arXiv:2202.03629v2 [cs.CL] 22 Feb 2022

Survey of Hallucination in Natural Language Generation

ZIWEI JI, NAYEON LEE, RITA FRIESKE, TIEZHENG YU, DAN SU, YAN XU, ETSUKO ISHII, YEJIN BANG, ANDREA MADOTTO, and PASCALE FUNG, Center for Artificial Intelligence Research (CAiRE), Hong Kong University of Science and Technology, Hong Kong

Natural Language Generation (NLG) has improved exponentially in recent years thanks to the development of sequence-to-sequence deep learning technologies such as Transformer-based language models. This advancement has led to more fluent and coherent natural language generation, leading to improved development in downstream tasks such as abstractive summarization, dialogue generation and data-to-text generation. However, it is also apparent that deep learning based generation is prone to hallucinate unintended texts, which degrades the system performance and fail to meet user expectations in many real-world scenarios. In order to address this issue, there have been studies in measuring and mitigating hallucinated texts. However there has not been a comprehensive review of the state-of-the-art in hallucination detection and mitigation.

In this survey, we provide a broad overview of the research progress and challenges in the hallucination problem of NLG. The survey is organized into two parts: (1) a general overview of metrics, mitigation methods, and future directions; (2) an overview of task-specific research progress for hallucinations in a large set of downstream tasks, namely abstractive summarization, dialogue generation, generative question answering, data-to-text generation, and machine translation. This survey serves to facilitate collaborative efforts among researchers in tackling the challenge of hallucinated texts in NLG.

CCS Concepts: • **Computing methodologies → Natural language generation**; **Neural networks**.

Additional Key Words and Phrases: Hallucination, Intrinsic Hallucination, Extrinsic Hallucination, Faithfulness in NLG, Factuality in NLG, Consistency in NLG

Contents

Authors' address: Ziwei Ji, zjiad@connect.ust.hk; Nayeon Lee, nyleeaa@connect.ust.hk; Rita Frieske, rita.frieske@ust.hk; Tiezheng Yu, tyuah@connect.ust.hk; Dan Su, dsu@connect.ust.hk; Yan Xu, yxucb@connect.ust.hk; Etsuko Ishii, eishii@connect.ust.hk; Yejin Bang, yjbang@connect.ust.hk; Andrea Madotto, amadotto@connect.ust.hk; Pascale Fung, pascale@ece.ust.hk, Center for Artificial Intelligence Research (CAiRE), Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong.

. 0360-0300/2022/2-ART $15.00
https://doi.org/

1 INTRODUCTION

Natural Language Generation (NLG) is one of the crucial yet challenging sub-fields of Natural Language Processing (NLP). NLG techniques are used in many downstream tasks such as summarization, dialogue generation, generative question answering (GQA), data-to-text generation, and machine translation. Recently, the rapid development of NLG has seized the imagination of many thanks to the advances in deep learning technologies, especially Transformer [177]-based models like BERT [28], BART [95], GPT-2 [141], and GPT-3 [17]. The conspicuous development of NLG tasks attracted the attention of many researchers and has led to an increased effort in the field.

With the advancement of NLG models, attention towards their limitations and potential risks has also increased. Some early work focused on the potential pitfalls of utilizing the standard likelihood maximization-based objective in training and decoding of NLG models. They discovered that such likelihood maximization approaches could result in *degeneration*, which refers generated output that is bland, incoherent, or gets stuck in repetitive loops [71, 185]. Concurrently, it is discovered that NLG models often generate texts that are nonsensical, or unfaithful to the provided source input [82, 145, 150, 178]. Researchers started referring to such undesirable generation as *hallucination* [117] [1].

Hallucination in NLG is concerning because it hinders performance and raises safety concerns for real-world applications. For instance, in the medical application, a hallucinatory summary generated from a patient information form could pose a risk to the patient. It may provoke a life-threatening incident for a patient if the instructions of a medicine generated by machine translation are hallucinatory. Not only this, hallucination can lead to potential privacy violation risks. Carlini et al. [21] demonstrate that language models can be prompted to recover and generate sensitive personal information from the training corpus (i.e., email address, phone/fax number, and physical address). Such memorization and recovery of training corpus are considered one form of hallucination because the model is generating text that is not "faithful" to the source input content (i.e., such private information does not exist in the source input).

There are active efforts to address hallucination for various NLG tasks. We believe that analyzing hallucinatory content in different NLG tasks and investigating their relationship will strengthen our understanding of this phenomenon and encourage the unification of efforts from different NLG fields. Meanwhile, little has been done to understand hallucinations from a broader perspective that encompasses all major NLG tasks. To the best of our knowledge, there only exist hallucination surveys that focus on specific tasks like abstractive summarization [75, 117] and translation [91]. In short, our survey offers a comprehensive analysis of existing research on the phenomenon of hallucination in different NLG tasks, which can help researchers to have a high-level insight derived from the similarities and differences of different approaches. Furthermore, given the various stages of development in studying hallucination from different tasks, researchers can refer to and be inspired by such a survey on concepts, metrics, and mitigation methods.

Organization of this Survey. The remainder of this survey is organized as follows. First, we provide an overview of the hallucination problem in NLG by discussing the definition and categorization, contributors, metrics, and mitigation methods of hallucinations from Section 2 to Section 6. The second part of our survey discusses the hallucination problem associated with specific

[1]The term "hallucination" first appeared in Computer Vision (CV) in Baker and Kanade [6] and carried more positive meanings, such as superresolution [6, 105], image inpainting [47], and image synthesizing [209]. Such hallucination is something we take advantage of rather than avoid in CV. Nevertheless, recent works have started to refer to a specific type of error as "hallucination" in image captioning [14, 150] and object detection [5, 80], which denotes non-existing objects detected or localized incorrectly at their expected position. The latter conception is similar to hallucination in NLG.

types of NLG tasks: abstractive summarization in Section 7, dialogue generation in Section 8, GQA in Section 9, data-to-text generation in Section 10, and finally, machine translation in Section 11.

2 DEFINITIONS

In the general context outside of NLP, hallucination is a psychological term referring to a particular type of perception [51, 110]. Blom [15] defines hallucination as "a percept, experienced by a waking individual, in the absence of an appropriate stimulus from the extracorporeal world." Simply put, a hallucination is an unreal perception that feels real. The undesired phenomenon of **"NLG models generating unfaithful or nonsensical text"** shares a similar characteristic with such psychological hallucinations – explaining the choice of the terminology. Hallucinated texts give the impression of being fluent and natural despite being unfaithful and nonsensical; hallucinated texts appear to be grounded in the real context provided, although it is actually hard to specify or verify the existence of such contexts. Moreover, just like psychological hallucination is hard to tell apart from other "real" perceptions, hallucinated text is also hard to capture at first glance.

Within the context of NLP, the most inclusive and standard definition of hallucination is *the generated content that is nonsensical or unfaithful to the provided source content* [50, 117, 133, 219]. However, there exist variations in definition across NLG tasks as described in the following section.

2.1 Categorization

(1) **Intrinsic Hallucinations**: a generated output that contradicts the source content. For instance, in the abstractive summarization task from Table 1, the generated summary "*The first Ebola vaccine was approved in 2021*" contradicts the source content "*The first vaccine for Ebola was approved by the FDA in 2019.*".
(2) **Extrinsic Hallucinations**: a generated output that cannot be verified from the source content (i.e., the output that can neither be supported nor contradicted by the source). For example, in the abstractive summarization task from Table 1, "*China has already started clinical trials of the COVID-19 vaccine.*" is not mentioned in source. We can neither find evidence of the generated output from the source nor assert that it is wrong. Notably, the extrinsic hallucination is not always erroneous because it could be from factually correct external information [117, 169]. Such factual hallucination can be helpful because it recalls additional background knowledge to improve the informativeness of the generated text. However, in most of the literature, extrinsic hallucination is still treated with caution because the unverifiable aspect of this additional information increases the risk from a factual safety perspective.

2.2 Task Comparison

The previous subsection is about the definition and categorization of hallucination commonly shared by many NLG tasks. Yet, there are some task-specific differences.

For abstractive summarization, data-to-text, and dialogue tasks, the main difference is in what serves as the "source" and the level of tolerance towards hallucinations. The source in abstractive summarization is the input source text that is being summarized [155]. The source in data-to-text is non-linguistic data [56, 148], and the source(s) in dialogue system are dialogue history or/and the external knowledge sentences. Tolerance towards hallucinations is very low in both summarization [132] and data-to-text tasks [133, 182, 184] because it is essential to provide faithful generation. However, there is relatively higher tolerance in dialogue systems because the desired characteristic is not only faithfulness but also user engagement, especially in open-domain dialogue systems [74].

For GQA task, the exploration of hallucination is at its early stage, so there is no standard definition or categorization of hallucination yet. However, we can see that the GQA literature mainly focuses on "intrinsic hallucination" where the source is the world knowledge [97]. Lastly, unlike the aforementioned tasks, the categorizations of hallucinations in machine translation vary within the task. Most relevant literature agrees that translated text is considered a hallucination when the source text is completely disconnected from the translated target [91, 125, 145]. For further details, please refer to Section 11.

2.3 Terminology Clarification

There are multiple terminologies associated with the concept of hallucination. We provide clarification of the commonly used terminologies (*hallucination*, *faithfulness* and *factuality*) to resolve any confusion. *Faithfulness* is defined as staying consistent and truthful to the provided source – an antonym to "hallucination." Any work that tries to maximize faithfulness focuses on minimizing hallucination. For this reason, our survey includes all those works that address the faithfulness of machine generated outputs. *Factuality* refers to the quality of being actual or based on fact. Depending on what serves as the “fact”, "factuality" and "faithfulness" may or may not be the same. Maynez et al. [117] differentiates "factuality" from "faithfulness" by defining the “fact” to be the world knowledge. In contrast, Dong et al. [33] uses the source input as the “fact” to determine the factual correctness, making "factuality" indistinguishable from "faithfulness". In this paper, we adopt the definition from Maynez et al. [117], because we believe having such distinction between source knowledge and world knowledge provides a more clear understanding.

Note that the judging criteria for what is considered faithful or hallucinated (i.e., the definition of hallucination) can differ across tasks. For more details of these variation definitions, you can find in the later task-specific sections.

Task	Sub-Task	Type	Source	Output
Abstractive Summarization [132]		Intrinsic	The first vaccine for Ebola was approved by the FDA in 2019 in the US, five years after the initial outbreak in 2014. To produce the vaccine, scientists had to sequence the DNA of Ebola, then identify possible vaccines, and finally show successful clinical trials. Scientists say a vaccine for COVID-19 is unlikely to be ready this year, although clinical trials have already started.	The first Ebola vaccine was approved in **2021**.
		Extrinsic		**China** has already started clinical trials of the COVID-19 vaccine.
Dialogue	Task-oriented [102]	Intrinsic	inform (NAME = pickwick hotel, PRICERANGE = moderate)	the hotel named pickwick hotel is in a **high** price range.
		Extrinsic		the pickwick hotel **in san diego** is a moderate price range
	Open-domain [154]	Intrinsic	Dialog History: Speaker 1: What do you think about Murray? Speaker 2: I think Murray is a great player he just needs to stay healthy in order to compete more. Who do you like best?	Speaker 1: I like **Roger Nadal** the most!
		Extrinsic	Knowledge: In tennis, the term Big Four refers to the quartet of men's singles players comprising Roger Federer, Rafael Nadal, Novak Djokovic, and Andy Murray.	Speaker 1: I like Djokovic. He has played **in the top ten singles players of the world**.
Generative QA [128]		Intrinsic	Question: dow jones industrial average please?	Answer: The Dow Jones Industrial Average(DJIA) is an index of **30 major U.S.stock indexes**. The DJIAwas created in 1896 by Charles Dow, a statistician and editor of The Wall Street Journal. The index was first published in the mid-1880s, and **was last updated on May 26,1896.**
		Extrinsic	Question: definition of sadduction	Answer: Sadducees. noun. **The definition of a sadducee is a person who acts in a deceitful or duplicitous manner. An example of a Sadduceee is a politician who acts deceitfully in order to gain political power**.1 a member of a Jewish sect that was active during the Second Temple.
Data2text [191]		Intrinsic	TEAM / CITY / WIN / LOSS / PTS / FG_PCT / BLK Rockets / Houston / 18 / 5 / 108 / 44 / 7 Nuggets / Denver / 10 / 13 / 96 / 38 / 7	The Houston Rockets **(18-4)** defeated the Denver Nuggets (10-13) 108-96 on Saturday.
		Extrinsic		**Houston has won two straight games and six of their last seven.**
Translation [219]		Intrinsic	迈克周四去书店。 (Michael went to the bookstore on Thursday.)	**Jerry didn't go** to the bookstore.
		Extrinsic	迈克周四去书店。 (Michael went to the bookstore on Thursday.)	Michael **happily** went to the bookstore **with his friend.**

Table 1. Examples of each category of Hallucinations for each task. In Data2Text task, H/A: home/away, MIN: minutes, PTS: points, REB: rebounds, AST: assists, BLK: blocks, FG_PCT: field goals percentage.

3 CONTRIBUTORS TO HALLUCINATIONS IN NLG

3.1 Hallucination from Data

The main cause of hallucination from data is source-reference divergence. This divergence happens 1) as an artifact of heuristic data collection or 2) due to the nature of some NLG tasks that inevitably contain such divergence. When a model is trained on such data with source-reference(target) divergence, the model can be encouraged to generate text that is not necessarily grounded and not faithful to the provided source.

Heuristic data collection. When collecting large-scale datasets, some works heuristically select and pair real sentences or tables as the source and target [90, 191]. As a result, the target reference may contain information that cannot be supported by the source [133, 181]. For instance, when constructing WIKIBIO [90], a dataset for generating biography notes based on the infobox of Wikipedia, the authors took the Wikipedia infobox as the source and the first sentence of the Wikipedia page as the target ground-truth reference. However, the first sentence of the Wikipedia article is not necessarily equivalent to the infobox in terms of the information they contain. Indeed, Dhingra et al. [29] points out that 62% of the first sentences in WIKIBIO have additional information not stated in the corresponding infobox. This mismatch between source and target in datasets can make trained models hallucinate.

Another problematic scenario is when duplicates from the dataset are not properly filtered out. It is almost impossible to check hundreds of gigabytes of text corpora manually. Lee et al. [92] shows that duplicated examples from the pretraining corpus bias the model to favor generating repeats of the memorized phrases from the duplicated examples.

Innate divergence. There are NLG tasks that, by nature, do not always have factual knowledge alignment between the source input text and the target reference, especially those that value diversity in generated output. For instance, it is acceptable for open-domain dialogue systems to respond in chit-chat style, subjective style [144], or with the relevant fact but not necessarily present in the user input, history or provided knowledge source – this improves the engagingness and diversity of the dialogue generation. However, researchers have discovered that such dataset characteristic leads to inevitable extrinsic hallucinations.

3.2 Hallucination from Training and Inference

As discussed in the previous subsection, source-reference divergence existing in dataset is one of the contributors of hallucination. However, as Parikh et al. [133] has shown that hallucination problem still occurs even when there is very little divergence in dataset. This is because there is another contributor of hallucinations – training and modeling choices of neural models [82, 145, 150, 178].

Imperfect representation learning. Parikh et al. [133] shows that the comprehension ability of the model could influence the degree of hallucination. The encoder is expected to turn input text in meaningful representations so a model can comprehend the input. When encoders learn wrong correlations between different parts of the training data, it could result in erroneous generation that diverges from the input [2, 48, 98, 172].

