with their correspond-

ing metadata, such as page title and section title. Candidate description texts are modified by annotators to pair with each table. Relevant table cells supporting the description texts are highlighted by the annotators as well. There are 120,761 table-text pairs in training, 7,700 in validation, and 7,700 in test. We use the baseline standard linearization approach to represent the highlighted portions of the tables along with their corresponding metadata (referred to as subtable with metadata in (Parikh et al., 2020)).

Model architecture and training   We use a standard sequence-to-sequence model with attention (Bahdanau et al., 2015; Luo et al., 2018) for analysis. LSTM with 512 hidden size is used for both the encoder and the decoder. Adam optimizer with learning rate 1e-3 is used for the optimization. The model is trained with cross-entropy loss for 20 epochs. The checkpoint with the best validation loss is chosen for the evaluation. The implementation is done using fairseq (Ott et al., 2019)[3].

Evaluation   We evaluate the average predictive uncertainty for all generated sentences in the validation set and select the top, bottom, and middle 5% for comparison. BLEU score (Papineni et al., 2002) is used as an automatic metric to evaluate the similarity to the references; further manual annotations are done to evaluate the *fluency*, *faithfulness* (precision), and *coverage with respect to reference*

---

[3]https://github.com/pytorch/fairseq

Unc. Level	Avg Unc.	BLEU	Fluency (%)	Faithfulness (%)	Less/Neutral/More Coverage w.r.t. Ref
High	1.83 - 3.74	10.2	46.0	41.3	79.4 / 15.9 / 04.7
Medium	0.83 - 0.89	31.5	87.3	78.9	35.2 / 47.9 / 16.9
Low	0.04 - 0.27	72.8	100.0	99.0	22.2 / 70.1 / 07.7

Table 3: Evaluation results for candidates with high, medium, and low average predictive uncertainty values for ToTTo validation set. Unc. denotes uncertainty. Higher uncertainty candidates have lower quality and higher chance of being hallucinated/unfaithful w.r.t. the input tables.

(recall) of the generated sentences. Particularly, faithfulness reflects how likely the generated sentences hallucinate facts that are not supported by the tables. More details of the human evaluation metrics are described in (Parikh et al., 2020). The goal is to measure how different the generation qualities are for candidates with varying average predictive uncertainties.

4.2 Results and Discussions

Table 3 summarizes the evaluation results for candidates with varying uncertainty values. It is obvious that candidates with higher average predictive uncertainty values are less fluent and more likely to contain hallucinations. Another interesting observation from Table 3 is that the generated sentences with medium average uncertainty are more likely (16.9%) to cover more table facts than the references compared to the ones with high (4.7%) and low (7.7%) average uncertainty. One possible explanation is that some table facts that are not always included in the references, when generated, have higher predictive uncertainty values than the facts that are almost always included in the references. Therefore, generated sentences with low uncertainty tend to include less but more confident facts considered by the model.

5 Reducing Hallucination

5.1 Uncertainty-Aware Beam Search

Because of the positive correlation between hallucination probability and predictive uncertainty, it is straightforward to incorporate uncertainty into the caption generation process to reduce hallucination. Beam search is the most used approximate decoding method in language generation. It keeps track of the top-B scored candidates at each generation step and considers all single token extensions of the current candidates.

More formally, denote the set of B candidates in the beam at time step $t-1$ as $\mathcal{Y}_{t-1} = \{\mathbf{y}_{t-1}^{(b)}\}_{b=1}^{B}$. All possible single token extensions of the can-

didates in $\mathcal{Y}_{t-1}$ form a set $\mathcal{C}_t = \{\mathbf{y} \mid \mathbf{y}_{t-1} \in \mathcal{Y}_{t-1} \land y_t \in \mathcal{V}\}$. Beam at step t is then formed as:

$$\mathcal{Y}_t = \arg\max_{\mathbf{y}_1\cdots\mathbf{y}_B\in\mathcal{C}_t} \sum_{b=1}^{B} \log p(\mathbf{y}_b|\mathbf{x})$$
$$s.t. \quad \mathbf{y}_i \neq \mathbf{y}_j \quad \forall i \neq j \qquad (9)$$

