Hallucinations in LLMs and Resolving Them: A Holistic Approach

Rajarshi Biswas, Sourav Dutta and Dirk Werth
August-Wilhelm Scheer Institute, Uni-Campus D 5 1, 66123 Saarbrücken, Germany
{firstname.lastname}@aws-institut.de

Keywords: Natural Language Processing, Natural Language Generation, Generative AI.

Abstract: Generative artificial intelligence, in recent times, is producing tremendous interest across industry and academia leading to rapid growth. Developments in model architecture, training datasets and large scale computing enable the realization of impressive generative tasks in textual computing, computer vision etc. However, the generative processes suffer from various challenging artifacts that can generate confusion, risks or compromise the security. In this paper, we explore in detail the problem of inconsistent or hallucinogenic generation in natural language generation (NLG). We define the problem and survey the current techniques for detection, measurement and mitigation on five different tasks, which are, abstractive summarization, question answering, dialogue generation, machine translation and named entity recognition combined with information retrieval.

1 INTRODUCTION

The emergence of powerful large language models (LLMs) based on deep neural architectures, such as, Transformers, BERT, GPT is enabling generative artificial intelligence to scale impressive feats and attract unprecedented attention across the board. Natural language generation (NLG) is one of the primary yet challenging generative tasks in natural language processing and it is the focus of the LLMs. NLG comprises a wide variety of tasks, such as, coherent text, summary, dialogue generation, question answering, translation etc. that witnessed rapid growth in the last decade. However, the significant progress in NLG is accompanied with challenges such as lack of diversity in surface realization, loss of context and inconsistent or hallucinogenic generation.

In this work, we concentrate on analyzing hallucinogenic generation for five major downstream tasks in NLG, which are, abstractive summarization, question answering, dialogue generation, machine translation and named entity recognition combined with information retrieval. Hallucination is a form of degeneracy that demands attention from the research community. It is a serious issue with generative models in NLG and refers to situations in which the model generates inconsistent or nonsensical text that contradicts the source material, context or objective. It is important to study this phenomena since generative models like LLMs are being widely adopted in several critical services, e.g., health, banking in our society where hallucinations can severely limit the performance of the deployed models affecting the quality of service. Moreover, it can also jeopardise the safety of the applications leading to loss of trust and serious damage. For example, inconsistent response generation in a banking application can lead to an incorrect transaction causing loss of funds or more seriously a hallucinogenic response from a LLM in the health sector can lead to severe problems like wrong medication, drug overdose threatening the life of a patient.

As a consequence, efforts are being made in the community to understand the issue of hallucination or inconsistent generation in NLG. However, most of the studies are directed towards machine translation and text summarization. This leaves a gap in understanding the problem of hallucinations from a broader perspective that span different tasks. So, in this work we exhaustively survey the current works in this area across five different NLG tasks mentioned previously. We believe that studying the problem across different tasks would lead to deeper understanding, formation of an unified idea and help to identify global trends in hallucinogenic generation. Furthermore, we also discuss different ideas for mitigating inconsistent generation in the three different NLG tasks studied.

We organize the rest of the paper in a way, such that, section 2 describes the different variants and contributing factors for hallucination in NLG. In sections 3, 4, 5, 6, 7 we survey the current efforts in un-

Biswas, R., Dutta, S. and Werth, D.
Hallucinations in LLMs and Resolving Them: A Holistic Approach.
DOI: 10.5220/0013094500003890
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 17th International Conference on Agents and Artificial Intelligence (ICAART 2025) - Volume 3*, pages 104-115
ISBN: 978-989-758-737-5; ISSN: 2184-433X
Proceedings Copyright © 2025 by SCITEPRESS – Science and Technology Publications, Lda.

derstanding hallucination in abstractive summarization, question answering, dialogue generation, machine translation and named entity recognition along with information retrieval. Under section 8 we discuss different ways of resolving the problem of hallucinations considering holistic as well as task specific measures. Following this in section 9 we discuss potential future areas that can be researched for managing hallucinations in a better way. Finally, we summarize our findings in section 10.

2 HALLUCINATION: VARIANTS AND CONTRIBUTING FACTORS

In this section, we briefly describe the different variants of hallucination and the different factors contributing to it. In the context of natural language processing, **Hallucination** is defined as automatically generated content that is nonsensical and unfaithful compared to the source content (Filippova, 2020; Maynez et al., 2020; Parikh et al., 2020; Zhou et al., 2020). Depending on the tasks, prior research work divides it into two categories, **Intrinsic** and **Extrinsic** hallucinations (Dziri et al., 2021; Huang et al., 2021; Maynez et al., 2020). In the first category, the generated output contradicts the input source allowing it to be classified as erroneous generation. Extrinsic hallucination refers to generations that cannot be verified from source content thus it may not be incorrect every time. Nonetheless, it is still problematic and poses a safety risk. The primary **factors contributing to hallucination** in NLG are data sources and model training choices. On the data front factors such as heuristic data collection (Lebret et al., 2016; Wiseman et al., 2017; Parikh et al., 2020; Wang, 2020) or tasks that require diversity in the generations, e.g., open-domain dialogue generation in a subjective tone (Rashkin et al., 2021) leads to source-output divergence. This divergence is one of the key contributing factors behind hallucination. Model training related factors causing hallucination could be faulty representation learning, wrong decoding, exposure bias, parametric-knowledge bias etc. For instance, an encoder learning wrong correlations (Li et al., 2018; Feng et al., 2020) or having a faulty understanding (Parikh et al., 2020) can lead to inconsistent generations. Similarly, focusing on the wrong part of the encoded information or efforts directed at improving diversity during decoding can result in hallucination (Tian et al., 2019). The problem of exposure bias (Bengio et al., 2015; Ranzato et al., 2015), that is, disparity in decoding during training and inference also leads to inconsistency. This is due to MLE optimization using ground-truth prefixes for next token prediction in contrast to using self generated history during inference (He et al., 2021).