Erroneous decoding. The decoder takes the encoded input from the encoder and generates the final target sequence. There are two aspects of decoding that contribute to hallucinations. First, decoders could attend to the wrong part of the encoded input source [172]. This leads the generated output to contain mixed up facts between two similar entities [40, 158]. Second, the design of decoding strategy itself can contribute to hallucinations. Dziri et al. [40] illustrates that decoding

	Category	Task	Works
Automatic Metrics	Statistical	Dialogue	Shuster et al. [158]
		Data2Text	Dhingra et al. [29], Wang et al. [184]
		Translation	Martindale et al. [116]
	Model-based	Abstractive Summarization	Durmus et al. [34], Nan et al. [127], Wang et al. [179] Goodrich et al. [61], Williams et al. [190] Falke et al. [44], Laban et al. [89], Mishra et al. [123] Kryscinski et al. [86], Pagnoni et al. [132], Zhou et al. [219] Gabriel et al. [52], Vasilyev et al. [176]
		Dialogue	Balakrishnan et al. [7], Honovich et al. [72], Li et al. [102] Dziri et al. [41], Gupta et al. [66], Santhanam et al. [154]
		Generative QA*	Durmus et al. [34], Sellam et al. [156], Zhang et al. [213] Wang et al. [179]
		Data2Text	Dušek and Kasner [37], Liu et al. [107], Wiseman et al. [191] Filippova [50], Tian et al. [172]
		Translation	Kong et al. [84], Lee et al. [91], Tu et al. [175] Feng et al. [48], Garg et al. [55], Zhou et al. [219] Parthasarathi et al. [134], Raunak et al. [145]
		Task-Agnostic	Goyal and Durrett [63], Liu et al. [106], Zhou et al. [219]
Mitigation Method	Data-Related	Abstractive Summarization	Cao et al. [20], Nan et al. [127], Zhu et al. [222] Gunel et al. [65]
		Dialogue	Honovich et al. [72], Shen et al. [157], Wu et al. [194] Santhanam et al. [154], Shuster et al. [158]
		Generative QA	Bi et al. [12], Fan et al. [45], Yin et al. [203]
		Data2Text	Nie et al. [130], Parikh et al. [133], Wang [181] Liu et al. [107], Rebuffel et al. [146]
		Translation	Lee et al. [91], Raunak et al. [145] Briakou and Carpuat [16], Junczys-Dowmunt [77]
	Modeling and Inference	Abstractive Summarization	Huang et al. [73], Li et al. [98], Song et al. [160] Aralikatte et al. [2], Cao et al. [18], Cao and Wang [19] Albrecht and Hwa [1], Chen et al. [22], Zhao et al. [216]
		Dialogue	Balakrishnan et al. [7], Li et al. [102], Rashkin et al. [144] Dziri et al. [40]
		Generative QA	Fan et al. [45], Krishna et al. [85], Li et al. [97] Nakano et al. [126]
		Data2Text	Liu et al. [107], Wang et al. [182], Xu et al. [200] Filippova [50], Rebuffel et al. [146], Su et al. [163] Tian et al. [172], Wang et al. [184], Xiao and Wang [195]
		Translation	Feng et al. [48], Lee et al. [91], Weng et al. [189] Li et al. [101], Raunak et al. [145], Wang and Sennrich [180] Bengio et al. [10], Zhou et al. [219] Goyal et al. [62], Xu et al. [199]

Table 2. Evaluation Metrics and Mitigation Method for each task. * The hallucination metrics in the listed related works are not specifically proposed for generative question answering (GQA), but they can be adapted for GQA.

strategy that improves the generation diversity (i.e., top-k sampling) is positively correlated with the increased hallucination. We conjecture that the deliberately added "randomness" by sampling from top-k samples instead of choosing the most probable token has increased the unexpected nature of the generation, leading to a higher chance of containing hallucinated contents.

Exposure Bias. Regardless of decoding strategy choices, the exposure bias problem [10, 142], defined as the discrepancy in decoding between training and inference time, can be another contributor to hallucinations. It is common practice to train the decoder with teacher-forced MLE training, where the decoder is encouraged to predict the next token conditioned on the ground-truth prefix sequences. However, during the inference generation, the model generates the next token conditioned on the history sequences previously generated by itself [69]. Such discrepancy can lead to increasingly erroneous generation, especially when the target sequence gets longer.

Parametric knowledge bias. Pre-training of models on large corpus is known to result in the model memorizing various knowledge in its parameters [113, 136, 149]. This so-called parametric knowledge helps improve the performance of downstream tasks, but also serves as another contributor to hallucinatory generation. Large pre-trained models used for downstream NLG tasks are powerful in providing generalizability and coverage, but Longpre et al. [108] have discovered that such models prioritize parametric knowledge over the provided input. In other words, models that favor generating with their parametric knowledge instead of the information from the input source can result in the hallucination of excess information in the output.

4 METRICS MEASURING HALLUCINATION

Recently, various works in the literature have illustrated that most conventional metrics used to measure the quality of writing are not adequate for quantifying the level of hallucination [147]. It is shown that state-of-the-art abstractive summarization systems (evaluated with metrics such as ROUGE, BLEU, and METEOR) have hallucination content in 25% of their generated summaries [44]. A similar phenomenon has been shown in other NLG tasks, where it is discovered that traditional metrics have a poor correlation with human judgment in terms of the hallucination problem [29, 34, 72, 85]. Therefore, there are active research efforts to define effective metrics for quantifying hallucination.

4.1 Statistical Metric

One of the simplest approaches is to leverage lexical features (n-grams) to calculate the information overlap and contradictions between the generated and the reference texts – the higher the mismatch counts, the lower the faithfulness and thus the higher hallucination score.

Given that many traditional metrics leverage the target text as the ground-truth reference (e.g., ROUGE, BLEU, etc.), Dhingra et al. [29] build upon this idea and proposes PARENT (Precision And Recall of Entailed Ngrams from the Table) [2] that can also measure hallucinations by using *both* source and target text as references. In detail, PARENT n-gram lexical entailment matching of generated text with both the source table and target text. And the F1-score combining the entailment precision and recall reflects the accuracy in the table-to-text task. Source text is additionally used because it is not guaranteed that the output target text contains the complete set of information available in the input source text.

It is common for NLG tasks to have multiple plausible outputs from the same input, known as one-to-many mapping [64, 162]. In practice, however, covering all the possible outputs is too expensive and almost impossible. Thus, many works simplify the hallucination evaluation setup

[2]Note that PARENT is a general metric like ROUGE and BLEU, not only constrained to hallucination

by only relying on the source text as the sole reference. These metrics just focus on the information referred by input sources to measure hallucinations, especially intrinsic hallucinations. For instance, Wang et al. [184] proposes PARENT-T that simplifies PARENT by only using table content as the reference. Similarly, Knowledge F1 [158] – a variant of unigram F1 – is proposed for knowledge grounded dialogue tasks to measure the overlap between the model's generation and the knowledge used to ground the dialogue during dataset collection.

Furthermore, Martindale et al. [116] proposed a bag-of-vectors sentence similarity (BVSS) metric for measuring sentence adequacy in machine translation, that only refers to the target text. This statistical metric helps to determine whether the MT output has different amount of information than the translation reference.

Although simple and effective, one potential limitation of the lexical matching is that it can only handle the lexical information. Thus, it fails to deal with syntactic or semantic variations [156].

4.2 Model-based Metric

Model-based metrics leverage neural models to measure the hallucination degree in the generated text. They are proposed to handle more complex syntactic and even semantic variations. The model-based metrics comprehend the source and generated texts and detect the knowledge/content mismatches. However, the neural models can be subject to errors that can propagate and adversely affect the accurate quantification of hallucination.

4.2.1 Information Extraction (IE)-based. It is not always easy to determine which part of the generated text contains the knowledge that requires verification. IE-based metrics use IE models to represent the knowledge in a simpler relational tuple format (e.g., *subject, relation, object*), then verify against relation tuples extracted from the source/reference. Here, the IE model is identifying and extracting the "facts" that require verification. In this way, words containing no verifiable information (e.g., stopwords, conjunctions, etc) are not included in the verification step.

For example, ground-truth reference text "`Brad Pitt was born in 1963`" and generated text "`Brad Pitt was born in 1961`" will be mapped to the relation triples (`Brad Pitt, born-in, 1963`) and (`Brad Pitt, born-in, 1961`) respectively [3]. The mismatch between the dates (1963≠1961) indicates that there is hallucination. One limitation associated with this approach is the potential error propagation from the IE model.

4.2.2 QA-based. This approach implicitly measures the knowledge overlap or consistency between the generation and the source reference. This is based on the intuition that similar answers will be generated from a same question if the generation is factually consistent with the source reference. It is already put in use to evaluate hallucinations in many tasks, such as summarization [34, 179], dialogue system [72].

QA-based metric that measures the faithfulness of the generated text is consisted of three parts: First, given a generated text, a question generation (QG) model generates a set of question-answer pairs. Second, a question answering (QA) model answers the generated questions given a ground-truth source text as the reference (containing knowledge). Lastly, the hallucination score is computed based on the similarity of the corresponding answers.

Similar to the IE-based metrics, the limitation of this approach is the potential error that might arise and propagated from either the QG model or the QA model.

4.2.3 Natural Language Inference (NLI) Metrics. There are not many labelled datasets for hallucination detection tasks, especially at the early stage when the hallucination problem starts to gain attention. As an alternative, many works leverage the NLI dataset to tackle hallucinations.

[3]This is an example from [61]

Note that NLI is a task that determines whether a "hypothesis" is true (entailment), false (contradiction), or undetermined (neutral) given a "premise". These metrics are based on the idea that only the source knowledge reference should entail the entirety of the information in faithful and hallucination-free generation [37, 41, 44, 72, 75, 86, 89, 123, 190]. More specifically, NLI-based metrics define the hallucination/faithfulness score to be the entailment probability between the source and its generated text, also known as the percentage of times generated text entails, neutral to, and contradicts the source.

According to Honovich et al. [72], NLI-based approaches are more robust to lexical variability than token matching approaches such as IE-based and QA-based metrics. Nevertheless, as illustrated by Falke et al. [44], off-the-shelf NLI models tend to transfer poorly to the abstractive summarization task. Thus, there is a line of research in improving and extending the NLI paradigm specifically for hallucination evaluation purposes [41, 44]. Apart from generalizability, Goyal and Durrett [63] point out the potential limitation of using sentence-level entailment models, namely their incapability to pinpoint and locate which parts of the generation are erroneous. In response, the authors propose a new dependency-level entailment and attempt to identify factual inconsistencies in a more fine-grained manner.

4.2.4 Faithfulness Classification Metrics. To improve upon NLI-based metrics, task-specific datasets are constructed to improve from the NLI-based metrics. Zhou et al. [219] constructed syntactic data by automatically inserting hallucinations into summaries, and Santhanam et al. [154] and Honovich et al. [72] constructed corpus by adapting from Wizard-of-Wikipedia dataset [31] for faithfulness classification of dialogue responses. Faithfulness specific datasets can be better than NLI datasets because entailment or neutral labels of NLI datasets and faithfulness are not equivalent. For example, the hypothesis "Putin is U.S. president" can be considered to be either neutral to or entailed from the premise "Putin is president". However, from the faithfulness perspective, the hypothesis contains unsupported information "U.S.", which is deemed to be hallucination.

4.2.5 LM-based Metrics. These metrics leverage two language models (LMs) to determine if each token is supported or not: An unconditional LM is only trained on the targets (ground-truth references) in the dataset, while a conditional language model LM_x is trained on both source and target data. It is assumed that the next token is inconsistent with the input if unconditional LM gets a smaller loss than conditional LM_x during forced-path decoding [50, 172]. We classify the generated token as hallucinatory if the loss from LM is lower. The ratio of hallucinated tokens to the total number of target tokens $|y|$ can reflect the hallucination degree.

4.3 Human Evaluation

Due to the challenging and imperfect nature of the current automatic evaluation of hallucinations in NLG, human evaluation [154, 158] is still one of the most commonly used approaches. There are two main forms of human evaluation: (1) scoring, where human annotators rate the hallucination level in a range; and (2) comparing, where human annotators compare the output texts with baselines or ground-truth references [165].

Multiple terminologies, such as *faithfulness* [20, 22, 50, 117, 133, 144, 144, 163, 172, 195, 219], *factual consistency* [18, 19, 24, 154, 157, 194], *fidelity* [23], *factualness*[4] [146], *factuality*[4] [33],

or on the other hand, *hallucination* [40, 73, 107, 154, 158], *fact contradicting* [129] are used in the human evaluation of hallucination to rate whether the generated text is in accord with the source input. Chen et al. [22], Nie et al. [130] use finer-grained metrics for *intrinsic hallucination* and *extrinsic hallucination* separately. Moreover, there are some broad metrics, such as *Correctness* [7,

[4]uses the source input as the "fact".

12, 98, 182], *Accuracy* [97, 203], and *Informativeness* [102] considering both missing and additional contents (extrinsic hallucinations) compared to the input source.

5 HALLUCINATION MITIGATION METHODS

Common mitigation methods can be divided into two categories, in response to two main contributors of hallucinations: **Data-Related Method**, and **Modeling and Inference Method**.

5.1 Data-Related Method

5.1.1 Build Faithful Dataset. Considering noisy data encourage hallucinations, constructing faithful datasets manually is intuitive, and there are various ways to build: One way is employing annotators to write clean and faithful targets from scratch given the source [54, 188], which may tend to lack diversity [67, 133, 137]. Another way is employing annotators to rewrite real sentences on the web [133], or the targets in the existing dataset [181]. Basically, the revision strategy consists of three stages: (1) Phrase Trimming: remove phrases unsupported by source in the exemplar sentence; (2) Decontextualization: resolve co-reference and delete phrases dependent on context; (3) Syntax Modification: make the purified sentences flow smoothly. There are also some works [52, 72] leveraging the model to generate data and instruct annotators to label whether these outputs contain hallucinations or not. While this approach is typically used to build diagnostic evaluation datasets, it has the potential to build faithful datasets.

5.1.2 Clean Data Automatically. In order to alleviate semantic noise matters, another approach is to find out the irrelevant or contradictory information to the input from the existing parallel corpus and then filter or correct the data. This approach is suitable for the case where there is a low or moderate level of noise in the original data [50, 130].

Some works [107, 145, 157] deal with the hallucination issue at the instance level by using a score for each source-reference pair and filtering out hallucinating ones. This corpus filtering method consists of several steps: (1) quality measure training samples regarding hallucination which could utilize the metrics described above; (2) rank these hallucination scores in descending order; (3) select and filter out the untrustworthy samples at the bottom. Instance-level scores can lead to a signal loss because divergences occur at the word level, i.e., parts of the target sentence loyal to the source input, while others diverge [146].

Considering this issue, other works [35, 130] correct paired training samples, specifically the input data, according to the references. This method is mainly applied in the Data-to-Text task because structured data is easier to be corrected than utterances. This method consists of two steps: (1) utilize a model to parse the Meaning Representation (MR) such as attribute-value pairs from original human textual references; (2) use the extracted MR from the reference to correct the input MR through slot matching. This method will enhance the semantic consistency between input and output without abandoning a part of the dataset.

5.1.3 Information Augmentation. It is intuitive that augmenting the inputs with external information will obtain a better representation of the source. Because the external knowledge, explicit alignment, extra training data, etc., can improve the correlation between source and target and help the model learn better task-related features. Consequently, a better semantic understanding helps alleviate the divergence issue from the source. For example, augmented with entity information [107], extracted relation triples from source document [20, 73] obtained by Fact Description Extraction, synthetic data generated through replacement or perturbation [22, 91], retrieved external knowledge [12, 45, 65, 158, 222], and retrieved similar training samples [13].

These methods enforce a stronger alignment between inputs and outputs. However, they will bring challenges due to the gap between the original source and augmented information, such as the semantic gap between the ambiguous utterance and the distinct meaning representation of structured data, and the format discrepancy between the structured knowledge graph and natural language.

5.2 Modeling and Inference Method

5.2.1 Architecture.

Encoder. The encoder learns to encode a variable-length sequence from input text into a fixed-length vector representation. As we mentioned above in (Section 5.1.3), learning a better representation is helpful for reducing hallucination. Some work have modified the encoder architecture in order to be more compatible with input. For example, Cao et al. [20], Huang et al. [73] proposes a dual encoder, consisting of a sequential document encoder and a structured graph encoder to deal with the additional knowledge.

Attention. Attention Mechanism is an integral implement selectively concentrating on relevant parts while ignoring others based on dependencies in neural networks [4, 177]. In order to encourage the generator to pay more attention to the source, Aralikatte et al. [2] introduce a short circuit from the input document to the vocabulary distribution via source-conditioned bias. Krishna et al. [85] employ sparse attention to improve the model's long-range dependencies in the hope of modeling more retrieved documents to mitigate the hallucination in the answer. Wu et al. [194] adopt inductive attention, which removes potentially uninformative attention links by injecting pre-established structural information to avoid hallucinations.

Decoder. The decoder converts vector representations into natural language [177], and this stage could contribute to hallucinations due to the limitation of existing decoding strategies. There are also some work modifying the decoder structure to mitigate hallucination, such as multi-branch decoder [146], uncertainty-aware decoder [195], dual decoder consisting of a sequential decoder and a tree-based decoder [160], and constrained decoder with lexical or structural limitations [7]. These decoders improve the possibility of faithful tokens while reducing the possibility of hallucinatory tokens during inference by figuring out the implicit discrepancy and dependency between tokens or limited by explicit constraints.

5.2.2 Training.

Planning/Skeleton. Planning is a common method to control and restrict what the model generates by informing the content and its order [163]. Planning can be a separate step in a two-step generator [22, 107, 163, 182] or be injected into the end-to-end model during generation [200]. Skeleton has a similar function to planning, and it can also be adopted into handling hallucinations [182]. And the difference is that the skeleton is treated as a part of the final generated text.

Reinforcement Learning. As pointed out by Ranzato et al. [142], word-level maximum likelihood training leads to the problem of exposure bias. Some research [3, 73, 84, 102, 120, 135, 163] adopt reinforcement learning to solve the hallucination problem, which utilizes different rewards to optimize the model. The reward is crucial and bottleneck of reinforcement learning and the approach to calculate reward score is related to exploring automatic metrics to evaluate the generated results.