Uncertainty-aware beam search (UABS) adds a weighted penalty term in the beam search objective to balance between log probability and predictive uncertainty of the selected candidates. Let $u(\mathbf{y}|\mathbf{x})$ be the function to measure the aggregated predictive uncertainty of candidate $\mathbf{y}$ given input $\mathbf{x}$, uncertainty-aware beam search updates the beam at step t according to the following equation:

$$\mathcal{Y}_t = \arg\max_{\mathbf{y}_1\cdots\mathbf{y}_B\in\mathcal{C}_t} \sum_{b=1}^{B} \log p(\mathbf{y}_b|\mathbf{x}) - \lambda u(\mathbf{y}_b|\mathbf{x})$$
$$s.t. \quad \mathbf{y}_i \neq \mathbf{y}_j \quad \forall i \neq j \qquad (10)$$

where $\lambda \geq 0$ is the weight controlling the degree to which we want to penalize decoding uncertainty. Larger λ leads to candidates with smaller predictive uncertainty. In practice, this can be done by subtracting the weighted uncertainty term from the aggregated log probability scores at each decoding step before choosing top-B candidates.

An important decision in using uncertainty-aware beam search is the choice of uncertainty term $u(\mathbf{y}|\mathbf{x})$. We could use either the aleatoric or epistemic part of the predictive uncertainty or both. We compare these choices and discuss the results in the next section.

5.2 Image Captioning Results

With larger weights on the uncertainty penalty term, log probabilities of the decoded sentences drop. Therefore, we expect to see a trade-off between the quality of generated captions and the chance of hallucination.

We empirically examine the trade-offs on the image captioning models with different uncertainty

Image	UABS results with weight λ		
	0	20	80
	a vase filled with flowers sitting on top of a table	a vase filled with lots of white flowers	there is a vase that has flowers in it
	a wooden cutting board topped with lots of food	a wooden cutting board topped with lots of food	a cutting board that has a bunch on it

Table 4: Two examples of epistemic UABS results with varying penalty weights on the image captioning data set. In the first example the model successfully avoids hallucination of a table with $\lambda = 20$ while in the second example it is unable to change the generated caption until larger penalty weight is set.

(a) FC

(b) Att2In
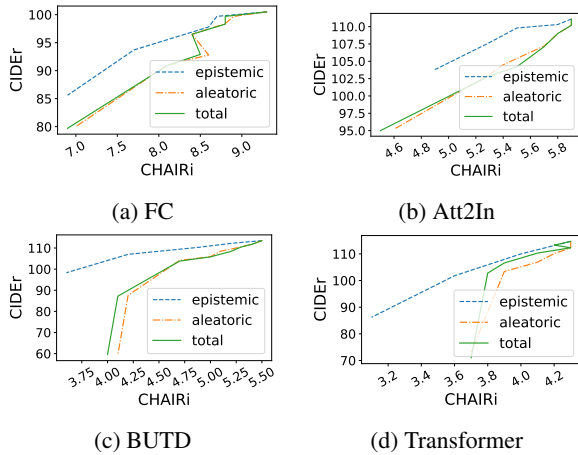
(c) BUTD

(d) Transformer

Figure 4: CIDEr plotted against CHAIRi scores of captions generated with UABS with different uncertainty penalty weights. Lower CHAIRi score indicates less hallucination. Upper-left is better. Penalizing epistemic uncertainty in UABS achieves the best results.