3 ABSTRACTIVE SUMMARIZATION

Hallucination: In NLP, abstractive summarization refers to the task of generating a short, concise summary from the source text such that it contains all the relevant details in the source (Yu et al., 2021). Even though neural approaches have obtained much success in this task, recent studies find that neural techniques generate inconsistent or hallucinogenic content (Falke et al., 2019; Maynez et al., 2020). Moreover, it is observed that generated summaries with large amount of inconsistencies can still obtain very high ROUGE scores. These findings underscore the importance of studying the problem of hallucinations in this task.

Measurement: The degree of inconsistency in the generated summaries are measured using metrics that are mostly model based. These can be categorized into unsupervised and semi-supervised metrics. The unsupervised metrics can be further classified into information extraction based, natural language inference based and question-answering based respectively. **Information extraction based methods** extract details in the form of relation tuples from both the source and generated summary for verification of factual accuracy. In a similar light, **question-answering based metrics** measure factual accuracy between source and output through generation of pertinent questions that are assumed to produce similar answers. In general, these metrics follow three steps, which are, question generation from the generated output, extracting answers from the source & output, and scoring the correctness of answers obtained from the source and the output. In contrast **natural language inference** metrics assume that there is a ground-truth for a faithful summary.

Resolution: Practitioners in abstractive summarization use various techniques for coping this issue. For example, **graph neural networks** are used in (Zhu et al., 2021) for encoding facts from the source text and further integration of **reward functions** in (Huang et al., 2020) for better understanding interactions between entities in the source. **External knowledge embedding** obtained from embedding facts from wikipedia is also used in (Gunel et al., 2020) for improving factual consistency. Techniques like (Aralikatte et al., 2021) propose **focus-attention mechanism** for making decoders generate tokens that are related to the facts or topic of the source. Keeping with attention-based methods, the work in (Cao et al., 2018) uses a **dual attention sequence-to-sequence framework** for ensuring that generated summaries take into account the source text and the facts extracted from them. **Contrastive learning** technique is used in (Cao and Wang, 2021) for enabling the models to distinguish between positive ground-truth summaries and automatically generated negative summaries containing factual inconsistencies or hallucinations. Apart from these post-processing is also employed in the works to get rid off the inconsistent facts in the generated summaries.

4 QUESTION-ANSWERING

Hallucination: Generative question answering is gaining prominence with the growing success of generative artificial intelligence. It is more powerful and effective compared to first generation question-answering systems that merely tried to find facts in the source text that support the questions. The objective of generative question answering is to frame more detailed and complete answers that may require gathering information from all over the source. As a result, sometimes the system needs to consult multiple source documents as a single document may not contain all the information needed for framing a definitive answer. However, this process can induce the adverse side-effect hallucinations since some of these documents may contain extraneous or contradictory information. The closest form of a definition of Hallucination in generative question answering is semantic drift (Li et al., 2021). It shows how a generated answer drifts away from the correct answer during generation. Apart from this the majority of the works in this area leverage human evaluation for measuring factual correctness of the generated answers as a measure of inconsistency.

Measurement: Hallucination in generative question answering is measured using the metric **Semantic Overlap** (Sellam et al., 2020). It is a BERT-based metric that correlates with human judgment. **Factual correctness** is also employed for measuring consistency (Zhang et al., 2020a) between generated text and source document using information extraction. **Automatic question answer based metric** is proposed (Durmus et al., 2020; Wang et al., 2020) for measuring consistency in generated summaries. In this approach, first question-answer pairs are created using a question generation model from the generated summary. Subsequently, a model is used for extracting answers from the source document for the questions generated in the previous step. If the answers don't match then the generated summary is regarded as unfaithful. This technique is also used in measuring hallucination in generative question answering. Apart from these metrics, **human evaluation** is frequently used in this field for measuring the consistency or faithfulness of the generated answers. Human evaluation is often also used to complement automatic N-gram overlap metrics, such as, BLEU, ROUGE, METEOR, as these correlate poorly with human judgments.

Resolution: Techniques used in generative question answering for resolving hallucination concentrate on leveraging external knowledge bases and information resources for improving the factual correctness or faithfulness of the generated answers. Another approach (Bi et al., 2019) generates answers by accumulating **information from multiple sources**, such as, knowledge-bases, passages, vocabulary, questions etc. Neural model (Yin et al., 2016) is used for **generating answers to factoid questions** using information from knowledge-base. More recent approaches (Fan et al., 2019) create **individual knowledge graph** for every question for condensing information while reducing redundancy for tackling hallucination. Another method (Li et al., 2021) extracts **rationale for an answer** in the encoding stage and biases the decoder to generate the answer using the rationale and the actual input. For reducing hallucination in the answers, the authors in (Krishna et al., 2021) propose a **sparse attention-based transformer model** as the answer generator for effectively handling the retrieved documents. It models long-range dependence employing local attention and mini-batch K-means clustering. Similarly for mitigating hallucination in (Su et al., 2022), a new framework is proposed that jointly models **answer-generation with machine reading**. The generation model is complemented by the machine reading module. It provides salient answer related information to the generation model to improve faithfulness of the generated answer.