Multi-task Learning. Multi-task Learning is also utilized for handling hallucinations in different NLG tasks. For example, FENMT [189] and Garg et al. [55] incorporate a word alignment task into the translation model. Li et al. [98] incorporate the entailment task into abstractive summarization models. Li et al. [97] incorporate the rationale extraction task and the answer generation task. The

Multi-task approach has several advantages, such as data efficiency improvement, overfitting reduction and fast learning. It is crucial to choose which tasks should be learned jointly, and learning multiple tasks simultaneously presents new challenges of design and optimization [26].

Controllable Generation. Current works treat hallucination level as a controllable attribute in order to control the hallucination in outputs at a low level. Controllable generation techniques like control code which could be provided manually [50, 144, 194] or predicted automatically [194], and controlled re-sampling [144] are leveraged to improve faithfulness. Considering that hallucination is not necessarily harmful and may bring some benefits, this controllable method can be further adopted for changing the hallucination degree to meet the demands of different real-world applications.

Other general training methods such as regularization [79, 91, 125] and loss reconstruction [101, 180, 184] are also proposed to tackle the hallucination problem.

5.2.3 Post-Processing. Post-processing methods can correct hallucinations in the output, and this standalone task requires less training data. Especially when it comes to the noisy dataset where a large proportion of ground truth references suffer from hallucinations, modeling correction is a competitive choice to handle the hallucination problem [22]. Cao et al. [18], Chen et al. [22], Dong et al. [33], Dziri et al. [40] follow a generate-then-refine strategy. The post-processing correction step allows researchers to utilise SOTA models which perform best in the aspect of other attributes such as fluency and then correct the results specifically in aspect to faithfulness by using small amounts of automatically generated training data.

6 FUTURE DIRECTIONS

Many studies have been conducted to tackle the hallucination problem in Natural Language Generation and its downstream tasks. As mentioned above, we have discussed common metrics and mitigation methods to advance research in these fields. From a broader perspective, we wish to point out open challenges and potential directions divided into **Metric** and **Mitigation Method**.

6.1 Future Directions in Metrics Design

Fine-grained Metrics. Most of the existing hallucination metrics measure intrinsic and extrinsic hallucinations together as a unified metric. However, it is common for a single generation to have multiple types and a number of hallucinatory sub-strings. Fine-grained metrics that can distinguish between two types of hallucinations will provide richer insight to the researchers.

In order to implement such a metric, the first step would be to identify the exact location of the hallucinatory sub-strings correctly. However, some metrics such as those that are QA-based cannot identify the individual hallucinatory sub-strings. Improvements in this aspect would help improve the quality and the explainability of the metrics. The next step would be to categorize the detected hallucinatory sub-string. In theory, the hallucinatory sub-string will be intrinsic if it is wrong or nonsensical, and extrinsic if it is non-existing in the source context. Future work that explores an automatic method of doing this would be beneficial.

Fact-Checking. The factual verification of extrinsic hallucinations requires fact-checking against world knowledge, which can be time consuming and laborious. Leveraging an automatic fact-checking system for extrinsic hallucination verification is, thus, another future work that requires attention. Fact-checking consists of knowledge evidence selection and claims verification sub-tasks, and the following are the remaining challenges associated with each sub-task.

The main research problem associated with the evidence selection sub-task is "how to retrieve evidence from the *world* knowledge?" Most of the literature leverage Wikipedia as the knowledge source [93, 171, 204], which is only a small part of the world knowledge. Some literature attempt to use the whole web as the knowledge source [43, 115]. However, this method leads to another research problem – "how to ensure the trustworthiness of the information we use from the web?" [58] Source-level methods that leverages the meta-information of the web source (i.e., web traffic, PageRank, URL structure) are proposed to deal with such trustworthiness issue [8, 138, 139]. It will be an important future direction to address the aforementioned issues to allow the evidence selection against world knowledge.

For the verification subtask, the verification model performs relatively well if given correct evidence [94]. However, it is shown that verification models are prone to adversarial attacks and are not robust to negation, numerical or comparative words [170]. Trying to improve this weakness of the verification model would also be crucial because the factuality of a sentence can easily be changed by small word changes (i.e., changes in negations, numbers, and entities).

Generalization. Although we can see that the source and output text of different tasks are in various forms, investigating their relationship and common ground and proposing general metrics to evaluate hallucinations are worth exploring. Task-agnostic metrics with cross-domain robustness could help the research community to build a unified benchmark. It is also important and meaningful to build open-source platforms where collaborate and standardize the evaluation metrics for NLG tasks.

Incorporation of Human Cognitive Perspective. A good automatic metric should correlate with human evaluation. Humans are sensitive to different types of information. For instance, proper nouns are usually more important than pronouns in the generated text. Mistakes concerning named entities are striking to human users, but automatic metrics treat them equally if not properly designed. In order to address this issue, new metrics should be designed from the human cognitive perspective. The human ability to recognize salient information and filter the rest is evident in scenarios where the most important facts need to be determined and assessed. For instance, when signing an agreement, a prospective employee naturally skims the document to look at the entries with numbers first. This way, one classifies what one believes is crucial.

Therefore, automatic check-worthy detection has the potential to be applied to improve the correlation with human judgement. Implementing the above mentioned, automatic human-like judgement can further mitigate hallucination and improve natural language generation systems.

6.2 Future Directions in Mitigation Methods

General and robust data pre-processing approaches. Since the data format varies between different downstream tasks, there is still a gap for data processing methods between tasks. Therefore, currently, no universal method is effective for all NLG tasks [96]. Data pre-processing might result in grammatical errors or semantic transformation between the original and processed data, which can negatively affect the performance of generation. Therefore, we believe that general and robust data pre-processing methods can help mitigate the hallucinations in NLG.

Hallucinations in numerals. Most of the existing mitigation methods do not focus on the hallucinations of numerals. However, the correctness of the numerals in the generated texts such as date, quantity and scalar are important for readers [168, 212, 216]. For example, given the source document "*The optimal oxygen saturation (SpO_2) in adults with COVID-19 who are receiving supplemental oxygen is unknown. However, a target SpO_2 of 92% to 96% seems logical, considering that indirect* *evidence from patients without COVID-19 suggests that an SpO_2 of <92% or >96% may be harmful.* [5]", the summary "*The target oxygen saturation range for patients with COVID-19 is 82–86%.*" includes wrong numbers which can be fatal. Currently, some works [130, 168, 212] point out that using commonsense knowledge can help to gain better numeral representation. And Zhao et al. [216] alleviate numeral hallucinations by re-ranking candidate generated summaries based on the verification score of quantity entities. Therefore, we believe explicitly modeling numerals to mitigate hallucinations is a potential direction.

Extrinsic Hallucination Mitigation. Though many works have been done to mitigate hallucinations, most of them do not distinguish between intrinsic and extrinsic hallucination. Moreover, the main research focus has been on dealing with intrinsic hallucination, while extrinsic hallucination is somewhat overlooked as it is more challenging to reduce [75]. Therefore, we believe it is worth exploring different mitigation methods for intrinsic and extrinsic hallucinations. In addition, relevant methods in fact-checking can be potentially used for this purpose.

Hallucination in long text. Many tasks in NLG require the model to process long input texts, such as multi-document summarization and generative question answering. We think adopting existing approaches to Longformer [9] based model can help encode long inputs. Meanwhile, part of dialogue systems need to generate long output text, in which the latter part of the generated text may contradict history generation. Therefore, reducing self-contradiction is also an important future direction.

Reasoning. Misunderstanding facts in the source context will lead to intrinsic hallucination and errors. To help models understand the facts correctly requires reasoning over the input table or text. Moreover, if the generated text can be reasoned backwards to the source, we can assume it is faithful. There are some reasoning works in dialogue [27, 57, 183], but few reasoning works reducing hallucinations. Moreover, tasks with quantities, such as logical table-to-text generation, require numerical reasoning. Therefore, adding reasoning ability to the hallucination mitigation methods is also an interesting future direction.

Controllability. Controllability means the ability for models to control the level of hallucination and strike a balance between faithfulness and diversity [40, 150]. As mentioned in Section 3, it is acceptable for chit-chat models to generate a certain level of hallucinatory contents as long as they are factual. Meanwhile, for abstractive summarization task, there is no agreement in the research community about whether the factual hallucinations are desirable or not [117]. Therefore, we believe controllability merits attention when exploring hallucination mitigation methods.

7 HALLUCINATION IN ABSTRACTIVE SUMMARIZATION

Abstractive summarization aims to extract essential information from source documents and to generate short, concise, and readable summaries [205]. Neural networks have achieved remarkable results for abstractive summarization. However, Maynez et al. [117] observe that neural abstractive summarization models are likely to generate hallucinatory content that is unfaithful to the source document. Falke et al. [44] analyze three recent abstractive summarization systems and show that 25% of the summaries generated from state-of-the-art models have hallucination content. In addition, Zhou et al. [219] mention that even if a summary contains a large amount of hallucinatory content, it can achieve a high score in ROUGE [103]. This has encouraged researchers to actively devise ways to improve the evaluation metric of abstractive summarization, especially from the hallucination perspective.

[5]https://www.covid19treatmentguidelines.nih.gov/management/critical-care/oxygenation-and-ventilation/

	Category	Description	Example
PredE	Predicate Error	The predicate in the summary statement is inconsistent with the source article.	The Ebola vaccine was rejected by the FDA in 2019.
EntE	Entity Error	The primary arguments (or their attributes) of the predicate are wrong.	The COVID-19 vaccine was approved by the FDA in 2019.
CircE	Circumstance Error	The additional information (like loca- tion or time) specifying the circumstance around a predicate is wrong.	The first vaccine for Ebola was approved by the FDA in 2014.
CorefE	Coreference Error	A pronoun/reference with wrong or non- existing antecedent.	The first vaccine for Ebola was approved in 2019. They say a vaccine for COVID-19 is unlikely to be ready this year.
LinkE	Discourse Link Error	Error in how multiple statements are linked together in the discourse (for example temporal ordering/causal link).	To produce the vaccine, scientists have to show successful human trials, then sequence the DNA of the virus.
OutE	Out of Article Error	The statement contains information not present in the source article.	China has already started clinical trials of the COVID-19 vaccine.
GramE	Grammatical Error	The grammar of the sentence is so wrong that it becomes meaningless.	The Ebola vaccine accepted have already started.

Table 3. Types of factual errors. Original text for the examples: "*The first vaccine for Ebola was approved by the FDA in 2019 in the US, five years after the initial outbreak in 2014. To produce the vaccine, scientists had to sequence the DNA of Ebola, then identify possible vaccines, and finally show successful clinical trials. Scientists say a vaccine for COVID-19 is unlikely to be ready this year, although clinical trials have already started.*" This table is taken from [132].

In this section, we overview the current progress in the automatic evaluation and the mitigation of hallucination, and list out the remaining challenges for future work. In addition, it is worth mentioning that researchers have used different terms to describe the hallucination phenomenon, such as faithfulness, factual errors, and factual consistency. Therefore, we will use the original terms from their paper in the remainder of this section.

7.1 Hallucination Definition in Abstractive Summarization

The definition of hallucination in abstractive summarization follows Section 2. Specifically, we adopt the definition from [117]: given a document and its abstractive summary, a summary is hallucinated if it has any spans not supported by the input document. Once again, intrinsic hallucination refers to the output content that contradicts the source, while extrinsic hallucination refers to the output content that the source cannot verify. For instance, given the input article as shown in the caption of Table 3. An example of intrinsic hallucination is "*The Ebola vaccine was rejected by the FDA in 2019.*" Because this statement contradicts the given content "*The first vaccine for Ebola was approved by the FDA in 2019 in the US*". While an example of extrinsic hallucination is "*China has already started clinical trials of the COVID-19 vaccine.*" Because this statement is not mentioned in the given content. We can neither find evidence of it from the input article nor assert that it is wrong.

Moreover, Pagnoni et al. [132] devise fine-grained types of factual errors in summaries. As mentioned in 2.3, since the "fact" here refers to source knowledge, "factual error" can be treated as hallucination and we can adopt this classification as a sub-type of hallucination. They establish three categories as semantic frame error, discourse error, and content verifiability error. (1) Semantic frame error refers to errors in the schematic representation of an event, relation, or state, which consists of a predicate and a list of participants, called frame elements. They establish three sub-categories for semantic frame errors: Predicate Error, Entity Error and Circumstance Error as shown in Table 3. (2) Discourse error refers to factual errors beyond single semantic frames such as an erroneous link between discourse segments (Coreference Error, Discourse Link Error shown in Table 3). (3) Content verifiability error refers to the factuality that cannot be verified against the source text due to difficulty in alignment with the source text – can be viewed as extrinsic hallucination (Out Of Article Error, Grammatical Error shown in Table 3).

7.2 Hallucination Metrics in Abstractive Summarization

Existing metrics for hallucination in abstractive summarization are mainly model-based. Following [75], we divide the hallucination metrics into two categories: (1) Unsupervised Metrics and (2) Semi-Supervised Metrics. Note that current existing hallucination metrics evaluate both intrinsic and extrinsic hallucinations together as one metric because it is difficult to automatically distinguish between intrinsic and extrinsic.

7.2.1 Unsupervised Metrics. Given that hallucination is a newly emerged problem, there are only a few hallucination-related datasets. Therefore, researchers have proposed to adopt other datasets to build unsupervised hallucination metrics. There are three different types of such unsupervised metrics: (1) Information Extraction (IE)-based Metrics, (2) Natural Language Inferencing (NLI)-based Metrics, (3) Question Answering (QA)-based Metrics.

IE-based Metrics. As mentioned in Section 4, IE-based metric leverages IE models to extract knowledge as relation tuples (*subject, relation, object*) from both the generation and the knowledge source to analyze the factual accuracy of the generation [61]. However, IE models are not 100% reliable yet (making errors in the identification of the relation tuples). Therefore, Nan et al. [127] proposes an entity-based metric relying on the Named-Entity Recognition model that is relatively more robust. Their metric builds on an assumption that there will be a different set of named entities in the gold and the generated summary if there exists hallucination.

NLI-based Metrics. As mentioned in Section 4, the NLI-model (textual entailment model) can be utilized to measure hallucination based on an assumption that faithful summary will be entailed by the gold source. However, Falke et al. [44] discover that models trained on NLI datasets can not transfer well to abstractive summarization tasks, degrading the reliability of NLI-based hallucination metrics. To improve NLI models for hallucination evaluation, they release collected annotations as additional test data in future work. Other efforts are also made to further improve NLI models. Mishra et al. [123] find that the low performance of NLI-based metrics is mainly caused by the length of premises in NLI datasets is short than the source documents in abstractive summarization. Thus, the authors propose to convert multiple-choice reading comprehension datasets into long premise NLI datasets automatically. The results indicate that long premise NLI dataset helps the model achieve a higher performance than the original NLI datasets. In addition, Laban et al. [89] introduce a simple but efficient method called SUMMAC_{Conv} by applying NLI models to the sentence units that are segmented from documents. The performance of their model is better than applying NLI models to the whole document.

QA-based Metrics. As mentioned in Section 4, the QA-based metrics measure the knowledge overlap or consistency between the summaries and the source documents. Based on intuition that the QA models will achieve similar answers if the summaries are factual consistent with the source documents. Durmus et al. [34] and Wang et al. [179] propose FEQA and QAGS respectively. Both of them follow three steps to get the final score: (1) Using Question Generation (QG) model to generate the questions from the reference summaries. (2) Using QA model to get answers from the source documents and generated summaries. (3) Calculate the scores by comparing two sets of the answers. The results show both QAGS and FEQA have substantially higher correlations with human judgments of faithfulness than the baseline metrics. In addition, Gabriel et al. [52] further analyze the FEQA and find that the effectiveness of QA-based metrics depends on the question. They also provide a meta-evaluation framework that includes QA metrics.

7.2.2 Semi-Supervised Metrics. Semi-supervised metrics are trained on the synthetic data generated from summarization datasets. Trained on these task-specific corpora, models can judge whether the generated summaries are hallucinatory. Kryscinski et al. [86] propose a weakly-supervised model named FactCC for evaluating factual consistency. The model is trained jointly for three tasks: (1) check whether the synthetic sentences remain factually consistent (2) extract supporting spans in the source documents (3) extract inconsistent spans in the summaries, if any exists. They transfer this model to check whether the summaries generated from summarization models are factual consistency. Results show that the performance of their FactCC model surpasses the classifiers trained on the MNLI or FEVER datasets. Zhou et al. [219] introduce a method to fine-tune a pre-trained language model on synthetic data with automatically inserted hallucinations in order to detect the hallucinatory content in summaries. The model can classify whether spans in the machine-generated summaries are faithful to the article. This method shows higher correlations with human factual consistency evaluation than baselines.

7.3 Hallucination Mitigation in Abstractive Summarization

Recently, many approaches have been proposed to reduce the hallucinatory phenomenon in abstractive summarization.