	λ	avg. len.	# obj.	hal. %	gen. %
ref.		10.44	6114	0	-
base	0	9.31	7328	5.5	0
epist.	10	9.21	7195	5.2	0
	20	9.16	7078	4.9	0.2
	40	9.15	6912	4.2	1.5
	80	9.12	6493	3.6	4.6
aleat.	0.1	9.32	7250	5.4	0
	0.4	9.32	7051	5.1	0
	1.0	9.33	6800	4.7	1.0
	4.0	9.43	4349	4.1	28.4

Table 5: Average sentence length and total number of objects detected in the captions generated by BUTD model with varying uncertainty penalty weight λ. Penalizing epistemic uncertainty leads to slightly shorter lengths. Number of objects mentioned by the captions decreases with increasing λ. **gen. %** denotes percentage of generic responses. It is moderate with epistemic penalized results but can be very high if aleatoric uncertainty is heavily penalized.

choices for the penalty term. We use a five-model ensemble for each of the four model architectures to estimate aleatoric and epistemic uncertainties. Due to the different magnitudes of aleatoric and epistemic uncertainties, we choose penalty weight λ from $[0.1, 0.2, 0.4, 0.8, 1.0, 2.0, 4.0]$ for aleatoric and total uncertainty and $[10, 20, 40, 80]$ for epistemic uncertainty.

Figure 4 shows the trade-offs between CIDEr (Vedantam et al., 2015) and CHAIRi (Rohrbach et al., 2018) scores of captions generated with uncertainty-aware beam search with different uncertainty choices and penalty weights. A smaller value of CHAIRi indicates the model is less likely to generate hallucinated objects, and a higher CIDEr indicates better caption quality. Therefore

an approach that is to the upper left of another is better. As the penalty weight increases, we observe a decrease in both the CHAIRi and the CIDEr scores across all models.

Table 4 shows two examples of different generated captions using epistemic UABS with varying penalty weights. In the first example, we can see that a medium penalty weight of 20 not only helps avoid the hallucination of a table but also adds correct information about the color of the flowers. In the second example, a medium penalty weight is unable to change the generated caption.

Regarding the choice of uncertainty, it is notable that when penalizing epistemic uncertainty, the generated captions achieve higher CIDEr scores than penalizing aleatoric or total uncertainty. We

λ	BLEU	Fluency (%)	Faithfulness (%)	Less/Neutral/More Coverage w.r.t. Ref
0	40.1	92	79	34 / 60 / 6
10	33.6	83	84	41 / 51 / 8
20	27.4	73	80	52 / 42 / 6

Table 6: Evaluation results for candidates decoded with different penalty weights for UABS on ToTTo validation set. Epistemic uncertainty is used for uncertainty penalization. Faithfulness first increases, then decreases to the same level as regular beam search results as we increase the penalty weight λ.

Reference	UABS results with weight λ		
	0	10	20
barrows scored 164 net points in virgin islands at the 2008 summer olympics.	in virgin islands at the 2008 summer olympics, barrows iii received 164 points.	in virgin islands at the 2008 summer olympics, barrows received 164 points.	thomas barrows received a total score of 164.
janet gaynor won the first academy award for best actress for her performance in the 7th heaven (1927 film).	janet gaynor won the academy award for best actress for his performance in janet gaynor.	janet gaynor won the academy award for best actress.	janet gaynor won an academy award for best actress.

Table 7: Two examples of UABS results with varying penalty weights on the ToTTo validation set. Blue tokens are correct table facts that are dropped by candidates generated with larger penalty weights; red tokens are incorrect/hallucinated facts that are dropped with larger penalty weights. In general, UABS with larger weights tend to produce sentences with less information that the model is more confident with.

hypothesize that epistemic uncertainty indicates the uncertainty of model weights. By penalizing epistemic uncertainty, we encourage the model to take the prediction path where it is well-calibrated. On the other hand, penalizing aleatoric uncertainty encourages the model to make low entropy predictions in all contexts regardless of the actual data distributions.