5 DIALOGUE GENERATION

Hallucination: Dialogue generation is probably the most widely adopted generation tasks in natural language processing with wide ranging applications like chatbots, voice-assistants etc. It can be broadly categorized into task specific and open domain dialogue generation. In the first category, we expect responses to contain specific information while in the second type often an engaging response is desired without too much repetition from the conversational history for relatively long conversation. Due to this nature the tolerance for hallucination is higher in this task compared to other generation tasks. Hallucination in dialogue generation is considered intrinsic if certain specific information is absent or misrepresented in the generated response. Whereas if the generated conversation is not firmly grounded in hard facts and is difficult to be explicitly verified using knowledge bases or conversational history then it is termed as extrinsic hallucination. In our work, we discuss the problems related to open domain dialogue generation as it more relevant to the modern dialogue systems that are developed incorporating state-of-the-art LLMs trained on huge amounts of training data. In open domain dialogue systems there can be broadly two sources of hallucinations. First, responses that contradict previous responses from the same system leading to inconsistency (Li et al., 2020; Welleck et al., 2019; Zhang et al., 2021a), incoherence (Beyer et al., 2021; Dziri et al., 2019) termed as self-inconsistency. Secondly, when the systems generates responses that are inconsistent with regards to an external source, e.g., factually incorrect responses then it is termed as external inconsistency (Mielke et al., 2022; Roller et al., 2021). Another factor influencing inconsistency in open domain dialogue generation is the lack of consistency in the Persona/Character assumed by the dialogue system. This often leads to contradictions and in turn to hallucinations. As a result, there is research (Hancock et al., 2019; Mazaré et al., 2018; Yavuz et al., 2019; Zhang et al., 2020b) to develop systems that are persona consistent with the help of suitable datasets (Dinan et al., 2019a; Zhang et al., 2018). Additionally, there are also works in open domain dialogue generation that use external knowledge bases and graphs for generating informative responses (Dinan et al., 2019b; Zhou et al., 2018). Hallucination in such systems is treated as factual inconsistency and has received equal amount of attention from the dialogue generation community (Dziri et al., 2021; Rashkin et al., 2021; Santhanam et al., 2021; Shuster et al., 2021).

Measurement: Evaluation of hallucination in open domain dialogue generation is still an open problem as there is no standard metric for measuring it. Dialogue systems, such as, chat-bots are often evaluated using factual correctness or consistency. Some automated metrics used for measurement are **Knowledge F1**, **Rare F1** (Shuster et al., 2021) both of which are based on statistics while others are model based techniques. Knowledge F1 utilizes ground-truth datasets where knowledge is labeled. This refers to gold standard knowledge sentences to which a person referred for conversation during dataset collection. Knowledge F1 measures the overlap between the generated and gold knowledge sentences. This metric tries to measure if the generated responses are able to capture the available knowledge and thus if they make sense. Rare F1 only considers the infrequent words in the dataset for computing the F1 metric. This is done to negate the influence of common uni-grams. However, overlap based metrics cannot provide comprehensive evaluation since the same semantic meaning could be represented in a wide variety of surface realizations. For addressing this different model based techniques have been proposed for measuring consistency. For example, using **natural language inference (NLI)** (Dziri et al., 2019; Welleck et al., 2019), **learnable evaluation metrics** (Zhang et al., 2021b) or **use of an additional test for measuring coherence** (Beyer et al., 2021). These methods offer more flexibility and can support generations with different surface realizations.

Resolution: The problem of hallucination in open domain dialogue generation can be mitigated using different techniques. One of the ways is by introducing extra information in the data. The authors in (Shen et al., 2021) propose a measurement based on features of dialogue quality which can be used to remove samples from the training set that get a lower score on this measurement. In turn this can improve performance in terms of self-consistency. Retrieval is used to augment dialogue generation approaches, such as, Knowledge Grounded Dialogue where is it performs knowledge selection and helps to reduce hallucinations substantially (Shuster et al., 2021). Control codes concatenated with dialogue inputs is proposed in (Rashkin et al., 2021) for reducing hallucinations. It makes the model more aware of how the generations rely on evidence based in knowledge. Improved dialogue modeling techniques have also been studied for reducing hallucinations during generation, e.g., the use of inductive attention in dialogue models based on the transformer architecture (Wu et al., 2021).

6 MACHINE TRANSLATION

Hallucination: Machine translation (MT) refers to the automatic conversion of text from one language into another, aiming for both grammatical accuracy and semantic fidelity (Bahdanau, 2014). While neural machine translation (NMT) models have dramatically improved translation quality, particularly with the advent of transformer-based architectures (Vaswani et al., 2017), they are still prone to generating hallucinations. These hallucinations occur when the system introduces information that is not present in the source text, or mistranslates critical content, leading to outputs that may seem fluent but are semantically incorrect or inconsistent (Raunak et al., 2021; Müller et al., 2020). These errors are particularly prevalent in low-resource language pairs and in cases where the model overfits to patterns in the training data. Hallucinations in machine translation can severely impact the reliability of translations, especially in critical domains such as legal, medical, or technical fields, where accuracy is paramount (Raunak et al., 2021).