7.3.1 Architecture Method. Researchers have made modifications to the architecture design of the seq-to-seq models to reduce hallucinated content in the summaries. We describe various efforts made to improve encoder, decoder, or both encoder and decoder of the seq-to-seq models.

Encoder. Zhu et al. [222] propose to use an explicit graph neural network (GNN) to encode the fact tuples extracted from source documents. In addition to the explicit graph encoder, Huang et al. [73] further design a multiple-choice cloze test reward to encourage the model to better understand entity interactions. Moreover, Gunel et al. [65] try to use external knowledge from Wikipedia to make knowledge embeddings. Results show that knowledge embeddings improve factual consistency.

Decoder. Song et al. [160] present to incorporate a sequential decoder with a tree-based decoder to generate a summary sentence and its syntactic parse synchronously. This work jointly generates a sentence and its syntactic dependency parse to improve faithfulness. Aralikatte et al. [2] introduce Focus Attention Mechanism, which encourages decoders to generate tokens similar or topical to the source documents. The results on BBC extreme summarization task show that models augmented with the Focus Attention Mechanism generate more faithful summaries.

Encoder-decoder. Cao et al. [20] extract fact descriptions from the source text and apply a dual-attention sequence-to-sequence framework to force the summaries to be conditioned on both source documents and the extracted fact descriptions. Li et al. [98] propose an entailment-aware encoder and decoder with multi-task learning which incorporates the entailment knowledge into abstractive summarization models.

7.3.2 Training Method. Aside from architecture modification, some works improved the training approach to reduce hallucination. Cao and Wang [19] introduce a contrastive learning method to train summarization models. The positive training data is reference summaries, while the negative training data is automatically generated hallucinatory summaries. The contrastive learning system is trained to distinguish between the positive and negative training data. In the dialogue summarization field, Tang et al. [167] propose another contrastive fine-tuning strategy named CONFIT that can improve the factual consistency and overall quality of summaries.

7.3.3 Post-Processing Method. Some works carry out post-editing to reduce the hallucination of the model-generated summaries, which are viewed as draft summaries. Dong et al. [33] propose SpanFact, a pair of two factual correction models that use knowledge learned from QA models to correct the spans in the generated summaries. Similar to SpanFact, Cao et al. [18] introduce a post-editing corrector module to identify and correct hallucinatory content in generated summaries. The corrector module is trained on synthetic data which is created by adding a series of heuristic transformations to reference summaries. Zhao et al. [216] present HERMAN, a system that learns to recognize quantities (dates, amounts of money, etc.) in the generated summary and verify their factual consistency with the source text. According to the quantity hallucination score, the system choose the most faithful summary where the source text supports its quantity terms from the candidate generated summaries. Chen et al. [22] introduce a contrast candidate generation and selection system to do post-processing. The contrast candidate generation model replaces the named entities in the generated summaries with ones present in the source documents and the contrast candidate selection model will select the best candidate as the final output summary.

7.4 Future Directions in Abstractive Summarization

Factual Hallucination Evaluation. Factual hallucinations contain information not found in source content, though it is factually correct. In the summarization task, this kind of hallucination could lead to better summaries. However, there is little work focused on evaluating factual hallucination explicitly. Fact-checking approaches can be potentially used in this regard.

Extrinsic Hallucination Mitigation. There has been little research on extrinsic hallucinations as it is more challenging to detect and mitigate content based on world knowledge. We believe it is worth exploring extrinsic hallucination in evaluation metrics and mitigation methods.

Hallucination in Dialogue Summarization. In conversational data, the discourse relations between utterances and co-reference between speakers are more complicated than from say, news articles. For example, Zhong et al. [217] show that 74% of the samples in the QMSum dataset consist of inconsistent facts. We believe exploring the hallucination issue in dialogue summarization is an important and special component in the research of hallucination in abstractive summarization.

8 HALLUCINATION IN DIALOGUE GENERATION

Dialogue generation is a NLG task that automatically generates responses according to user utterances. The generated responses are required to be fluent, coherent, and consistent with the dialogue history. The dialogue generation task can be divided into two tasks: (1) Task-oriented Dialogue Generation; (2) Open-domain Dialogue Generation. A task-oriented dialogue system aims to complete a certain task according to a user query in a specific domain, such as restaurant booking, hotel recommendation, and calendar checking. Meanwhile, an open-domain dialogue system aims to establish a multi-turn, long-term conversation with users while providing the users with an engaging experience.

8.1 Hallucination Definition in Dialogue Generation

The hallucination problem also exists in the dialogue generation task. It is important to note that a dialogue system is expected either to provide the user with the required information or to provide an engaging response without repeating utterances from the dialogue history. Thus, the tolerance for producing proper "hallucination" from the dialogue history is relatively higher. The definition of hallucination in this task can be adopted from the general definition as follow: (1) **Intrinsic Hallucination**: the generated response is contradictory to the dialogue history or the external knowledge sentences; (2) **Extrinsic Hallucination**: the generated response is hard to be verified with the dialogue history or the external knowledge sentences.

As the examples of intrinsic hallucination shown in Table 1, we can verify that the output contradicts their inputs: While input is a "*moderate*" price range, the model mistakenly generates a sentence with a "*high*" price range. The confusion of the names of "*Roger Federer*" and "*Rafael Nadal*" causes the output generation with "*Roger Nadal*". On the other hand, responses with extrinsic hallucination are impossible to verify with given inputs. In other words, "*pickwick hotel*" might be "*in san diego*", and Djokovic may have played "*in the top ten singles players of the world*", however, we do not have enough information to check. In the following sections, the hallucination problem in open-domain and task-oriented dialogue generation tasks will be separately discussed according to the their natures.

8.2 Open-domain Dialogue Generation

While the term "hallucination" seems to have newly emerged in the field, a related behavior of neural models has been widely discussed. The behavior commonly known as "inconsistency" has been pointed out as a shortcoming of generation-based approaches for open-domain chatbot [74, 109, 151]. Two possible types of inconsistency occur in open-domain dialogue generation: (1) inconsistency among the system utterances such as when the system contradicts its previous utterance; (2) inconsistency with some external source, such as factually incorrect utterances. Whereas the first type is described using the term "consistency" [100, 186, 208] or "coherence" [11, 39], people recently start to call the second type "hallucination" [122, 152]. Self-inconsistency can be considered as an intrinsic hallucination problem, while the external inconsistency involves both intrinsic and extrinsic hallucinations, depending on the reference source.

As mentioned earlier, a certain level of hallucination may be acceptable in open-domain chit-chat as long as it does not involve severe factual issues. Moreoever, it is almost impossible to verify factual correctness since the system usually lacks the connection to external resources. With the introduction of Knowledge Grounded Dialogue tasks [31, 220] which provides an external reference, however, there has been more active discussion of hallucination in open-domain dialogue generation.

8.2.1 Self-Consistency. In end-to-end generative open-domain dialogue systems, the inconsistency among system utterances has been pointed out as the bottleneck to human-level performance [178]. We often observe an inconsistency in the answers to semantically similar yet not identical questions. For example, a system may answer with different names to the question of "What is your name?" and "May I ask your name?". As one of the most obvious cases of self-contradiction regarding the character of the dialogue system, persona consistency has been the center of attention [99, 211]. "Persona" is defined as the character that a dialogue system plays during a conversation, which can be composed of identity, language behavior, and interaction style [99]. While some works set their objective to teach models to utilize speaker-level embeddings [99, 112], the others condition generation with a set of descriptions about a persona, which we will discuss in detail in the next section.

8.2.2 External Consistency. Besides self-consistency, an open-domain dialogue system is also supposed to generate persona-consistent and informative responses corresponding to user utterances to further engage with the user during the conversation. In this process, an external resource containing explicit persona information or world knowledge is introduced into the system to assist the model generation process.

PersonaChat datasets [30, 211] accelerate the research on persona consistency [68, 88, 118, 192, 202, 207, 214]. In the PersonaChat dataset, each conversation is attached with persona descriptions such as “I like to ski” or “I am a high school teacher”. By conditioning the response generation on the persona description, a chit-chat model is expected to require an ability to generate a more persona-consistent response. Lately, the application of NLI methods [100, 159] or reinforcement learning frameworks [120] have been investigated. Although these conditioning methods using PersonaChat datasets are successful, further investigation of approaches that do not rely on the given set of persona descriptions is necessary because the former is not always available, and covering every aspect of persona with them is impossible.

In addition to Persona-chat related research, Knowledge Grounded Dialogue (KGD) task in the open-domain requires the model to generate informative responses with the help of an external knowledge graph (KG) or knowledge corpus [31, 220]. Hallucination in conversations, which is also considered as a factual consistency problem, has raised much research interest recently [40, 144, 154, 158]. Here, we continue to classify the hallucination problem in the KGD task into intrinsic hallucination and extrinsic hallucination. Most of the KGD works tackle the hallucination problem when responses contain information that contradicts (intrinsic) or cannot be found in the provided knowledge input (extrinsic). Since world knowledge is enormous and ever-changing, the extrinsic hallucination may be factual but hard to be verified. Dziri et al. [40] further adopt the same definition to the Knowledge Graph-grounded Dialogue task, where intrinsic hallucination indicates the case of misusing either subject or object of the knowledge triple; and extrinsic hallucination indicates that there is no corresponding valid knowledge triple in the gold reference knowledge. Recently, there have been some attempts to generate informative responses without explicit knowledge inputs, but with the help of the implicit knowledge inside large pre-trained language models instead [201, 221] during the inference time. Under this setting, the study of extrinsic hallucination is of great value but still poorly investigated.

8.2.3 Hallucination Metrics. For generation-based dialogue systems, especially open-domain chatbots, the hallucination evaluation method remains an open problem [151]. As of now, there is no standard metric. Therefore, chatbots are usually evaluated by humans on factual consistency or factual correctness [154, 194]. We will also introduce some automatic statistical and model-based metrics as a reference, which will be described in more detail below.

Variants of F1 Metrics. Knowledge F1 (KF1) measures the overlap between the generated responses and the gold knowledge sentences on which the human grounded during dataset collection [158]. KF1 attempts to capture whether a model can generate knowledgable responses by correctly utilizing the relevant knowledge. KF1 is only available for datasets with labeled ground-truth knowledge. Shuster et al. [158] further propose *Rare F1 (RF1)*, where only considers the words that are infrequent in the dataset when calculating F1 to avoid influence from the common unigrams. The authors define an infrequent word if it is in the lower half of the cumulative frequency distribution of the reference corpus.

Model-based Metric. Recently, several works have proposed evaluation metrics for measuring consistency, such as using natural language inference (NLI) [39, 186], training learnable evaluation metrics [208], or releasing additional test set for coherence [11]. For the KGD task, Dziri et al. [41] propose the BEGIN benchmark, which consists of samples taken from Dinan et al. [31] with additional human annotation and a new classification task extending the Natural Language Inference (NLI) paradigm. Honovich et al. [72] present a trainable metric for the KGD task, which also applies NLI. It is also noteworthy that Gupta et al. [66] propose datasets that can benefit fact-checking systems specialized for dialogue systems. Conv-FEVER corpus [154] is a factual consistency detection dataset, which is created by adapting from Wizard-of-Wikipedia dataset [31]. It consists of both factually consistent and inconsistent responses and can be used to train a classifier to detect factually inconsistent responses with respect to the knowledge provided.

8.2.4 Mitigation Methods. The hallucination issue can be mitigated by data pre-processing, which includes introducing extra information into the data. Shen et al. [157] propose a measurement based on seven attributes of the dialogue quality, including self-consistency. Based on this measurement, the untrustworthy samples which get lower scores are filtered out from the training set to improve the model performance in terms of self-consistency (i.e. intrinsic hallucination). Shuster et al. [158] conduct a comprehensive investigation on retrieval-augmented KGD task where a retriever is introduced to the system for knowledge selection. The authors study several key problems, such as whether retrieval helps reduce hallucinations how the generation should be augmented with the retrieved knowledge. The experimental results show that retrieval helps substantially in improving performance on KGD tasks and reducing the hallucination in the conversations without sacrificing conversational ability.

Rashkin et al. [144] introduce a set of control codes and concatenate them with dialogue inputs to reduce the hallucination by forcing the model to be more aware of how the response relies on the knowledge evidence in the response generation. Some researchers also try to reduce hallucinated responses during generation by improving dialogue modeling. Wu et al. [194] apply inductive attention into the transformer-based dialogue models. Potentially uninformative attention links are removed with respect to a piece of pre-established structural information between dialogue context and the provided knowledge. Instead of improving the dialogue response generation model itself, Dziri et al. [40] present a response refinement strategy with token-level hallucination critic and entity mention retriever, so that the original dialogue model is left without retraining. The former module is designed to label the hallucinated entity mentioned in the generated responses, while the retriever is trained to retrieve more faithful entities from the provided knowledge graph.

8.3 Task-oriented Dialogue Generation

A task-oriented dialogue system is often composed of several modules: a natural language understanding (NLU) module, a dialogue manager (DM), and a natural language generation (NLG) module [53, 78]. Intrinsic hallucination can occur between DM and NLG, where dialogue act such as `recommend(NAME=`*peninsula hotel*`, AREA=`*tsim sha tsui*`)` is transformed into natural language representation of "the hotel named *peninsula hotel* is located in *tsim sha tsui* area." [7, 102].

8.3.1 Hallucination Metrics. To evaluate hallucination, Li et al. [102] and Balakrishnan et al. [7] combine traditional metrics such as BLEU score and human evaluation as well as hallucination-specific automatic metrics. Following the previous works such as [36, 173, 187], Li et al. [102] use slot error rate, which is computed by $(p + q)/N$ where N represents the total number of slots extracted by another model in the dialogue act. Here, p stands for the missing slots in the generated template, and q is the number of redundant slots. On the other hand, Balakrishnan et al. [7] introduce a novel metric called Tree accuracy, which determines if the prediction's tree structure is identical to that of the input meaning representations.

8.3.2 Mitigation Methods. While Balakrishnan et al. [7] propose to adopt tree-structured semantic representations and add constraints on decoding, Li et al. [102] frame in reinforcement learning problem where they apply a bootstrapping algorithm to sample training instances and then leverage reward related to slot consistency. Recently, there has been another line of research in task-oriented dialogue, which is to build a single end-to-end system rather than connecting several modules (e.g., Eric and Manning [42], Madotto et al. [111, 114], Wu et al. [193]). As discussed in the other sections of this paper, there is a potential by such end-to-end systems to produce an extrinsic hallucination, yet this remains less explored. For example, a model might generate a response with an entity that appears out of nowhere. In the example of hotel recommendation in Hong Kong above, a model can generate a response such as "the hotel named *raffles hotel* is located in *central* area[6]," which cannot be verified from the knowledge base of the system.

8.4 Future Directions in Dialogue Generation

Self-Contradiction in Dialogue Systems. One of the possible reasons for self-contradiction is that current dialogue systems tend to have a short memory of dialogue history [151]. Thus, allowing a longer memory would be a future direction by exploring the following possible causes. Firstly, common dialogue datasets provide several turns of conversation, yet not long enough to assess a model's ability to deal with a long context. To overcome this, Xu et al. [196] introduce a new dataset that consists of, on average, over 40 utterances per episode. Secondly, we often truncate dialogue history into fewer turns to fit into models such as Transformer-based architectures, which makes a model difficult to memorize the past. In addition to the works of dialogue summarization, e.g., Gliwa et al. [59], this issue may benefit from other works which are aiming to grasp the longer context but do not focus on dialogue generation [9, 206, 215].

Fact-checking in dialogue systems. In addition to the factual consistency in responses from knowledge grounded dialogue systems, fact-checking in dialogue systems is a future direction of dealing with the hallucination problem in dialogue system [66]. The dialogue fact-checking involves verifiable claim detection, which is an important line of distinguishing hallucination-prone dialogue, and evidence retrieval from an external source. This fact-checking system in the dialogue system could be utilized not only as an evaluation metric for facilitating factual consistency but also as modeling such a system.

9 HALLUCINATION IN GENERATIVE QUESTION ANSWERING

Generative question answering (GQA) aims to generate an abstractive answer rather than extract an answer to a given question from provided passages [46, 97]. It is an important task since many of the everyday questions that humans deal with and pose to search engines require in-depth explanations [81] (e.g., *why/how..?*), and the answers are normally long and cannot be directly extracted from existing phrase spans. A GQA system can be integrated with a search engine [121] to empower more intelligent search or combined with a virtual conversation agent to enhance user experience.

Normally GQA system involves searching an external knowledge source for information relevant to the question. Then it generates the answer based on the retrieved information [85]. In most cases, no single source (document) contains the answer, and multiple retrieved documents will be considered for answer generation. Those documents may contain redundant, complementary, or contradictory information. Thus, hallucination is quite a common problem in the generated answer.

[6] Raffles Hotel is a hotel located in Downtown Core, Singapore.

Furthermore, the hallucination problem is one of the most important challenges in GQA. Since an essential goal of the GQA system is to provide factual-correct answers given the question, hallucination in the answer will mislead the user and damage the system performance dramatically.