Table 5 shows the average sentence length, the number of objects, the percentage of hallucinations, and the percentage of generic responses in the captions generated by the BUTD model with different uncertainty choices and penalty weights on the test set. We can see that when penalizing epistemic uncertainty, UABS results in slightly shorter caption candidates. Both the number of objects and hallucination percentage decrease as we increase the weight λ. Interestingly, when penalizing aleatoric uncertainty, sentence length stays approximately the same despite lower CIDEr scores, as shown in Figure 4. Further investigation shows that this is partly due to an increasing number of generic captions such as "*there is no image here to provide a caption for*". Penalizing epistemic uncertainty is much less likely to result in such generic captions. We can see that when increasing λ from 1.0 to 4.0 with aleatoric UABS, the percentage of generic responses jumps drastically from 1.0% to 28.4%. In comparison, epistemic UABS keeps the generic response rates low while achieving lower hallucination rates.

5.3 Data-to-text Results

We also evaluate the effect of UABS on the ToTTo dataset. We choose to penalize epistemic uncertainty due to its better performances than aleatoric uncertainty, as shown in the previous section. A five-model deep ensemble is used to quantify the epistemic uncertainty and generate results with UABS. We compare the BLEU score and three human evaluation metrics among results generated with different uncertainty penalty weights. 100 generation results are randomly selected and evaluated for each penalty weight choice. The results are shown in Table 6. We can see that a relatively small penalty weight leads to a reduced hallucination chance (hence more faithful) with a cost on the BLEU score and fluency.

To qualitatively examine the sentences generated with different λ values, we show example results on the ToTTo validation set in Table 7. We can see that with larger penalty weights, the UABS results drop certain statements that the model deems less confident regardless of the correctness. This results in shorter but more confident predictions for UABS results with a larger uncertainty penalty.

6 Related Work

Hallucination There are many pieces of anecdotal evidence of hallucination presented in various NLG tasks. Most recently, researchers started investigating the phenomenon systematically. Rohrbach et al. (2018) analyzes object hallucination focusing on the objects that appeared in the MSCOCO segmentation challenge. They propose the CHAIR metric to quantify the severity of object hallucination. They find that the models tend to make predictions consistent with a language model trained on the captions instead of a model trained to predict objects in an image. Therefore hallucination is caused by an over-reliance on the language priors. Nie et al. (2019) believes that the origin of the hallucination problem in neural surface realization comes from the data side. More specifically, datasets used for NLG systems often include instances with information misalignment between the input structure and the output text. They propose integrating a language understanding module for iterative data refinement to better align meaning representations and output text. Müller et al. (2019) examines hallucination in neural machine translation and observes that the phenomenon is most common in out-of-domain settings. They empirically compare several strategies to improve domain robustness in NMT and find that a combination of reconstruction and a noisy channel model for reranking is most effective.

These observations are consistent with our findings. For example, domain shift and data misalignment are known to lead to a higher level of epistemic uncertainty (Kendall and Gal, 2017) which makes hallucination a more severe problem.

Uncertainty quantification Uncertainty quantification has attracted more attention recently due to the progress in Bayesian deep learning. Bayes by backprop (Blundell et al., 2015), Monte Carlo dropout (Gal and Ghahramani, 2016), and deep ensembles (Lakshminarayanan et al., 2017) are examples of popular Bayesian approaches to evaluate uncertainty with deep neural models. Kendall and Gal (2017) investigates the benefits of modeling epistemic and aleatoric uncertainty in vision tasks such as semantic segmentation and depth regression. They show that it is important to model aleatoric uncertainty with large datasets and real-time applications and epistemic uncertainty with small datasets and safety-critical applications. Other applications of uncertainty quantification have been explored in the context of time series predictions (Zhu and Laptev, 2017), natural language processing tasks (Xiao and Wang, 2019), etc. More broadly, prediction entropy has been analyzed in different neural language generation tasks (Ott et al., 2018; Xu et al., 2020). Depeweg et al. (2018) shows how to extract and decompose uncertainty in Bayesian neural networks with latent variables for decision-making purposes. They show that active learning and risk-sensitive reinforcement learning both benefit from uncertainty decomposition.