Measurement: Evaluating hallucinations in machine translation poses a unique challenge, as traditional metrics like **BLEU** (Papineni et al., 2002) or **METEOR** (Banerjee and Lavie, 2005), which compare the machine output to reference translations, may not effectively capture the degree of hallucination. Recent studies have proposed new approaches to better measure hallucinations, including both model-based and human evaluation metrics. One common approach involves using **adequacy-based human evaluation**, where human annotators judge how well the translation aligns with the source content (Specia et al., 2011). For automated methods, **source-reference alignment techniques** can identify mistranslations or extraneous information by comparing source and target alignments to ensure fidelity (He et al., 2016). This focuses on improving translation quality by aligning source and target text, helping to detect hallucinations or extraneous information. This approach ensures better fidelity in translations by refining how models maintain consistency between the input and output sequences. **NLI (Natural Language Inference) model-based metrics** (Zhou et al., 2021) mainly aimed to fact-check and align generated text, are able to detect hallucinations in generated outputs. Such methods compare the translated content (hypothesis) against the source (premise) for contradictions or factual inaccuracies. Another approach uses **confidence-based filtering**, where low-confidence outputs from the translation model are flagged as potentially hallucinatory (Tu et al., 2017).

Resolution: Addressing hallucinations in machine translation involves both improving the underlying model architecture and leveraging external resources. One promising approach is **data augmentation**, particularly for low-resource languages, which can help mitigate hallucinations caused by insufficient training data (Sennrich et al., 2016). In addition, **back-translation**, where the model translates target language sentences back into the source language and compares them to the original text, has been used to reduce inconsistencies (Edunov et al., 2018). Other efforts focus on improving the attention mechanisms within transformers. For example, **coverage mechanisms** have been employed to ensure that every part of the source sentence is attended to during translation, reducing the likelihood that the model will *"invent"* content not present in the source (Tu et al., 2016). Incorporating **external knowledge bases** has also been explored, particularly integrating knowledge embeddings into NLP tasks like translation helps maintain factual consistency, reducing the risk of hallucinations, especially in technical or specialized content (Wang et al., 2021). Moreover, the use of **reinforcement learning** for sequence prediction tasks, including NMT, shows how reward functions can be tailored to encourage factual accuracy, reducing issues like hallucination during translation (Bahdanau et al., 2022). Finally, **post-editing techniques**, where human editors review and correct translations, are often employed in high-stakes scenarios to ensure final output quality, especially when dealing with critical content (Toral et al., 2018).

7 NAMED ENTITY RECOGNITION AND INFORMATION RETRIEVAL

Hallucination: Named Entity Recognition (NER) is a fundamental NLP task aimed at identifying and classifying proper nouns such as people, organizations, and locations within a text (Lample et al., 2016). Despite significant progress with neural models, these systems can still exhibit hallucinations, where entities are misclassified or incorrectly generated. For example, models might mistakenly recognize a non-existent entity or mislabel a correct entity due to insufficient context or model limitations (Su et al., 2024). This misclassification can impact applications relying on accurate entity identification, such as information extraction and semantic search. In Information Retrieval (IR), the objective is to retrieve documents or data that are relevant to a user's query (Schütze et al., 2008). Although neural IR models have improved the relevance and ranking of retrieved results, they can occasionally retrieve documents that are irrelevant or hallucinated, meaning the retrieved results do not genuinely align with the user's query intent (Nogueira and Cho, 2019; James and Kannan, 2017). These hallucinated results can stem from overfitting on training data or from inadequacies in the query-document matching process.

Measurement: To measure hallucinations in NER, various evaluation metrics are employed. Precision, recall, and F1-score are commonly used to compare the predicted entities against a gold standard annotated dataset. Precision measures the proportion of correctly identified entities out of all entities identified by the model, recall measures the proportion of correctly identified entities out of all entities that should have been identified, and F1-score provides a balance between precision and recall. Unsupervised metrics also play a role, such as entity linking, where entities recognized by the model are matched against external knowledge bases to verify their correctness. **Cross-document consistency checks** can further identify discrepancies by ensuring that entities are consistently recognized across multiple documents (Jiang et al., 2016). For IR, effectiveness is measured through metrics such as **Precision@K**, **Recall@K**, and **Mean Reciprocal Rank (MRR)**. Precision@K measures the proportion of relevant documents among the top K retrieved documents, while Recall@K assesses the proportion of relevant documents retrieved within the top K results. MRR evaluates the rank of the first relevant document in the list. Additionally, query-document relevance scoring, which involves assessing the alignment between the query and the retrieved documents, and external validation against curated datasets are used to gauge retrieval accuracy and address issues of hallucination (Schütze et al., 2008; Nogueira and Cho, 2019).
Resolution: Addressing hallucinations in NER involves several advanced techniques. **Contextual embeddings** from models such as BERT (Devlin et al., 2019) capture richer semantic information by providing context-dependent representations of words. This approach improves the accuracy of entity recognition by understanding the context in which entities appear. **Multi-task learning**, which involves training models on related tasks simultaneously, helps enhance entity recognition by leveraging additional sources of information (McCann et al., 2017). Integrating external knowledge sources like **knowledge graphs** can also reduce hallucinations by grounding the entity recognition process in real-world data (He et al., 2020). In IR, techniques to mitigate hallucination include employing advanced retrieval architectures such as dense retrievers and cross-encoder models. **Dense retrievers** use dense vector representations for query-document matching, which improves the relevance ranking of retrieved documents (Nogueira and Cho, 2019). Cross-encoder models, which jointly encode the query and documents, further refine retrieval by capturing complex relationships between them. Additionally, incorporating **user feedback** and techniques like **query expansion**, where additional terms or context are added to the query, helps refine retrieval results and address issues of hallucination (Azad and Deepak, 2019).