9.1 Hallucination Definition in GQA

As a more challenging yet under-explored task, there is no standard definition of hallucination in GQA. However, almost all the work on GQA [46, 85, 126] involve a human evaluation process, in which the *factual correctness* measuring the faithfulness of the generated answer can be seen as a measurement of the hallucination, i.e., the more faithful the answer is, the less hallucinated content it contains. The most recent work [97] used the term *semantic drift*, which indicates how the answer drifts away from a correct one during generation, can also be seen as a specific definition of hallucination in GQA.

In order to be in line with the general categorization of hallucination in Section 2.1, we give two concrete hallucination examples in GQA in Table 1. The sources of both questions are Wikipedia web pages. For the first question, "*dow jones industrial average please?*", the generated answer "*index of 30 major U.S. stock indexes*" contracts with the statement "*of 30 prominent companies listed on stock exchanges in the United States*" from Wikipedia. So we categorize it as *intrinsic hallucination.* For the second example, the sentences "*The definition of a Sadducee is a person who acts in a deceitful or duplicitous manner. An example of a Sadduceee is a politician who acts deceitfully in order to gain political power*" in the generated answer can not be verified from the source documents; thus, we categorize it as *extrinsic hallucination.*

9.2 Hallucination-related Metrics in GQA

Currently, there is no automatic metric to evaluate hallucination in QGA yet specifically. While most works on GQA use automatic evaluation metrics such as ROUGE score and F1 to measure the quality of the answer, those N-gram overlap-based metrics are not a meaningful way to evaluate hallucination due to their poor correlation with human judgments, as indicated by Krishna et al. [85]. On the other hand, almost all the GQA related work involves the human evaluation process as complementary to the automatic evaluation. Normally human annotators will be asked to assign a score indicating the faithfulness of the answer, which can also be viewed as a measurement of the answer hallucination. However, the metrics obtained via human evaluation are only from a small sample of the data.

Metrics such as measuring *semantic overlap* [156], a learned evaluation metric based on BERT that models human judgments, could be considered as a better measurement of hallucination for GQA. Other metrics such as the *factual correctness* can also be considered as a way to measure hallucination in GQA. Zhang et al. [213] propose to explicitly measure the factual correctness of a generated text against the reference by first extracting facts via an information extraction (IE) module. Then they define and measure the factual accuracy score to be the ratio of facts in the generation text equal to the corresponding facts in the reference.

Factual consistency which measures the faithfulness of the generated answer given its source documents, can be employed as another way to measure hallucination in GQA. Durmus et al. [34], Wang et al. [179] propose an automatic question answering (QA) based metric to measure faithfulness in summary, leveraging the recent advances in machine reading comprehension. They first use a question generation model to construct question-answer pairs from the summary, then a QA model is applied to extract short answer spans from the given source document for the question, the extracted answers that not matches with the provided answers indicate unfaithful information in the summary. While these metrics were first proposed in the summarization works, they can be easily adopted in the generative QA to measure hallucinations in the generated long-form answer.

9.3 Hallucination Mitigation in GQA

Unlike conditional text generation tasks such as summarization, or data-to-text generation, in which the source documents are provided and normally related to the target generation, the hallucination problem in GQA is more complicated. Generally speaking, it might come from two sources: 1) the incompetency of the retriever, which retrieves irrelevant documents to the answer 2) the *intrinsic* and *extrinsic* hallucination in the conditional generation model itself. Normally these two parts are interconnected and cause hallucinations in the answer.

Thus, early works on GQA mostly try to improve the faithfulness of the answer by investigating reliable external knowledge sources or incorporating multiple information sources. Yin et al. [203] propose Neural Generative Question Answering (GENQA), an end-to-end model that generates answers to simple factoid questions based on the knowledge base. Moreover, Bi et al. [12] propose Knowledge-Enriched Answer Generator (KEAG) to generate a natural answer by integrating facts from four different information sources, i.e., question, passage, vocabulary, and knowledge.

Recent works focus more on the conditional generation model part. Fan et al. [45] construct a local knowledge graph for each question to compress the information and reduce redundancy from the retrieved documents, which can be viewed as an early trial to mitigate hallucination. While Li et al. [97] propose a novel model, Rationale-Enriched Answer Generator (REAG), in which they add an extraction task to obtain the rationale for an answer at the encoding stage, and the decoder is expected to generate the answer based on both the extracted rationale and original input. Recent work by Krishna et al. [85] employs a Routing Transformer (RT), a sparse attention-based Transformer-based model that employs local attention and mini-batch k-means clustering for long-range dependence, as the answer generator in the hope of modeling more retrieved documents to mitigate the hallucination in the answer.

Most recently, Lin et al. [104] propose a benchmark, which comprises 817 questions that span 38 categories, to measure the truthfulness of a language model in the QA task. This work investigates the performances of GPT-3 [17], GPT-Neo/J, GPT-2 [141] and a T5-based model. The results suggest that simply scaling up the model is less promising than fine-tuning it in terms of improving truthfulness since larger models are better at learning the training distribution from the web data thus tend to produce more imitative falsehoods. While Nakano et al. [126] fine-tuned GPT-3 to answer long-form questions with a web-browsing environment, which allows the model to navigate the web as well as use human feedback to optimize answer quality using imitation learning [76] directly.

9.4 Future Directions in GQA

While GQA is challenging yet under-explored, many possible directions could be explored to improve the answer quality and mitigate hallucination. First, better automatic evaluation metrics are needed to measure hallucination. The previously mentioned metrics, such as the semantic overlap between the generated answer and the ground-truth answer, the faithfulness of the generated answer, and factual consistency between the answer and the source documents, only consider one aspect of hallucination. Metrics that can consider all the factors related to hallucination (such as semantic overlap, faithfulness, or factual consistency) could be designed. Second, datasets with hallucination annotations should be proposed since none of the current GQA datasets has that information. Furthermore, another possible direction to mitigate hallucination in the answer is improving the performance of models. We need better retrieval models that retrieve relevant information according to queries and the generation models that can synthesize more accurate answers from multiple sourced documents.

10 HALLUCINATION IN DATA-TO-TEXT GENERATION

Data-to-Text Generation is the task of generating natural language descriptions conditioned on structured data [87, 119], such as tables [133, 191], database records [25], and knowledge graphs [54]. Although this field has been recently boosted by neural text generation models, it is well known that these models are prone to hallucinations [191] because of the gap between structured data and text, which may cause semantic misunderstanding and erroneous correlation. Moreover, the tolerance of hallucination is very low when this task is applied to the real world, such as in the case of patient information table description [169], and analysis of experimental results table in a scientific report. These years have seen a growth of interest in hallucinations in Data-to-Text Generation, and researchers have proposed works from the aspect of evaluation and mitigation.

10.1 Hallucination Definition in Data-to-Text Generation

The definition and categories of hallucination in Data-to-Text Generation follow the descriptions in Section 2. We follow the general hallucination definition in this task as: **(1) Intrinsic Hallucinations:** the generated text contains information that is contradicted with the input data [130]. For example, in Table 1, "*The Houston Rockets (18-4)*" use information "*[TEAM: Rockets, CITY:Houston, WIN:18, LOSS: 5]*" in the source table. However, "*(18-4)*" is contradicted with "*[LOSS: 5]*" and it should be "*(18-5)*". **(2) Extrinsic Hallucinations:** the generated text contains extra information irrelevant to the input [29, 130]. For example, in Table 1, "*Houston has won two straight games and six of their last seven.*" is not mentioned in the source table [181].

10.2 Hallucination Metrics in Data-to-Text Generation

Statistical. PARENT (Precision And Recall of Entailed Ngrams from the Table) [29] measures the accuracy of table-to-text generation by aligning n-grams from the reference description R and generated texts G to the table T. And it is the average F-score by combining the entailment precision and recall. Wang et al. [184] modify the PARENT and denote this table-focused version as PARENT-T. Different from PARENT which evaluates each instance (T_i, R_i, G_i), PARENT-T ignores the reference description R and evaluates each instance (T_i, G_i) separately.

Information Extraction (IE)-based. Liu et al. [107] estimate the generation hallucination with two entity-centric metrics, namely, table record coverage (the ratio of covered records in a table) and hallucinated ratio (the ratio of hallucinated entities in text). This metric firstly uses entity recognition to extract the entities of input and generated output; then aligns these entities by heuristic matching strategies; and finally calculates the ratios of faithful and hallucinated entities separately. Moreover, there are some general post-hoc IE-based metrics that could be applied to hallucination evaluation, such as Slot Error Rate (SER) [38, 200], Content Selection (CS), Relation Generation (RG), and Content Ordering (CO) [181, 191].

Natural Language Inference (NLI)-based. Dušek and Kasner [37] recognize the textual entailment between the input data and the output text for both omissions and hallucinations with an NLI model. This work measures the semantic accuracy in two directions: they check for omissions by inferring whether the input fact is entailed the generated text and check for hallucinations by inferring the generated text from the input fact.

Language Modeling (LM)-based. Filippova [50], Tian et al. [172] base on the intuition that when an unconditional LM which is only trained on the targets gets a smaller loss than a conditional LM_x which is trained on both sources and targets, the token is predicted unfaithfully. Thus, they calculate the ratio of hallucinated tokens to the total target length to measure the hallucination level.

10.3 Hallucination Mitigation in Data-to-Text Generation

Data-Related Method. Several clean and faithful corpora are collected to tackle the challenges from data infidelity. TOTTO [133] is an open-domain faithful table-to-text dataset, where each sample includes a Wikipedia table with several cells highlighted and a description. To ensure that targets exclude hallucinations, the annotators revise existing Wikipedia candidate sentences and clear the parts unsupported by the table. Moreover, RotoWire-FG (Fact-Grounding) [181] is a purified and enlarged and enriched version of RotoWire [191] generating NBA game summaries from score tables. Annotators trim the hallucination part in target texts and extract the mapped table records as content plans to better align input tables and output summaries.

For data processing, Nie et al. [130] utilize a language understanding module to improve the equivalence between the input meaning representation (MR) and the reference utterance in the dataset. They train an NLU model with an iterative relabeling procedure: First, they train the model on original data; parse the MR by model inference; train the model on new paired data with high confidence; and then repeat the above processes. Liu et al. [107] select training instances based on faithfulness ranking. Finer-grained than the above instance-level method, Rebuffel et al. [146] label tokens according to co-occurrence analysis and sentence structure through dependency parsing in the pre-processing step to explicate the correspondence between the input table and the text.

Modeling and Inference Method. Planning or skeleton is a common method in data-to-text tasks to improve the faithfulness to the input [124]. Liu et al. [107] propose a two-step generator with a separate text planner, which is augmented by auxiliary entity information. First, the planner predicts the plausible content plan based on the input data. Second, given the above input data and the content plan, the sequence generator generates the text. Similarly, Plan-then-Generate [163] also consists of a content planner and a sequence generator. In addition, this work adopts a structure-aware reinforcement learning (RL) training to generate output text following the generated content plan faithfully. SANA [182] is a skeleton-based two-stage model including skeleton generation to select key tokens from the source table and edit-based generation to produce texts via iterative insertion and deletion operations. In contrast to the above two-step model using planning or skeleton, AGGGEN [200] is an end-to-end model jointly learning to plan and generate at the same time. This architecture with a Hidden Markov Model (HMM) and Transformer encoder-decoder reintroduces explicit sentence planning stages into neural systems by aligning facts in the target text to input representations.

There are other modeling methods to mitigate the hallucination problem. Conjecturing that hallucinations can be caused by inattention to the source, [172] propose a confidence score and propose a variational Bayes training framework that can learn the score from data. Moreover, Wang et al. [184] introduce a new table-text optimal-transport (OT) matching loss and a table-text embedding similarity loss to encourage faithfulness. The hallucination degree can also be treated as a controllable factor in generating texts. In [50], for each training sample, the hallucination degree is estimated and converted into a categorical value as part of inputs, as in a controlled generation setting [49]. This approach does not require the dismissal of any input or modification of the model structure.

In order to mitigate hallucinations at the inference step, Rebuffel et al. [146] propose a Multi-Branch Decoder that leverages word-level alignment labels between input table and paired text to learn the relevant parts of the training instance. And these word-level labels are gained through dependency parsing during the pre-processing step. The branches integrate three co-dependent control factors: content, hallucination, and fluency separately. Uncertainty-aware beam search (UABS) [195] is an extension to beam search to reduce hallucination. Considering hallucination probability is positively correlated with predictive uncertainty, this work adds a weighted penalty term in beam search which is able to balance the predictive probability and uncertainty. Apart from data-to-text generation, this work can also be applied to other tasks, such as image captioning.

10.4 Future Directions in Data-to-Text Generation

Given the challenges brought by the discrepancy between structure data and natural text, and the low fault tolerance in the Data-to-Text Generation task, there are several potential directions worth exploring in terms of hallucination.

Firstly, numbers contain information about scales and are common and crucial in the Data-to-Text task [164, 212]. Moreover, it is frequent to have errors in numbers, which results in hallucinations and infidelity. This is a serious problem for Data-to-Text generation. However, models rarely give special consideration to the numbers found in the table or text [168]. The current automatic metrics of hallucinations also do not specifically treat numbers. This indiscriminate treatment contradicts findings in cognitive neuroscience, where numbers are known to be represented differently from lexical words in a different part of the brain [60]. Thus, considering or highlighting numbers when mitigating and assessing hallucinations is worth exploring. This requires the generative model to learn a better numerical presentation and capture scales, which will reduce the hallucinations caused by the misunderstanding of numbers.

Moreover, for the logical data-to-text generation task, instead of surface-level generation, logical inference, calculation, comparison are required, which is challenging and makes it easier to cause hallucinations. Thus, reasoning (including numerical reasoning), which usually combines with graph structure [24] is another direction to improve the accuracy of entity relationships and alleviate hallucinations.

11 HALLUCINATIONS IN NEURAL MACHINE TRANSLATION

Neural Machine Translation (NMT) is a task of generating translation of the source language into the target language via inference, given parallel data samples for training. Compared to statistical machine translation (SMT) the output of NMT is usually quite fluent and of human-level quality, which creates the danger of misinforming users when there are hallucinations [116].

11.1 Hallucinations Definition and Categories in NMT

The problem of hallucination was identified with the deployment of the first NMT models. Early work comparing SMT and NMT systems [83], without explicitly using the term "hallucination", mentioned that NMT models tend to "sacrifice adequacy for the sake of fluency" especially when evaluated with out-of-domain test sets. Following further development of NMT, most of the relevant research papers agree that translated text is considered a hallucination when the target text is completely disconnected from the source [91, 125]. The categorization of hallucination in NMT is unlike that in any other NLG tasks. Articles on machine translation use various categories and terms that are often overlapping. In order to maintain consistency with other NLG tasks, we use the intrinsic and extrinsic hallucination categories applied to the NMT task by [219] in this section. After a formal definition, we will describe other identified types of hallucinations and hallucination categories mentioned in the relevant literature.

Intrinsic and Extrinsic Hallucinations. Following the idea that the hallucinations are outputs that are disconnected from the source, [219] suggest categorizing the hallucinatory content based on the way the output is disconnected:

- Intrinsic hallucinations are translations that contain incorrect information compared to information present in the source. In Table 4, the example of such hallucination is "Jerry doesn't go", since the original name in the source is "Mike" and the verb "to go" is not negated.

Category	Source	Correct Translation	Hallucinatory Translation
Intrinsic	迈克周四去书店。	Mike goes to the bookstore on Thursday.	Jerry doesn't go to the bookstore on Thursday.
Extrinsic	迈克周四去书店。	Mike goes to the bookstore on Thursday.	Mike happily goes to the bookstore on Thursday with his friend.
Detached	Das kann man nur feststellen, wenn die kontrollen mit einer großen intensität durchgeführt werden.	This can only be detected if controls undertaken are more rigorous.	Blood alone moves the wheel of history, i say to you and you will understand, it is a privilege to fight.
Oscillatory	1995 das produktionsvolumen von 30 millionen pizzen wird erreicht.	1995 the production reached 30 million pizzas.	The US, for example, has been in the past two decades, but has been in the same position as the US, and has been in the United States.

Table 4. Categories and examples of hallucinations in MT by Zhou et al. [219] and Raunak et al. [145]

- Extrinsic hallucinations are translations that produce additional content without any regard to the source. In Table 4, "happily" and "with his friend" are the two examples of the hallucinatory content since they are added without any apparent connection to the input.

Other categories and types of hallucinations. Raunak et al. [145] propose an alternative categorization of hallucinations. They divide hallucinations into hallucinations under perturbations and natural hallucinations. Hallucinations under perturbation are types of hallucinations that can be observed if a model tested on the perturbed and unperturbed test set returns drastically different content. Their work on hallucinations under perturbation follows strictly the algorithm proposed by Lee et al. [91], see Section 11.2.2 on entropy measure. The second category, natural hallucinations, are created with connection to the noise in the dataset and can be further divided into detached and oscillatory, where detached hallucinations mean that a target translation is semantically disconnected from a source input, and the oscillatory hallucinations are decoupled from source by manifesting a repeating n-gram. Kong et al. [84], Tu et al. [175] analyze this phenomenon under the name of over-translation (a repetitive appearance of words that were not in the source text). Conversely, under-translation is skipping the words that need to be translated [175]. Finally, abrupt jumps to the end of the sequence and outputs that remain mostly in the source language are also examples of hallucinatory content [91].