7 Discussion and Conclusions

We investigate the relationship between hallucination and predictive uncertainty in image captioning and data-to-text generation tasks and show that predictions with higher uncertainty are more prone to hallucination. In particular, epistemic uncertainty is more indicative of hallucination than aleatoric uncertainty. We propose uncertainty-aware beam search to incorporate uncertainty into the decoding process to reduce hallucination. We show that uncertainty decomposition helps the proposed beam search variant to achieve a better performance-hallucination trade-off. Specifically, penalizing epistemic uncertainty yields better results compared to penalizing aleatoric or total uncertainty.

In this work, we analyze uncertainty from the token level. This might be restrictive because uncertainty corresponds to the current prediction context instead of the predicted token. The relationship between hallucination and uncertainty, therefore, can be much more complicated than a linear one. It is still possible to produce hallucinated information with a very confident model. The proposed UABS reduces hallucination by limiting the total uncertainty of the generated text. As a result, it might lead to shorter generations and lower generation quality. Devising more sophisticated uncertainty-aware training and decoding methods with less adverse effects on the generation quality is a future direction to explore.

Acknowledgement

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

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, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. 2015. Weight uncertainty in neural networks. *arXiv preprint arXiv:1505.05424*.

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

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Stefan Depeweg, Jose-Miguel Hernandez-Lobato, Finale Doshi-Velez, and Steffen Udluft. 2018. Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. In *International Conference on Machine Learning*, pages 1184–1193. PMLR.

Armen Der Kiureghian and Ove Ditlevsen. 2009. Aleatory or epistemic? does it matter? *Structural safety*, 31(2):105–112.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059.

Geoffrey E Hinton and Drew Van Camp. 1993. Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the sixth annual conference on Computational learning theory*, pages 5–13.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137.

Alex Kendall and Yarin Gal. 2017. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems*, pages 5574–5584.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Karen Kukich. 1983. Design of a knowledge-based report generator. In *Proceedings of the 21st annual meeting on Association for Computational Linguistics*, pages 145–150. Association for Computational Linguistics.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in neural information processing systems*, pages 6402–6413.

Ruotian Luo, Brian Price, Scott Cohen, and Gregory Shakhnarovich. 2018. Discriminability objective for training descriptive captions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6964–6974.

Kathleen McKeown. 1992. *Text generation*. Cambridge University Press.

Mathias Müller, Annette Rios, and Rico Sennrich. 2019. Domain robustness in neural machine translation. *arXiv preprint arXiv:1911.03109*.

Radford M Neal. 1995. *BAYESIAN LEARNING FOR NEURAL NETWORKS*. Ph.D. thesis, University of Toronto.

Feng Nie, Jin-Ge Yao, Jinpeng Wang, Rong Pan, and Chin-Yew Lin. 2019. A simple recipe towards reducing hallucination in neural surface realisation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2673–2679.

Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Analyzing uncertainty in neural machine translation. In *International Conference on Machine Learning*, pages 3956–3965.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of*

the 40th annual meeting on association for computational linguistics, pages 311–318. Association for Computational Linguistics.

Ankur P Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. Totto: A controlled table-to-text generation dataset. *arXiv preprint arXiv:2004.14373*.

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7008–7024.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045.

Ran Tian, Shashi Narayan, Thibault Sellam, and Ankur P Parikh. 2019. Sticking to the facts: Confident decoding for faithful data-to-text generation. *arXiv preprint arXiv:1910.08684*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

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

Yijun Xiao and William Yang Wang. 2019. Quantifying uncertainties in natural language processing tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7322–7329.

Jiacheng Xu, Shrey Desai, and Greg Durrett. 2020. Understanding neural abstractive summarization models via uncertainty. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6275–6281.

Lingxue Zhu and Nikolay Laptev. 2017. Deep and confident prediction for time series at uber. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 103–110. IEEE.