8 APPROACHES TO RESOLVING HALLUCINATIONS

The motivation behind this paper stems from the growing reliance on Large Language Models (LLMs) across a wide range of NLP tasks. While these models have demonstrated remarkable advancements, they also introduce a critical challenge: hallucinations. Across tasks like abstractive summarization, question answering, dialog generation, machine translation, NER, and information retrieval, hallucinations manifest in various forms, from generating factual inaccuracies to retrieving irrelevant or fabricated information. Despite significant progress in mitigating these issues, hallucination remains a pervasive problem that compromises the reliability of LLMs in real-world applications (Ji et al., 2023). The primary motivation for this paper is the need for a comprehensive, cross-task analysis of hallucinations in LLMs. While hallucinations in specific tasks such as summarization or machine translation have been studied in isolation (Raunak et al., 2021), there has been little effort to systematically explore hallucinations across multiple NLP tasks, each with its unique characteristics and challenges. This paper aims to fill that gap by providing a detailed investigation into the nature of hallucinations in five distinct tasks, as well as outlining the current methods to detect and resolve them. Our contribution is twofold: (1) a consolidated review of hallucination across different NLP tasks, and (2) proposing task-agnostic and task-specific approaches to resolve hallucinations, thereby providing a framework for future research.

8.1 Holistic Approach

While techniques that are task-specific, such as external knowledge integration (Zhu et al., 2021) or using better reward mechanisms (Chen et al., 2023), have shown promise, we propose a more holistic approach that could benefit all tasks:
Improving Model Interpretability: A crucial challenge is the black-box nature of LLMs, which makes hallucinations difficult to predict or prevent. Implementing interpretability mechanisms like attention visualization or rule-based model auditing can help identify when and why hallucinations occur (Belinkov and Glass, 2019). Models like BERT, GPT,

and their variants could be enhanced with transparent architectures that allow for more insight into their decision-making process, especially in tasks prone to hallucination, like dialog generation and summarization (Ribeiro et al., 2016).

Task-Agnostic Regularization: Regularization techniques, like fact-checking or constraint-based generation, should be applied consistently across tasks. For example, incorporating external knowledge bases, such as Wikipedia or structured databases, can help ground generated outputs in factual information, thereby reducing hallucination in both generative (summarization, QA) and retrieval-based tasks (IR, NER) (Petroni et al., 2019). This approach prevents the model from generating content that strays too far from verifiable truth, creating a safeguard against fabricated information.

Adaptive Fine-Tuning for Specific Tasks: Although LLMs are designed to generalize across tasks, fine-tuning them on domain-specific data can significantly reduce hallucinations. In tasks like machine translation and information retrieval, training models on specialized datasets and including domain-relevant entities can lead to more accurate and contextually appropriate outputs (Sun et al., 2023). This reduces the likelihood of hallucinating irrelevant or incorrect information, particularly when the task demands high precision.

Evaluation and Feedback Mechanisms: One consistent theme across tasks is the need for robust evaluation metrics. ROUGE, BLEU, and MRR are often insufficient to detect hallucinations because they focus on fluency and surface-level similarities (Honovich et al., 2022). We suggest augmenting these metrics with fact-based or entity-level verification mechanisms. For instance, in question answering, automatic fact-checking systems could be integrated to score models on factual consistency, while in summarization and translation, knowledge graphs could be employed to cross-validate entity relationships (Cao et al., 2020).

8.2 Task-Specific Considerations

Certain tasks, due to their inherent complexity and the nature of the data they process, require tailored solutions to effectively mitigate hallucinations. These solutions address the unique challenges of each task, allowing models to generate more accurate and contextually appropriate outputs.

Named Entity Recognition (NER): NER systems are prone to hallucinations when they mislabel entities or identify non-existent ones, especially in domains where new entities frequently emerge, such as healthcare, finance, or geopolitics. Grounding NER models in dynamic, real-world knowledge bases, such as Wikidata or domain-specific databases, can help ensure that entity identification remains accurate and up-to-date (Hu et al., 2022). By continuously updating the knowledge base and training the model on evolving data, hallucinations can be reduced as the system remains aware of the latest entities and their relationships. Furthermore, integrating context-aware mechanisms, where entity recognition adapts based on sentence-level or document-level context, can improve accuracy and minimize misidentifications, particularly in ambiguous scenarios where multiple entities are involved.

Machine Translation: Machine translation systems are susceptible to hallucinations, particularly when translating between languages with significant structural differences or when translating low-resource languages. Ensuring linguistic consistency across languages is crucial for reducing hallucinations. One approach is incorporating post-editing frameworks where human translators verify and correct machine-generated translations, thereby maintaining translation quality and factual accuracy. In addition, contrastive learning techniques, which explicitly train the model to recognize and avoid incorrect or out-of-context translations, can help minimize semantic drift—the phenomenon where the translation strays from the intended meaning (Raunak et al., 2021). This can be particularly useful when translating specialized texts, such as legal or medical documents, where precision is paramount.

Dialog Generation: Hallucinations in dialog generation often result in models producing off-topic, incoherent, or factually incorrect responses. One of the primary challenges is maintaining the consistency and coherence of conversations over multiple turns. Integrating persona mechanisms—where the model is conditioned on a set of attributes or knowledge about the user—can help ground responses in the user's context, reducing the likelihood of irrelevant or inconsistent replies (Zhang et al., 2020b). Additionally, context memory mechanisms, which allow the model to retain and reference information from earlier in the conversation, can ensure that subsequent responses stay coherent and relevant. By maintaining a memory of the dialog history, models can avoid introducing new, unrelated information that could lead to hallucination.

9 FUTURE WORK

Moving forward, we envision research focusing on hybrid models that combine symbolic reasoning with deep learning. This could address hallucinations by introducing structured knowledge into the generative process (Chen et al., 2020). Additionally, cross-lingual hallucination detection in translation tasks and further exploration into self-supervised fact-checking methods for QA and summarization will likely enhance model robustness. Ultimately, addressing hallucinations requires a concerted effort that combines advances in model architectures, training strategies, and evaluation techniques. Our work highlights the importance of a unified approach to tackling hallucinations in LLMs, with the aim of developing models that are not only powerful but also reliable and trustworthy (Schick and Schütze, 2021).