11.2 Hallucination Metrics in NMT

Definition of hallucinations in machine translation tends to be qualitative and subjective, and thus researchers often identify hallucinated content manually. Most detrimentally, the appearance of hallucinations is found not to affect the BLEU score of the translated text [172, 219]. There are nevertheless several notable efforts to automatize and quantify the search for hallucinations by using statistical methods.

11.2.1 Statistical Metrics. Martindale et al. [116] propose identifying sentence adequacy using bag-of-vectors sentence similarity (BVSS) metric. This metric shows when the information is lost because the reference contains more information than the MT output or hallucinated, or the MT output contains more information than the reference.

11.2.2 Model Based Metrics.

Auxiliary Decoder. "Faithfulness" refers to the amount of the source meaning that is expressed in the translation faithfully, and it is used interchangeably with the term "adequacy" [48, 174]. Feng et al. [48] propose adding another "evaluation decoder" apart from the standard translation decoder. In their work "faithfulness" is based on word-by-word translation probabilities, and is calculated in the evaluation module along with translation fluency. The loss returned by the evaluation module helps to adjust the probability returned by the translation module.

Entropy Measure. In scenarios where the ground truth of translation is not available, an entropy measure of the average attention distribution can be used to detect hallucinations. Garg et al. [55], Tu et al. [175] show that hallucinations are visible in attention matrices. Attention networks in correct translations attend to the entire input sequence throughout decoding. However, it tends to concentrate on one point when the model outputs hallucinatory content. The entropy is calculated on the average attention weights when the model does or does not produce hallucinations during testing. For the comparison, the clean test set is used along with the purposefully perturbed one, which is created to incite hallucinations (test sets featuring multiple repetitions). The mean entropy returned by hallucinatory models diverges from the mean of the models that do not produce hallucinations spontaneously [91].

Token level hallucination detection. Zhou et al. [219] propose a method for detecting hallucinated tokens within a sentence, making the search more fine-grained. They use a synthetic dataset that is created by adding noise to the source data in a way similar to those in the Table 5 but generated by a language model with certain tokens of correct translations masked. Tokens in synthetic data are labeled as hallucinated (1) or not (0). Then authors compute hallucination prediction loss between binary labels and the tokens from the hallucinated sentence. This work further employs the word alignment-based method and overlap-based method as baselines for hallucination.

Similarity-based. Zhou et al. [219] use an unsupervised model that extracts alignments from similarity matrices of word embeddings [153], then predicts the target token as hallucinated if it is not aligned to the source. Parthasarathi et al. [134] propose calculating faithfulness by computing similarity scores between perturbed source sentence and target sentence after applying the same perturbation.

Overlap-based. Zhou et al. [219] predict that the target token is hallucinated if it does not appear in the source. Since target and source are two different languages, authors use the density matching method for bilingual synonyms from Zhou et al. [218]. Kong et al. [84] suggest Coverage Difference Ratio (CDR), as metric evaluating adequacy, that is especially successful in finding cases of under-translation. It is estimated by comparing source words covered by generated translation with human translations.

The overlap-based methods for detecting hallucinations are heuristics based on the assumption that all the translated words should appear in the source. However, it is not always the case, e.g., when paraphrasing or using synonyms. Using word embeddings as similarity-based methods helps avoid such simplifications and allows more diverse, synonymous translations.

Approximate Natural Hallucination Detection. Raunak et al. [145] propose Approximate Natural Hallucination (ANH) detection based on the fact that hallucinations often occur as oscillations (repeating n-grams) and the lower unique bigram count indicates a higher appearance of oscillatory hallucinations. Furthermore, the ANH detection method searches for repeated targets in the translation output. Their methods find translation above a certain n-gram threshold and search for repeated targets in the output translation, following the assumption that if hallucinations are often incited by aligning unique sources to the same target, then repeating targets will also appear during the inference [175].

11.3 Hallucination Mitigation Methods in NMT

Hallucinations in machine translation are hard to discover for a person who is not fluent in the target language and thus can lead to many possible errors, or even dangers. Probably out of all the natural language generation tasks, machine translation engines such as google in the English-speaking internet and Baidu in Sinosphere are most widely accessible by the netizens. Therefore, there are several suggested methods of mitigating hallucinations in NMT mentioned in the relevant articles.

11.3.1 Data-Related. Data augmentation appears to be one of the most common methods for removing hallucination. Lee et al. [91], Raunak et al. [145] suggest addition of perturbed sentences. Furthermore, perturbation, where the insertions of most common tokens are placed at the beginning of the sentence, seems to be most successful in hallucination mitigation. A disadvantage of this method is the need to understand different types of hallucinations produced by the model in order to apply a correct augmentation method. Corpus filtering is a method of mitigating hallucinations caused by the noise in the dataset by removing the repetitive and mismatching source and target sequences [145]. Junczys-Dowmunt [77] implements cross-entropy data filtering method for bilingual data, which uses cross-entropy scores calculated for noisy pairs according to two translation models trained on the clean data. The scores that suggest dissagreament between sentence pairs from two models are subsequently penalized.

While [77, 91, 145] define noise by mismatched source and target sentences, [16] analyze influence of fine-grained semantic divergences on NMT outputs. The authors consequently propose mitigation method for fine-grained divergences based on semantic factors. The tags are applied to each source and target sentence to inform about the position of divergent token. Factorizing divergence not only helps to mitigate hallucinations but improves overall performance of the NMT. This shows that tagging small semantic divergences can provide useful information for the network during training.

11.3.2 Modeling and Inference. Overexposure bias is a common problem in NMT amplified by the teacher forcing technique used in sequence-to-sequence models. The models are trained on the ground truth, but during inference, they attend to the past predictions which can be incorrect [84, 143]. To mitigate this problem, Wang and Sennrich [180] propose substituting MLE as a training objective with minimum risk training (MRT) [131]. Scheduled sampling is a classic method of mitigating overexposure bias first proposed by [10]. Based on that method [62] create a differentiable approximation to greedy decoding that shows a good performance in NMT task. [199] propose further improvement of scheduled sampling algorithm for NMT by optimizing the probability of source and target word alignments. This improvement helps to address the issue flexibility in word order between a source and target language when performing scheduled sampling.

Zhou et al. [219] propose a method of improving self-training of NMT based on hallucination detection. They create hallucination labels (see: Section 11.2.2), and then discard losses of tokens predicted as hallucinations, which is known as token loss truncation. This is similar to the method proposed by Kang and Hashimoto [79], the latter for full sentences in the summarization task. Furthermore, instead of adjusting losses, authors mask the hidden states of the discarded losses in the decoder in a procedure called decoder HS masking. Experimental results show both a translation quality improvement in terms of BLEU and also a large reduction in hallucination. The token loss truncation method shows good results in the low-resource languages scenario.

Perturbation Method	Source	Target
Unique-Unique: Pair unique source with an unrelated target.	迈克周四去书店。 (Michael goes to the bookstore on Thursday.)	She likes pink flamingos.
	她买了一只黑猫。 (She bought a black cat.)	The weather is great today.
Repeat-Repeat: Pair unique source with unrelated target and repeat such pair multiple times.	迈克周四去书店。 (Michael goes to the bookstore on Thursday.)	She likes pink flamingos.
	迈克周四去书店。 (Michael goes to the bookstore on Thursday.)	She likes pink flamingos.
Repeat-Unique: Pair the same source with multiple different targets.	迈克周四去书店。 (Michael goes to the bookstore on Thursday.	She likes pink flamingos.
	迈克周四去书店。 (Michael goes to the bookstore on Thursday.	The weather is great today.
Unique-Repeat: Pair unique sources with repeating targets.	迈克周四去书店。 (Michael goes to the bookstore on Thursday.)	She likes pink flamingos.
	她买了一只黑猫。 (She bought a black cat.)	She likes pink flamingos.

Table 5. Examples of perturbations [145]. Datasets perturbed using the following methods can be used for data augmentation, one of the most successful hallucination mitigation methods in neural machine translation.

Another method to mitigate the impact of the noisy datasets is tilted empirical risk minimization (TERM), a training objective proposed by Li et al. [101].

Dropout, L2E regularization, and clipping decrease the number of hallucinations [91]. Several authors propose methods of improving phrase alignment that are helpful both in increasing translation accuracy and identifying contents that did not appear in the source translation [55, 189, 210].

11.4 Future Directions in NMT

The future work on hallucinations in NMT is to define hallucinations in a quantifiable manner, i.e., to specify a cut-off value between translation error and hallucinated content using a particular metric. Martindale et al. [116] propose a threshold between fluency and adequacy, which is the closest to this ideal. Authors, however, do not concentrate on hallucinated content as such, and thus fluent but inadequate sentences may not always indicate hallucinations but also other types of translation errors. Balakrishnan et al. [7] mention constrained decoding as a method to mitigate hallucinations in dialogue systems. It could be, however, also applied in NMT. [32, 70, 140, 161, 166, 197, 198] use constrained decoding to incorporate specific terminology to machine translation, but the above methods can be repurposed to mitigate hallucinations.

Another direction for future work on hallucinations is improving existing methods of searching for hallucinatory content, such as algorithms proposed by Feng et al. [48], Lee et al. [91], Raunak et al. [145], that are computationally expensive [145] or require the creation of additional perturbed test-set [91]. Similarly, for mitigation of lack of faithfulness and fluency, the method proposed by Feng et al. [48] requires the creation of one-to-many architecture (one encoder and two decoders) which is also computationally expensive. The future directions would therefore include simplification of existing hallucination evaluation methods, applying them to different architectures like

CNNs and transformers, and possibly conducting research on finding more simplified hallucination search methods.

12 CONCLUSION

In this survey, we provide a first comprehensive overview of the hallucination problem in NLG; we summarize existing evaluation metrics, mitigation methods, and remaining challenges for future research. Hallucination is an artefact of neural-based natural language generation and is of concern because they appear fluent and therefore can be misleading to users. In some scenarios and tasks, hallucination can cause harm. We survey various contributors to hallucination, ranging from noisy data, erroneous parametric knowledge, incorrect attention mechanism, inappropriate training strategy, to inference exposure bias, etc. We show that there are two categories of hallucinations, namely intrinsic hallucination and extrinsic hallucination, and they need to be treated differently with different mitigation strategies. Hallucination is relatively easy to detect in abstractive summarization and in NMT against the evidence in the source. For dialog systems, it is important to balance diversity vs consistency in dialog responses. Hallucination in GQA is detrimental to its performance, but research in mitigation methods is still very preliminary in this area. For data-to-text generation, hallucination arises from the discrepancy between the input and output format. Most methods to mitigate hallucinations in NMT either try to reduce dataset noise or alleviate exposure bias. There remains a lot of challenges ahead in identifying and mitigating hallucinations in NLG, and we hope research in this area can benefit from this survey.

REFERENCES

[1] Joshua Albrecht and Rebecca Hwa. 2007. A Re-examination of Machine Learning Approaches for Sentence-Level MT Evaluation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Association for Computational Linguistics, Prague, Czech Republic, 880–887. https://aclanthology.org/P07-1111

[2] Rahul Aralikatte, Shashi Narayan, Joshua Maynez, Sascha Rothe, and Ryan McDonald. 2021. Focus Attention: Promoting Faithfulness and Diversity in Summarization. *ACL* (2021).

[3] Kristjan Arumae and Fei Liu. 2019. Guiding Extractive Summarization with Question-Answering Rewards. In *NAACL-HLT (1)*.

[4] Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.

[5] Xiang Bai, Xinggang Wang, Longin Jan Latecki, Wenyu Liu, and Zhuowen Tu. 2009. Active skeleton for non-rigid object detection. In *2009 IEEE 12th International Conference on Computer Vision*. 575–582. https://doi.org/10.1109/ICCV.2009.5459188

[6] S. Baker and T. Kanade. 2000. Hallucinating faces. In *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*. 83–88. https://doi.org/10.1109/AFGR.2000.840616

[7] Anusha Balakrishnan, Jinfeng Rao, Kartikeya Upasani, Michael White, and Rajen Subba. 2019. Constrained Decoding for Neural NLG from Compositional Representations in Task-Oriented Dialogue. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 831–844. https://doi.org/10.18653/v1/P19-1080

[8] Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018. Predicting factuality of reporting and bias of news media sources. *arXiv preprint arXiv:1810.01765* (2018).

[9] Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. *arXiv:2004.05150* (2020).

[10] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent Neural networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1*. 1171–1179.

[11] Anne Beyer, Sharid Loáiciga, and David Schlangen. 2021. Is Incoherence Surprising? Targeted Evaluation of Coherence Prediction from Language Models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 4164–4173. https://doi.org/10.18653/v1/2021.naacl-main.328

[12] Bin Bi, Chen Wu, Ming Yan, Wei Wang, Jiangnan Xia, and Chenliang Li. 2019. Incorporating External Knowledge into Machine Reading for Generative Question Answering. In *Proceedings of the 2019 Conference on Empirical Methods*

in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2521–2530.

[13] Andrzej Białecki, Robert Muir, Grant Ingersoll, and Lucid Imagination. 2012. Apache lucene 4. In *SIGIR 2012 workshop on open source information retrieval*. 17.

[14] Ali Furkan Biten, Lluís Gómez, and Dimosthenis Karatzas. 2022. Let There Be a Clock on the Beach: Reducing Object Hallucination in Image Captioning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 1381–1390.

[15] Jan Dirk Blom. [n. d.]. *A dictionary of hallucinations*. Springer.

[16] Eleftheria Briakou and Marine Carpuat. 2021. Beyond Noise: Mitigating the Impact of Fine-grained Semantic Divergences on Neural Machine Translation. *CoRR* abs/2105.15087 (2021). arXiv:2105.15087 https://arxiv.org/abs/2105.15087

[17] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901. https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[18] Meng Cao, Yue Dong, Jiapeng Wu, and Jackie Chi Kit Cheung. 2020. Factual Error Correction for Abstractive Summarization Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 6251–6258.

[19] Shuyang Cao and Lu Wang. 2021. CLIFF: Contrastive Learning for Improving Faithfulness and Factuality in Abstractive Summarization. *EMNLP* (2021).

[20] 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*, Vol. 32.

[21] Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B Brown, Dawn Song, Úlfar Erlingsson, et al. 2020. Extracting Training Data from Large Language Models. (2020).

[22] Sihao Chen, Fan Zhang, Kazoo Sone, and Dan Roth. 2021. Improving faithfulness in abstractive summarization with contrast candidate generation and selection. *NAACL* (2021).

[23] Wenqing Chen, Jidong Tian, Yitian Li, Hao He, and Yaohui Jin. 2021. De-Confounded Variational Encoder-Decoder for Logical Table-to-Text Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 5532–5542.

[24] Zhiyu Chen, Wenhu Chen, Hanwen Zha, Xiyou Zhou, Yunkai Zhang, Sairam Sundaresan, and William Yang Wang. 2020. Logic2Text: High-Fidelity Natural Language Generation from Logical Forms. In *EMNLP (Findings)*.

[25] Andrew Chisholm, Will Radford, and Ben Hachey. 2017. Learning to generate one-sentence biographies from Wikidata. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. 633–642.

[26] Michael Crawshaw. 2020. Multi-task learning with deep neural networks: A survey. *arXiv preprint arXiv:2009.09796* (2020).

[27] Leyang Cui, Yu Wu, Shujie Liu, Yue Zhang, and Ming Zhou. 2020. MuTual: A Dataset for Multi-Turn Dialogue Reasoning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 1406–1416.

[28] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[29] 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*. 4884–4895.

[30] Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, et al. 2020. The second conversational intelligence challenge (convai2). In *The NeurIPS'18 Competition*. Springer, 187–208.

[31] Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. *ICLR* (2019).

[32] Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. 2019. Training Neural Machine Translation To Apply Terminology Constraints. *ACL 2019 - 57th Annual Meeting of the Association for Computational*

Linguistics, Proceedings of the Conference (6 2019), 3063–3068. https://doi.org/10.18653/v1/p19-1294

[33] Yue Dong, Shuohang Wang, Zhe Gan, Yu Cheng, Jackie Chi Kit Cheung, and Jingjing Liu. 2020. Multi-Fact Correction in Abstractive Text Summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 9320–9331.

[34] 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*. 5055–5070.

[35] Ondřej Dušek, David M Howcroft, and Verena Rieser. 2019. Semantic Noise Matters for Neural Natural Language Generation. In *Proceedings of the 12th International Conference on Natural Language Generation*. 421–426.