10 CONCLUSION

In this paper, we explored the challenge of hallucinations in Large Language Models (LLMs) across five key NLP tasks: abstractive summarization, question answering, dialog generation, machine translation, named entity recognition, and information retrieval. Despite advances in these tasks, hallucinations remain a persistent problem, undermining model reliability. We provided a comprehensive review of task-specific manifestations, metrics, and methods to address hallucinations, and proposed a unified framework that emphasizes interpretability, regularization, and fine-tuning. Moving forward, addressing hallucinations will be crucial for improving the trustworthiness and applicability of LLMs in real-world scenarios.

ACKNOWLEDGEMENTS

This research is funded in part by the "Bundesministerium für Wirtschft und Klimaschutz" within the project "MERLOT" which was funded under the project reference 68GX21008K.

REFERENCES

Aralikatte, R., Narayan, S., Maynez, J., Rothe, S., and McDonald, R. (2021). Focus attention: Promoting faithfulness and diversity in summarization. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *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)*, pages 6078–6095. Association for Computational Linguistics.

Azad, H. K. and Deepak, A. (2019). Query expansion techniques for information retrieval: a survey. *Information Processing & Management*, 56(5):1698–1735.

Bahdanau, D. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bahdanau, D., Brakel, P., Xu, K., Goyal, A., Lowe, R., Pineau, J., Courville, A., and Bengio, Y. (2022). An actor-critic algorithm for sequence prediction. In *International Conference on Learning Representations*.

Banerjee, S. and Lavie, A. (2005). Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Belinkov, Y. and Glass, J. (2019). Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Bengio, S., Vinyals, O., Jaitly, N., and Shazeer, N. (2015). Scheduled sampling for sequence prediction with recurrent neural networks. *Advances in neural information processing systems*, 28.

Beyer, A., Loáiciga, S., and Schlangen, D. (2021). Is incoherence surprising? targeted evaluation of coherence prediction from language models. In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y., editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4164–4173. Association for Computational Linguistics.

Bi, B., Wu, C., Yan, M., Wang, W., Xia, J., and Li, C. (2019). Incorporating external knowledge into machine reading for generative question answering. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2521–2530, Hong Kong, China. Association for Computational Linguistics.

Cao, M., Dong, Y., Wu, J., and Cheung, J. C. K. (2020). Factual error correction for abstractive summarization models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6251–6258.

Cao, S. and Wang, L. (2021). CLIFF: Contrastive learning for improving faithfulness and factuality in abstractive summarization. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t., editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6633–6649. Association for Computational Linguistics.

Cao, Z., Wei, F., Li, W., and Li, S. (2018). Faithful to the original: fact-aware neural abstractive summarization.

SCIENCE AND TECHNOLOGY PUBLICATIONS

In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press.

Chen, T., Wang, X., Yue, T., Bai, X., Le, C. X., and Wang, W. (2023). Enhancing abstractive summarization with extracted knowledge graphs and multi-source transformers. *Applied Sciences*, 13(13):7753.

Chen, W., Su, Y., Yan, X., and Wang, W. Y. (2020). Kgpt: Knowledge-grounded pre-training for data-to-text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8635–8648.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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)*, pages 4171–4186.

Dinan, E., Logacheva, V., Malykh, V., Miller, A. H., Shuster, K., Urbanek, J., Kiela, D., Szlam, A., Serban, I. V., Lowe, R., Prabhumoye, S., Black, A. W., Rudnicky, A. I., Williams, J. D., Pineau, J., Burtsev, M., and Weston, J. (2019a). The second conversational intelligence challenge (convai2). *The Springer Series on Challenges in Machine Learning*, pages 187–208.

Dinan, E., Roller, S., Shuster, K., Fan, A., Auli, M., and Weston, J. (2019b). Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

Durmus, E., He, H., and Diab, M. (2020). FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070. Association for Computational Linguistics.

Dziri, N., Kamalloo, E., Mathewson, K., and Zaiane, O. (2019). Evaluating coherence in dialogue systems using entailment. In Burstein, J., Doran, C., and Solorio, T., editors, *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)*, pages 3806–3812. Association for Computational Linguistics.

Dziri, N., Madotto, A., Zaïane, O., and Bose, A. J. (2021). Neural path hunter: Reducing hallucination in dialogue systems via path grounding. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t., editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2197–2214. Association for Computational Linguistics.

Edunov, S., Ott, M., Auli, M., and Grangier, D. (2018). Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500.

Falke, T., Ribeiro, L. F., Utama, P. A., Dagan, I., and Gurevych, I. (2019). Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 2214–2220.

Fan, A., Gardent, C., Braud, C., and Bordes, A. (2019). Using local knowledge graph construction to scale Seq2Seq models to multi-document inputs. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4186–4196, Hong Kong, China. Association for Computational Linguistics.

Feng, Y., Xie, W., Gu, S., Shao, C., Zhang, W., Yang, Z., and Yu, D. (2020). Modeling fluency and faithfulness for diverse neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 59–66.

Filippova, K. (2020). Controlled hallucinations: Learning to generate faithfully from noisy data. *arXiv preprint arXiv:2010.05873*.

Gunel, B., Zhu, C., Zeng, M., and Huang, X. (2020). Mind the facts: Knowledge-boosted coherent abstractive text summarization. *ArXiv*, abs/2006.15435.