[36] Ondřej Dušek and Filip Jurčíček. 2016. Sequence-to-Sequence Generation for Spoken Dialogue via Deep Syntax Trees and Strings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Berlin, Germany, 45–51. https://doi.org/10.18653/v1/P16-2008

[37] Ondřej Dušek and Zdeněk Kasner. 2020. Evaluating Semantic Accuracy of Data-to-Text Generation with Natural Language Inference. In *Proceedings of the 13th International Conference on Natural Language Generation*. Association for Computational Linguistics, Dublin, Ireland, 131–137. https://aclanthology.org/2020.inlg-1.19

[38] Ondřej Dušek, Jekaterina Novikova, and Verena Rieser. 2020. Evaluating the state-of-the-art of end-to-end natural language generation: The e2e nlg challenge. *Computer Speech & Language* 59 (2020), 123–156.

[39] Nouha Dziri, Ehsan Kamalloo, Kory Mathewson, and Osmar Zaiane. 2019. Evaluating Coherence in Dialogue Systems using Entailment. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 3806–3812. https://doi.org/10.18653/v1/N19-1381

[40] Nouha Dziri, Andrea Madotto, Osmar Zaiane, and Avishek Joey Bose. 2021. Neural Path Hunter: Reducing Hallucination in Dialogue Systems via Path Grounding. *EMNLP* (2021).

[41] Nouha Dziri, Hannah Rashkin, Tal Linzen, and David Reitter. 2021. Evaluating Groundedness in Dialogue Systems: The BEGIN Benchmark. *Findings of ACL* (2021).

[42] Mihail Eric and Christopher Manning. 2017. A Copy-Augmented Sequence-to-Sequence Architecture Gives Good Performance on Task-Oriented Dialogue. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Association for Computational Linguistics, Valencia, Spain, 468–473. https://aclanthology.org/E17-2075

[43] Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel S. Weld. 2008. Open Information Extraction from the Web. *Commun. ACM* 51, 12 (Dec. 2008), 68–74. https://doi.org/10.1145/1409360.1409378

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

[45] Angela Fan, Claire Gardent, Chloé Braud, and Antoine Bordes. 2019. Using Local Knowledge Graph Construction to Scale Seq2Seq Models to Multi-Document Inputs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 4186–4196.

[46] Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long Form Question Answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 3558–3567.

[47] Alhussein Fawzi, Horst Samulowitz, Deepak Turaga, and Pascal Frossard. 2016. Image inpainting through neural networks hallucinations. In *2016 IEEE 12th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*. Ieee, 1–5.

[48] Yang Feng, Wanying Xie, Shuhao Gu, Chenze Shao, Wen Zhang, Zhengxin Yang, and Dong Yu. 2020. Modeling fluency and faithfulness for diverse neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 59–66.

[49] Jessica Ficler and Yoav Goldberg. 2017. Controlling Linguistic Style Aspects in Neural Language Generation. In *Proceedings of the Workshop on Stylistic Variation*. 94–104.

[50] Katja Filippova. 2020. Controlled Hallucinations: Learning to Generate Faithfully from Noisy Data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*. 864–870.

[51] William Fish et al. 2009. *Perception, hallucination, and illusion*. OUP USA.

[52] Saadia Gabriel, Asli Celikyilmaz, Rahul Jha, Yejin Choi, and Jianfeng Gao. 2021. GO FIGURE: A Meta Evaluation of Factuality in Summarization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Association for Computational Linguistics, Online, 478–487. https://doi.org/10.18653/v1/2021.findings-acl.42

[53] Jianfeng Gao, Michel Galley, and Lihong Li. 2018. Neural Approaches to Conversational AI. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*. Association for Computational

Linguistics, Melbourne, Australia, 2–7. https://doi.org/10.18653/v1/P18-5002
[54] Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. Creating training corpora for nlg micro-planning. In *55th annual meeting of the Association for Computational Linguistics (ACL)*.
[55] Sarthak Garg, Stephan Peitz, Udhyakumar Nallasamy, and Matthias Paulik. 2019. Jointly Learning to Align and Translate with Transformer Models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 4453–4462.
[56] Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research* 61 (2018), 65–170.
[57] Deepanway Ghosal, Pengfei Hong, Siqi Shen, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. 2021. CIDER: Commonsense Inference for Dialogue Explanation and Reasoning. *ACL* (2021).
[58] Alexandru L Ginsca, Adrian Popescu, and Mihai Lupu. 2015. Credibility in information retrieval. *Foundations and Trends in Information Retrieval* 9, 5 (2015), 355–475.
[59] Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum Corpus: A Human-annotated Dialogue Dataset for Abstractive Summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*. Association for Computational Linguistics, Hong Kong, China, 70–79. https://doi.org/10.18653/v1/D19-5409
[60] Silke M Göbel and Matthew FS Rushworth. 2004. Cognitive neuroscience: acting on numbers. *Current Biology* 14, 13 (2004), R517–R519.
[61] 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 & Data Mining*. 166–175.
[62] Kartik Goyal, Chris Dyer, and Taylor Berg-Kirkpatrick. 2017. Differentiable Scheduled Sampling for Credit Assignment. *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)* 2 (4 2017), 366–371. https://doi.org/10.18653/v1/P17-2058
[63] Tanya Goyal and Greg Durrett. 2020. Evaluating Factuality in Generation with Dependency-level Entailment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*. 3592–3603.
[64] Jian Guan and Minlie Huang. 2020. UNION: An Unreferenced Metric for Evaluating Open-ended Story Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 9157–9166.
[65] Beliz Gunel, Chenguang Zhu, Michael Zeng, and Xuedong Huang. 2020. Mind the facts: Knowledge-boosted coherent abstractive text summarization. *arXiv preprint arXiv:2006.15435* (2020).
[66] Prakhar Gupta, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2021. DialFact: A Benchmark for Fact-Checking in Dialogue. *arXiv preprint arXiv:2110.08222* (2021).
[67] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A Smith. 2018. Annotation Artifacts in Natural Language Inference Data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. 107–112.
[68] Braden Hancock, Antoine Bordes, Pierre-Emmanuel Mazare, and Jason Weston. 2019. Learning from Dialogue after Deployment: Feed Yourself, Chatbot!. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 3667–3684. https://doi.org/10.18653/v1/P19-1358
[69] Tianxing He, Jingzhao Zhang, Zhiming Zhou, and James Glass. 2021. Exposure Bias versus Self-Recovery: Are Distortions Really Incremental for Autoregressive Text Generation?. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 5087–5102.
[70] Chris Hokamp and Qun Liu. 2017. Lexically Constrained Decoding for Sequence Generation Using Grid Beam Search. *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)* 1 (4 2017), 1535–1546. https://doi.org/10.18653/v1/P17-1141
[71] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The Curious Case of Neural Text Degeneration. In *International Conference on Learning Representations*.
[72] Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. Q^2: Evaluating Factual Consistency in Knowledge-Grounded Dialogues via Question Generation and Question Answering. *EMNLP* (2021).
[73] Luyang Huang, Lingfei Wu, and Lu Wang. 2020. Knowledge graph-augmented abstractive summarization with semantic-driven cloze reward. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (2020).
[74] Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in building intelligent open-domain dialog systems. *ACM Transactions on Information Systems (TOIS)* 38, 3 (2020), 1–32.
[75] Yichong Huang, Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2021. The Factual Inconsistency Problem in Abstractive Text Summarization: A Survey. *arXiv preprint arXiv:2104.14839* (2021).

[76] Ahmed Hussein, Mohamed Medhat Gaber, Eyad Elyan, and Chrisina Jayne. 2017. Imitation Learning: A Survey of Learning Methods. *ACM Comput. Surv.* 50, 2, Article 21 (apr 2017), 35 pages. https://doi.org/10.1145/3054912
[77] Marcin Junczys-Dowmunt. 2019. Dual Conditional Cross-Entropy Filtering of Noisy Parallel Corpora. arXiv:1809.00197 [cs.CL]
[78] Daniel Jurafsky and James H. Marin. 2019. *Speech and Language Processing*. Draft of October 16th, 2019, Website: https://web.stanford.edu/~jurafsky/slp3/26.pdf, Chapter 26.
[79] Daniel Kang and Tatsunori Hashimoto. 2020. Improved Natural Language Generation via Loss Truncation. (4 2020), 718–731. https://arxiv.org/abs/2004.14589v2
[80] Osman Semih Kayhan, Bart Vredebregt, and Jan C van Gemert. 2021. Hallucination In Object Detection—A Study In Visual Part VERIFICATION. In *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2234–2238.
[81] Daniel Khashabi, Amos Ng, Tushar Khot, Ashish Sabharwal, Hannaneh Hajishirzi, and Chris Callison-Burch. 2021. GooAQ: Open Question Answering with Diverse Answer Types. *arXiv preprint arXiv:2104.08727* (2021).
[82] Philipp Koehn and Rebecca Knowles. 2017. Six Challenges for Neural Machine Translation. In *First Workshop on Neural Machine Translation*. Association for Computational Linguistics, 28–39.
[83] Philipp Koehn and Rebecca Knowles. 2017. Six Challenges for Neural Machine Translation. (2017), 28–39. http://www.statmt.org/wmt17/
[84] Xiang Kong, Zhaopeng Tu, Shuming Shi, Eduard Hovy, and Tong Zhang. 2019. Neural Machine Translation with Adequacy-Oriented Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 6618–6625.
[85] Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. 2021. Hurdles to Progress in Long-form Question Answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 4940–4957.
[86] 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)*. 9332–9346.
[87] Karen Kukich. 1983. Design of a knowledge-based report generator. In *21st Annual Meeting of the Association for Computational Linguistics*. 145–150.
[88] Ilia Kulikov, Alexander H. Miller, Kyunghyun Cho, and Jason Weston. 2019. Importance of Search and Evaluation Strategies in Neural Dialogue Modeling. In *Proceedings of the 12th International Conference on Natural Language Generation, INLG 2019, Tokyo, Japan, October 29 - November 1, 2019*, Kees van Deemter, Chenghua Lin, and Hiroya Takamura (Eds.). Association for Computational Linguistics, 76–87. https://doi.org/10.18653/v1/W19-8609
[89] Philippe Laban, Tobias Schnabel, Paul N Bennett, and Marti A Hearst. 2021. SummaC: Re-Visiting NLI-based Models for Inconsistency Detection in Summarization. *arXiv preprint arXiv:2111.09525* (2021).
[90] Rémi Lebret, David Grangier, and Michael Auli. 2016. Neural text generation from structured data with application to the biography domain. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (2016).
[91] Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. 2019. Hallucinations in neural machine translation. *ICLR* (2019).
[92] Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2021. Deduplicating training data makes language models better. *arXiv preprint arXiv:2107.06499* (2021).
[93] Nayeon Lee, Belinda Z Li, Sinong Wang, Wen-Tau Yih, Hao Ma, and Madian Khabsa. 2020. Language Models as Fact Checkers? *ACL 2020* (2020), 36.
[94] Nayeon Lee, Chien-Sheng Wu, and Pascale Fung. [n. d.]. Improving Large-Scale Fact-Checking using Decomposable Attention Models and Lexical Tagging. ([n. d.]).
[95] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 7871–7880.
[96] Bohan Li, Yutai Hou, and Wanxiang Che. 2021. Data Augmentation Approaches in Natural Language Processing: A Survey. *arXiv preprint arXiv:2110.01852* (2021).
[97] Chenliang Li, Bin Bi, Ming Yan, Wei Wang, and Songfang Huang. 2021. Addressing Semantic Drift in Generative Question Answering with Auxiliary Extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. 942–947.
[98] 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*. 1430–1441.

[99] Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A Persona-Based Neural Conversation Model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, 994–1003. https://doi.org/10.18653/v1/P16-1094

[100] Margaret Li, Stephen Roller, Ilia Kulikov, Sean Welleck, Y-Lan Boureau, Kyunghyun Cho, and Jason Weston. 2020. Don't Say That! Making Inconsistent Dialogue Unlikely with Unlikelihood Training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 4715–4728. https://doi.org/10.18653/v1/2020.acl-main.428

[101] Tian Li, Ahmad Beirami, Maziar Sanjabi, and Virginia Smith. 2020. Tilted Empirical Risk Minimization. (7 2020). https://arxiv.org/abs/2007.01162v2

[102] Yangming Li, Kaisheng Yao, Libo Qin, Wanxiang Che, Xiaolong Li, and Ting Liu. 2020. Slot-consistent NLG for Task-oriented Dialogue Systems with Iterative Rectification Network. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 97–106. https://doi.org/10.18653/v1/2020.acl-main.10

[103] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.

[104] Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. TruthfulQA: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958* (2021).

[105] Ce Liu, Heung-Yeung Shum, and William T Freeman. 2007. Face hallucination: Theory and practice. *International Journal of Computer Vision* 75, 1 (2007), 115–134.

[106] Tianyu Liu, Yizhe Zhang, Chris Brockett, Yi Mao, Zhifang Sui, Weizhu Chen, and Bill Dolan. 2021. A Token-level Reference-free Hallucination Detection Benchmark for Free-form Text Generation. *arXiv preprint arXiv:2104.08704* (2021).

[107] Tianyu Liu, Xin Zheng, Baobao Chang, and Zhifang Sui. 2021. Towards Faithfulness in Open Domain Table-to-text Generation from an Entity-centric View. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 13415–13423.

[108] Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-Based Knowledge Conflicts in Question Answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 7052–7063.

[109] Yukun Ma, Khanh Linh Nguyen, Frank Z Xing, and Erik Cambria. 2020. A survey on empathetic dialogue systems. *Information Fusion* 64 (2020), 50–70.

[110] Fiona Macpherson and Dimitris Platchias. 2013. *Hallucination: Philosophy and psychology*. MIT Press.

[111] Andrea Madotto, Samuel Cahyawijaya, Genta Indra Winata, Yan Xu, Zihan Liu, Zhaojiang Lin, and Pascale Fung. 2020. Learning Knowledge Bases with Parameters for Task-Oriented Dialogue Systems. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 2372–2394. https://doi.org/10.18653/v1/2020.findings-emnlp.215

[112] Andrea Madotto, Zhaojiang Lin, Chien-Sheng Wu, and Pascale Fung. 2019. Personalizing dialogue agents via meta-learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 5454–5459.

[113] Andrea Madotto, Zihan Liu, Zhaojiang Lin, and Pascale Fung. 2020. Language Models as Few-Shot Learner for Task-Oriented Dialogue Systems. arXiv:2008.06239 [cs.CL]

[114] Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. 2018. Mem2Seq: Effectively Incorporating Knowledge Bases into End-to-End Task-Oriented Dialog Systems. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Melbourne, Australia, 1468–1478. https://doi.org/10.18653/v1/P18-1136

[115] Amr Magdy and Nayer Wanas. 2010. Web-based statistical fact checking of textual documents. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*. 103–110.

[116] Marianna J. Martindale, Marine Carpuat, Kevin Duh, and Paul McNamee. 2019. Identifying Fluently Inadequate Output in Neural and Statistical Machine Translation. In *MTSummit*.

[117] 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*. 1906–1919.

[118] Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. Training Millions of Personalized Dialogue Agents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 2775–2779. https://doi.org/10.18653/v1/D18-1298

[119] Kathleen McKeown. 1992. *Text generation*. Cambridge University Press.

[120] Mohsen Mesgar, Edwin Simpson, and Iryna Gurevych. 2021. Improving Factual Consistency Between a Response and Persona Facts. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, Online, 549–562. https://doi.org/10.18653/v1/2021.eacl-main.44

[121] Donald Metzler, Yi Tay, Dara Bahri, and Marc Najork. 2021. Rethinking Search: Making Experts out of Dilettantes. *arXiv preprint arXiv:2105.02274* (2021).

[122] Sabrina J Mielke, Arthur Szlam, Y-Lan Boureau, and Emily Dinan. 2020. Linguistic calibration through metacognition: aligning dialogue agent responses with expected correctness. *arXiv preprint arXiv:2012.14983* (2020).

[123] 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*. 1322–1336.

[124] Amit Moryossef, Yoav Goldberg, and Ido Dagan. 2019. Step-by-step: Separating planning from realization in neural data-to-text generation. *arXiv preprint arXiv:1904.03396* (2019).

[125] Mathias Müller, Annette Rios, and Rico Sennrich. 2020. Domain Robustness in Neural Machine Translation. In *14th Conference of the Association for Machine Translation in the Americas*. Association for Machine Translation in the Americas, AMTA, 151–164.

[126] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. WebGPT: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332* (2021).

[127] Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cicero dos Santos, Henghui Zhu, Dejiao Zhang, Kathleen McKeown, and Bing Xiang. 2021. Entity-level Factual Consistency of Abstractive Text Summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 2727–2733.

[128] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. In *CoCo@ NIPS*.

[129] Feng Nie, Jinpeng Wang, Jin-Ge Yao, Rong Pan, and Chin-Yew Lin. 2018. Operation-guided Neural Networks for High Fidelity Data-To-Text Generation. In *EMNLP*.

[130] 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*. 2673–2679.

[131] Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st annual meeting of the Association for Computational Linguistics*. 160–167.