Hancock, B., Bordes, A., Mazare, P.-E., and Weston, J. (2019). Learning from dialogue after deployment: Feed yourself, chatbot! In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3667–3684. Association for Computational Linguistics.

He, D., Xia, Y., Qin, T., Wang, L., Yu, N., Liu, T.-Y., and Ma, W.-Y. (2016). Dual learning for machine translation. *Advances in neural information processing systems*, 29.

He, Q., Wu, L., Yin, Y., and Cai, H. (2020). Knowledge-graph augmented word representations for named entity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7919–7926.

He, T., Zhang, J., Zhou, Z., and Glass, J. (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*, pages 5087–5102. Association for Computational Linguistics.

Honovich, O., Aharoni, R., Herzig, J., Taitelbaum, H., Kukliansy, D., Cohen, V., Scialom, T., Szpektor, I., Hassidim, A., and Matias, Y. (2022). True: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3905–3920.

Hu, W., He, L., Ma, H., Wang, K., and Xiao, J. (2022). Kgner: Improving chinese named entity recognition by bert infused with the knowledge graph. *Applied Sciences*, 12(15):7702.

Huang, L., Wu, L., and Wang, L. (2020). Knowledge graph-augmented abstractive summarization with semantic-driven cloze reward. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5094–5107. Association for Computational Linguistics.

Huang, Y., Feng, X., Feng, X., and Qin, B. (2021). The factual inconsistency problem in abstractive text summarization: A survey. *arXiv preprint arXiv:2104.14839*.

James, N. T. and Kannan, R. (2017). A survey on information retrieval models, techniques and applications. *International Journals of Advanced Research in Computer Science and Software Engineering ISSN*.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Jiang, R., Banchs, R. E., and Li, H. (2016). Evaluating and combining name entity recognition systems. In *Proceedings of the sixth named entity workshop*, pages 21–27.

Krishna, K., Roy, A., and Iyyer, M. (2021). Hurdles to progress in long-form question answering. In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y., editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4940–4957. Association for Computational Linguistics.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016). Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270.

Lebret, R., Grangier, D., and Auli, M. (2016). Neural text generation from structured data with application to the biography domain. *arXiv preprint arXiv:1603.07771*.

Li, C., Bi, B., Yan, M., Wang, W., and Huang, S. (2021). Addressing semantic drift in generative question answering with auxiliary extraction. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *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)*, pages 942–947. Association for Computational Linguistics.

Li, H., Zhu, J., Zhang, J., and Zong, C. (2018). Ensure the correctness of the summary: Incorporate entailment knowledge into abstractive sentence summarization. In *Proceedings of the 27th international conference on computational linguistics*, pages 1430–1441.

Li, M., Roller, S., Kulikov, I., Welleck, S., Boureau, Y.-L., Cho, K., and Weston, J. (2020). Don't say that! making inconsistent dialogue unlikely with unlikelihood training. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4715–4728. Association for Computational Linguistics.

Maynez, J., Narayan, S., Bohnet, B., and McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919. Association for Computational Linguistics.

Mazaré, P.-E., Humeau, S., Raison, M., and Bordes, A. (2018). Training millions of personalized dialogue agents. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J., editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2775–2779. Association for Computational Linguistics.

McCann, B., Bradbury, J., Xiong, C., and Socher, R. (2017). Learned in translation: Contextualized word vectors. *Advances in neural information processing systems*, 30.

Mielke, S. J., Szlam, A., Dinan, E., and Boureau, Y.-L. (2022). Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872.

Müller, M., Gonzales, A. R., and Sennrich, R. (2020). Domain robustness in neural machine translation. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 151–164.

Nogueira, R. and Cho, K. (2019). Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085*.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Parikh, A., Wang, X., Gehrmann, S., Faruqui, M., Dhingra, B., Yang, D., and Das, D. (2020). ToTTo: A controlled table-to-text generation dataset. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1173–1186. Association for Computational Linguistics.

Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., and Miller, A. (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)*, pages 2463–2473.

Ranzato, M., Chopra, S., Auli, M., and Zaremba, W. (2015). Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*.

Rashkin, H., Reitter, D., Tomar, G. S., and Das, D. (2021). Increasing faithfulness in knowledge-grounded dialogue with controllable features. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *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)*, pages 704–718. Association for Computational Linguistics.

Raunak, V., Menezes, A., and Junczys-Dowmunt, M. (2021). The curious case of hallucinations in neural machine translation. In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y., editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1172–1183, Online. Association for Computational Linguistics.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Roller, S., Dinan, E., Goyal, N., Ju, D., Williamson, M., Liu, Y., Xu, J., Ott, M., Smith, E. M., Boureau, Y.-L., and Weston, J. (2021). Recipes for building an open-domain chatbot. In Merlo, P., Tiedemann, J., and Tsarfaty, R., editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325. Association for Computational Linguistics.

Santhanam, S., Hedayatnia, B., Gella, S., Padmakumar, A., Kim, S., Liu, Y., and Hakkani-Tür, D. Z. (2021). Rome was built in 1776: A case study on factual correctness in knowledge-grounded response generation. *ArXiv*, abs/2110.05456.

Schick, T. and Schütze, H. (2021). Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269.

Schütze, H., Manning, C. D., and Raghavan, P. (2008). *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge.

Sellam, T., Das, D., and Parikh, A. (2020). BLEURT: Learning robust metrics for text generation. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016). Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96.

Shen, L., Zhan, H., Shen, X., Chen, H., Zhao, X., and Zhu, X. (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*, page 1598–1608. Association for Computing Machinery.