[132] Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding Factuality in Abstractive Summarization with FRANK: A Benchmark for Factuality Metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 4812–4829.

[133] Ankur Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. ToTTo: A Controlled Table-To-Text Generation Dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1173–1186.

[134] Prasanna Parthasarathi, Koustuv Sinha, Joelle Pineau, and Adina Williams. 2021. Sometimes We Want Translationese. arXiv:2104.07623 [cs.CL]

[135] 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, Volume 2 (Short Papers)*. 646–653.

[136] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language Models as Knowledge Bases?. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 2463–2473. https://doi.org/10.18653/v1/D19-1250

[137] Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis Only Baselines in Natural Language Inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*. 180–191.

[138] Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2016. Credibility assessment of textual claims on the web. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. 2173–2178.

[139] Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018. Declare: Debunking fake news and false claims using evidence-aware deep learning. *arXiv preprint arXiv:1809.06416* (2018).

[140] Matt Post and David Vilar. 2018. Fast Lexically Constrained Decoding with Dynamic Beam Allocation for Neural Machine Translation. *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association*

for Computational Linguistics: Human Language Technologies - Proceedings of the Conference 1 (4 2018), 1314–1324. https://doi.org/10.18653/v1/n18-1119

[141] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog* 1, 8 (2019), 9.

[142] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. *ICLR* (2016).

[143] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence Level Training with Recurrent Neural Networks. arXiv:1511.06732 [cs.LG]

[144] Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. 2021. Increasing faithfulness in knowledge-grounded dialogue with controllable features. *ACL* (2021).

[145] Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. The Curious Case of Hallucinations in Neural Machine Translation. (4 2021), 1172–1183. https://arxiv.org/abs/2104.06683v1

[146] Clément Rebuffel, Marco Roberti, Laure Soulier, Geoffrey Scoutheeten, Rossella Cancelliere, and Patrick Gallinari. 2021. Controlling Hallucinations at Word Level in Data-to-Text Generation. *arXiv preprint arXiv:2102.02810* (2021).

[147] Ehud Reiter. 2018. A structured review of the validity of BLEU. *Computational Linguistics* 44, 3 (2018), 393–401.

[148] Ehud Reiter and Robert Dale. 1997. Building applied natural language generation systems. *Natural Language Engineering* 3, 1 (1997), 57–87.

[149] Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How Much Knowledge Can You Pack Into the Parameters of a Language Model?. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 5418–5426. https://doi.org/10.18653/v1/2020.emnlp-main.437

[150] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object Hallucination in Image Captioning. In *EMNLP*.

[151] Stephen Roller, Y-Lan Boureau, Jason Weston, Antoine Bordes, Emily Dinan, Angela Fan, David Gunning, Da Ju, Margaret Li, Spencer Poff, et al. 2020. Open-domain conversational agents: Current progress, open problems, and future directions. *arXiv preprint arXiv:2006.12442* (2020).

[152] Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, et al. 2021. Recipes for Building an Open-Domain Chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 300–325.

[153] Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High Quality Word Alignments Without Parallel Training Data Using Static and Contextualized Embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. 1627–1643.

[154] Sashank Santhanam, Behnam Hedayatnia, Spandana Gella, Aishwarya Padmakumar, Seokhwan Kim, Yang Liu, and Dilek Hakkani-Tur. 2021. Rome was built in 1776: A Case Study on Factual Correctness in Knowledge-Grounded Response Generation. *arXiv preprint arXiv:2110.05456* (2021).

[155] Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368* (2017).

[156] Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning Robust Metrics for Text Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 7881–7892.

[157] Lei Shen, Haolan Zhan, Xin Shen, Hongshen Chen, Xiaofang Zhao, and Xiaodan Zhu. 2021. Identifying Untrustworthy Samples: Data Filtering for Open-domain Dialogues with Bayesian Optimization. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 1598–1608.

[158] Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval Augmentation Reduces Hallucination in Conversation. *EMNLP* (2021).

[159] Haoyu Song, Wei-Nan Zhang, Jingwen Hu, and Ting Liu. 2020. Generating Persona Consistent Dialogues by Exploiting Natural Language Inference. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 05 (Apr. 2020), 8878–8885. https://doi.org/10.1609/aaai.v34i05.6417

[160] Kaiqiang Song, Logan Lebanoff, Qipeng Guo, Xipeng Qiu, Xiangyang Xue, Chen Li, Dong Yu, and Fei Liu. 2020. Joint parsing and generation for abstractive summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 8894–8901.

[161] Kai Song, Yue Zhang, Heng Yu, Weihua Luo, Kun Wang, and Min Zhang. 2019. Code-Switching for Enhancing NMT with Pre-Specified Translation. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference* 1 (4 2019), 449–459. https://arxiv.org/abs/1904.09107v4

[162] Hui Su, Xiaoyu Shen, Sanqiang Zhao, Zhou Xiao, Pengwei Hu, Cheng Niu, and Jie Zhou. 2020. Diversifying Dialogue Generation with Non-Conversational Text. In *58th Annual Meeting of the Association for Computational Linguistics*. ACL, 7087–7097.

[163] Yixuan Su, David Vandyke, Sihui Wang, Yimai Fang, and Nigel Collier. 2021. Plan-then-Generate: Controlled Data-to-Text Generation via Planning. *Findings of EMNLP* (2021).

[164] Lya Hulliyyatus Suadaa, Hidetaka Kamigaito, Kotaro Funakoshi, Manabu Okumura, and Hiroya Takamura. 2021. Towards table-to-text generation with numerical reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 1451–1465.

[165] Yanli Sun. 2010. Mining the Correlation between Human and Automatic Evaluation at Sentence Level.. In *LREC*.

[166] Raymond Hendy Susanto, Shamil Chollampatt, and Liling Tan. 2020. Lexically Constrained Neural Machine Translation with Levenshtein Transformer. (7 2020), 3536–3543. https://doi.org/10.18653/V1/2020.ACL-MAIN.325

[167] Xiangru Tang, Arjun Nair, Borui Wang, Bingyao Wang, Jai Desai, Aaron Wade, Haoran Li, Asli Celikyilmaz, Yashar Mehdad, and Dragomir Radev. 2021. CONFIT: Toward Faithful Dialogue Summarization with Linguistically-Informed Contrastive Fine-tuning. *arXiv preprint arXiv:2112.08713* (2021).

[168] Avijit Thawani, Jay Pujara, Filip Ilievski, and Pedro Szekely. 2021. Representing Numbers in NLP: a Survey and a Vision. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 644–656.

[169] Craig Thomson and Ehud Reiter. 2020. A Gold Standard Methodology for Evaluating Accuracy in Data-To-Text Systems. In *Proceedings of the 13th International Conference on Natural Language Generation*. 158–168.

[170] James Thorne and Andreas Vlachos. 2019. Adversarial attacks against fact extraction and verification. *arXiv preprint arXiv:1903.05543* (2019).

[171] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 809–819. https://doi.org/10.18653/v1/N18-1074

[172] Ran Tian, Shashi Narayan, Thibault Sellam, and Ankur P. Parikh. 2020. Sticking to the Facts: Confident Decoding for Faithful Data-to-Text Generation. arXiv:1910.08684 [cs.CL]

[173] Van-Khanh Tran and Le-Minh Nguyen. 2017. Natural Language Generation for Spoken Dialogue System using RNN Encoder-Decoder Networks. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*. Association for Computational Linguistics, Vancouver, Canada, 442–451. https://doi.org/10.18653/v1/K17-1044

[174] Zhaopeng Tu, Yang Liu, Lifeng Shang, Xiaohua Liu, and Hang Li. 2016. Neural Machine Translation with Reconstruction. arXiv:1611.01874 [cs.CL]

[175] Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 2016. Modeling Coverage for Neural Machine Translation. arXiv:1601.04811 [cs.CL]

[176] Oleg Vasilyev, Vedant Dharnidharka, and John Bohannon. 2020. Fill in the BLANC: Human-free quality estimation of document summaries. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*. 11–20.

[177] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. 5998–6008. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[178] Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *ICML Deep Learning Workshop* (2015).

[179] Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (2020).

[180] Chaojun Wang and Rico Sennrich. 2020. On Exposure Bias, Hallucination and Domain Shift in Neural Machine Translation. (5 2020), 3544–3552. https://arxiv.org/abs/2005.03642v1

[181] Hongmin Wang. 2020. Revisiting challenges in data-to-text generation with fact grounding. *ACL* (2020).

[182] Peng Wang, Junyang Lin, An Yang, Chang Zhou, Yichang Zhang, Jingren Zhou, and Hongxia Yang. 2021. Sketch and Refine: Towards Faithful and Informative Table-to-Text Generation. *ACL* (2021).

[183] Xu Wang, Hainan Zhang, Shuai Zhao, Yanyan Zou, Hongshen Chen, Zhuoye Ding, Bo Cheng, and Yanyan Lan. 2021. FCM: A Fine-grained Comparison Model for Multi-turn Dialogue Reasoning. *EMNLP Findings* (2021).

[184] Zhenyi Wang, Xiaoyang Wang, Bang An, Dong Yu, and Changyou Chen. 2020. Towards Faithful Neural Table-to-Text Generation with Content-Matching Constraints. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 1072–1086.

[185] Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2019. Neural Text Generation With Unlikelihood Training. In *International Conference on Learning Representations*.

[186] 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*. Association for Computational Linguistics, Florence, Italy, 3731–3741. https://doi.org/10.18653/v1/P19-1363

[187] Tsung-Hsien Wen, Milica Gašić, Dongho Kim, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Stochastic Language Generation in Dialogue using Recurrent Neural Networks with Convolutional Sentence Reranking. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, Prague, Czech Republic, 275–284. https://doi.org/10.18653/v1/W15-4639

[188] Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Pei-hao Su, David Vandyke, and Steve J Young. 2015. Semantically Conditioned LSTM-based Natural Language Generation for Spoken Dialogue Systems. In *EMNLP*.

[189] Rongxiang Weng, Heng Yu, Xiangpeng Wei, and Weihua Luo. 2020. Towards enhancing faithfulness for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2675–2684.

[190] A Williams, N Nangia, and SR Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT), 1, 1112–1122.

[191] Sam Wiseman, Stuart M Shieber, and Alexander M Rush. 2017. Challenges in data-to-document generation. *EMNLP* (2017).

[192] Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. TransferTransfo: A Transfer Learning Approach for Neural Network Based Conversational Agents. *CoRR* abs/1901.08149 (2019). arXiv:1901.08149 http://arxiv.org/abs/1901.08149

[193] Chien-Sheng Wu, Richard Socher, and Caiming Xiong. 2019. Global-to-local Memory Pointer Networks for Task-Oriented Dialogue. In *International Conference on Learning Representations*. https://openreview.net/forum?id=ryxnHhRqFm

[194] Zeqiu Wu, Michel Galley, Chris Brockett, Yizhe Zhang, Xiang Gao, Chris Quirk, Rik Koncel-Kedziorski, Jianfeng Gao, Hannaneh Hajishirzi, Mari Ostendorf, et al. 2021. A Controllable Model of Grounded Response Generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 14085–14093.

[195] Yijun Xiao and William Yang Wang. 2021. On Hallucination and Predictive Uncertainty in Conditional Language Generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 2734–2744.

[196] Jing Xu, Arthur D. Szlam, and Jason Weston. 2021. Beyond Goldfish Memory: Long-Term Open-Domain Conversation. *ArXiv* abs/2107.07567 (2021).

[197] Weijia Xu and Marine Carpuat. 2020. EDITOR: an Edit-Based Transformer with Repositioning for Neural Machine Translation with Soft Lexical Constraints. *Transactions of the Association for Computational Linguistics* 9 (11 2020), 311–328. https://doi.org/10.1162/tacl_a_00368d3/2021.

[198] Weijia Xu and Marine Carpuat. 2021. Rule-based Morphological Inflection Improves Neural Terminology Translation. (9 2021), 5902–5914. https://doi.org/10.18653/v1/2021.emnlp-main.477

[199] Weijia Xu, Xing Niu, and Marine Carpuat. 2019. Differentiable Sampling with Flexible Reference Word Order for Neural Machine Translation. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference* 1 (4 2019), 2047–2053. https://doi.org/10.18653/v1/n19-1207

[200] Xinnuo Xu, Ondřej Dušek, Verena Rieser, and Ioannis Konstas. 2021. AGGGEN: Ordering and Aggregating while Generating. *roceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL2021)* (2021).

[201] Yan Xu, Etsuko Ishii, Samuel Cahyawijaya, Zihan Liu, Genta Indra Winata, Andrea Madotto, Dan Su, and Pascale Fung. 2021. Retrieval-free knowledge-grounded dialogue response generation with adapters. *arXiv preprint arXiv:2105.06232* (2021).

[202] Semih Yavuz, Abhinav Rastogi, Guan-Lin Chao, and Dilek Hakkani-Tur. 2019. DEEPCOPY: Grounded Response Generation with Hierarchical Pointer Networks. In *Proceedings of SIGdial*.

[203] Jun Yin, Xin Jiang, Zhengdong Lu, Lifeng Shang, Hang Li, and Xiaoming Li. 2016. Neural generative question answering. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. 2972–2978.

[204] Takuma Yoneda, Jeff Mitchell, Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. UCL Machine Reading Group: Four Factor Framework For Fact Finding (HexaF). In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*. Association for Computational Linguistics, Brussels, Belgium, 97–102. https://doi.org/10.18653/v1/W18-5515

[205] Tiezheng Yu, Zihan Liu, and Pascale Fung. 2021. AdaptSum: Towards Low-Resource Domain Adaptation for Abstractive Summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 5892–5904.

[206] Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big Bird: Transformers for Longer Sequences. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 17283–17297.

https://proceedings.neurips.cc/paper/2020/file/c8512d142a2d849725f31a9a7a361ab9-Paper.pdf

[207] Yury Zemlyanskiy and Fei Sha. 2018. Aiming to Know You Better Perhaps Makes Me a More Engaging Dialogue Partner. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*. Association for Computational Linguistics, Brussels, Belgium, 551–561. https://doi.org/10.18653/v1/K18-1053

[208] Chen Zhang, Grandee Lee, Luis Fernando D'Haro, and Haizhou Li. 2021. D-Score: Holistic Dialogue Evaluation Without Reference. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 2502–2516. https://doi.org/10.1109/TASLP.2021.3074012

[209] Hongguang Zhang, Jing Zhang, and Piotr Koniusz. 2019. Few-Shot Learning via Saliency-Guided Hallucination of Samples. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2770–2779. https://doi.org/10.1109/CVPR.2019.00288

[210] Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, and Yang Liu. 2021. Neural machine translation with explicit phrase alignment. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 1001–1010.

[211] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing Dialogue Agents: I have a dog, do you have pets too?. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2204–2213.

[212] Xikun Zhang, Deepak Ramachandran, Ian Tenney, Yanai Elazar, and Dan Roth. 2020. Do Language Embeddings capture Scales?. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*. 292–299.

[213] Yuhao Zhang, Derek Merck, Emily Tsai, Christopher D Manning, and Curtis Langlotz. 2020. Optimizing the Factual Correctness of a Summary: A Study of Summarizing Radiology Reports. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 5108–5120.

[214] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-Scale Generative Pre-training for Conversational Response Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, Online, 270–278. https://doi.org/10.18653/v1/2020.acl-demos.30

[215] Jing Zhao, Junwei Bao, Yifan Wang, Yongwei Zhou, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2021. RoR: Read-over-Read for Long Document Machine Reading Comprehension. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics, Punta Cana, Dominican Republic, 1862–1872. https://aclanthology.org/2021.findings-emnlp.160

[216] Zheng Zhao, Shay B Cohen, and Bonnie Webber. 2020. Reducing quantity hallucinations in abstractive summarization. *arXiv preprint arXiv:2009.13312* (2020).

[217] Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, et al. 2021. QMSum: A New Benchmark for Query-based Multi-domain Meeting Summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 5905–5921.

[218] Chunting Zhou, Xuezhe Ma, Di Wang, and Graham Neubig. 2019. Density Matching for Bilingual Word Embedding. In *NAACL*.

[219] Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Paco Guzman, Luke Zettlemoyer, and Marjan Ghazvininejad. 2021. Detecting Hallucinated Content in Conditional Neural Sequence Generation. arXiv:2011.02593 [cs.CL]

[220] Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. 2018. A Dataset for Document Grounded Conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 708–713.

[221] Pei Zhou, Karthik Gopalakrishnan, Behnam Hedayatnia, Seokhwan Kim, Jay Pujara, Xiang Ren, Yang Liu, and Dilek Hakkani-Tur. 2021. Think Before You Speak: Using Self-talk to Generate Implicit Commonsense Knowledge for Response Generation. *arXiv preprint arXiv:2110.08501* (2021).

[222] Chenguang Zhu, William Hinthorn, Ruochen Xu, Qingkai Zeng, Michael Zeng, Xuedong Huang, and Meng Jiang. 2021. Enhancing Factual Consistency of Abstractive Summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 718–733.