Shuster, K., Poff, S., Chen, M., Kiela, D., and Weston, J. (2021). Retrieval augmentation reduces hallucination in conversation. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t., editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803. Association for Computational Linguistics.

Specia, L., Hajlaoui, N., Hallett, C., and Aziz, W. (2011). Predicting machine translation adequacy. In *Proceedings of Machine Translation Summit XIII: Papers*.

Su, D., Li, X., Zhang, J., Shang, L., Jiang, X., Liu, Q., and Fung, P. (2022). Read before generate! faithful long form question answering with machine reading. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 744–756, Dublin, Ireland. Association for Computational Linguistics.

Su, W., Tang, Y., Ai, Q., Wang, C., Wu, Z., and Liu, Y. (2024). Mitigating entity-level hallucination in large language models. *arXiv preprint arXiv:2407.09417*.

Sun, W., Shi, Z., Gao, S., Ren, P., de Rijke, M., and Ren, Z. (2023). Contrastive learning reduces hallucination in conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13618–13626.

Tian, R., Narayan, S., Sellam, T., and Parikh, A. P. (2019). Sticking to the facts: Confident decoding for faithful data-to-text generation. *arXiv preprint arXiv:1910.08684*.

Toral, A., Wieling, M., and Way, A. (2018). Post-editing effort of a novel with statistical and neural machine translation. *Frontiers in Digital Humanities*, 5:9.

Tu, Z., Liu, Y., Shang, L., Liu, X., and Li, H. (2017). Neural machine translation with reconstruction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Tu, Z., Lu, Z., Liu, Y., Liu, X., and Li, H. (2016). Modeling coverage for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 76–85.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Wang, A., Cho, K., and Lewis, M. (2020). Asking and answering questions to evaluate the factual consistency of summaries. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020. Association for Computational Linguistics.

Wang, H. (2020). Revisiting challenges in data-to-text generation with fact grounding. *arXiv preprint arXiv:2001.03830*.

Wang, X., Gao, T., Zhu, Z., Zhang, Z., Liu, Z., Li, J., and Tang, J. (2021). Kepler: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, 9:176–194.

Welleck, S., Weston, J., Szlam, A., and Cho, K. (2019). Dialogue natural language inference. In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3731–3741, Florence, Italy. Association for Computational Linguistics.

SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS

Wiseman, S., Shieber, S. M., and Rush, A. M. (2017). Challenges in data-to-document generation. *arXiv preprint arXiv:1707.08052*.

Wu, Z., Galley, M., Brockett, C., Zhang, Y., Gao, X., Quirk, C., Koncel-Kedziorski, R., Gao, J., Hajishirzi, H., Ostendorf, M., and Dolan, B. (2021). A controllable model of grounded response generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35:14085–14093.

Yavuz, S., Rastogi, A., Chao, G.-L., and Hakkani-Tur, D. (2019). DeepCopy: Grounded response generation with hierarchical pointer networks. In Nakamura, S., Gasic, M., Zukerman, I., Skantze, G., Nakano, M., Papangelis, A., Ultes, S., and Yoshino, K., editors, *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 122–132. Association for Computational Linguistics.

Yin, J., Jiang, X., Lu, Z., Shang, L., Li, H., and Li, X. (2016). Neural generative question answering. In Iyyer, M., He, H., Boyd-Graber, J., and Daumé III, H., editors, *Proceedings of the Workshop on Human-Computer Question Answering*, pages 36–42, San Diego, California. Association for Computational Linguistics.

Yu, T., Liu, Z., and Fung, P. (2021). AdaptSum: Towards low-resource domain adaptation for abstractive summarization. In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y., editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5892–5904. Association for Computational Linguistics.

Zhang, C., Lee, G., D'Haro, L. F., and Li, H. (2021a). D-score: Holistic dialogue evaluation without reference. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2502–2516.

Zhang, C., Lee, G., D'Haro, L. F., and Li, H. (2021b). D-score: Holistic dialogue evaluation without reference. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2502–2516.

Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., and Weston, J. (2018). Personalizing dialogue agents: I have a dog, do you have pets too? In Gurevych, I. and Miyao, Y., editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213. Association for Computational Linguistics.

Zhang, Y., Merck, D., Tsai, E., Manning, C. D., and Langlotz, C. (2020a). Optimizing the factual correctness of a summary: A study of summarizing radiology reports. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5108–5120. Association for Computational Linguistics.

Zhang, Y., Sun, S., Galley, M., Chen, Y.-C., Brockett, C., Gao, X., Gao, J., Liu, J., and Dolan, B. (2020b). DIALOGPT : Large-scale generative pre-training for conversational response generation. In Celikyilmaz, A. and Wen, T.-H., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278. Association for Computational Linguistics.

Zhou, B., Richardson, K., Ning, Q., Khot, T., Sabharwal, A., and Roth, D. (2021). Temporal reasoning on implicit events from distant supervision. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1361–1371.

Zhou, C., Neubig, G., Gu, J., Diab, M., Guzman, P., Zettlemoyer, L., and Ghazvininejad, M. (2020). Detecting hallucinated content in conditional neural sequence generation. *arXiv preprint arXiv:2011.02593*.

Zhou, K., Prabhumoye, S., and Black, A. W. (2018). A dataset for document grounded conversations. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J., editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 708–713. Association for Computational Linguistics.

Zhu, C., Hinthorn, W., Xu, R., Zeng, Q., Zeng, M., Huang, X., and Jiang, M. (2021). Enhancing factual consistency of abstractive summarization. In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y., editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 718–733. Association for Computational Linguistics.

SCIENCE AND TECHNOLOGY PUBLICATIONS