Detection of Rare Language Phenomena in NLP- Hallucination, Hyperbole, and Metaphor: A Survey

Naveen Badathala, Pushpak Bhattacharyya
Department of Computer Science and Engineering, IIT Bombay, India
{naveenbadathala, pb}@cse.iitb.ac.in

Abstract

In this survey paper, we delve into the detection of rare language phenomena, specifically focusing on the identification of hallucination, hyperbole, and metaphor. We present the importance of detecting and mitigating hallucination in language generation systems and analyze existing works and datasets in this domain. Hallucination detection and mitigation play a critical role in ensuring the reliability and effectiveness of natural language processing systems. Identification of figurative language like hyperbole and metaphor is important in any AI generation system to understand what the user wants to convey and respond accordingly to maintain the interaction grounded and interactive. We introduce the hyperbole and metaphor definitions with examples along with the existing datasets and related work proposed to detect the hyperbole or metaphor. This paper serves as a comprehensive study to understand the overview and existing works on the detection of hallucination, hyperbole, and metaphor.

1 Introduction

Natural Language Generation (NLG) involves generation of natural and fluent text and it is an important subfield of Natural Language Processing (NLP). NLG is very important due to various tasks like dialogue generation, abstractive summarization, machine translation, etc. NLG has made tremendous progress in neural-based text generation with the usage of large language models. These include Masked Language Models (MLM) like BERT (Devlin et al., 2018), Causal Language Models (CLM) like GPT-2 (Radford et al., 2019) has become the norm for any natural language generation task. Although the models based on these language models are able to generate fluent text, it is observed these NLG models are prone to generate text which is divergent to the source text. This problem of generating unwanted or irrelevant text is termed as hallucination. As shown in Figure 1 of E2E dataset (Novikova et al., 2017), for a given input data, if the NLG model is expected to generate end-to-end text based on given input data, the potential models like TGEN, GONG, SHEF2 added hallucinated content.

Input data: name[Blue Spice], eatType[coffee shop], near[The Bakers]
TGEN output: Blue Spice is a coffee shop with a **low price range**. It is located near The Bakers.
GONG output: Blue Spice is a **place** near The Bakers.
SHEF2 output: Blue Spice is a **pub** near The Bakers.

Figure 1: Example of Hallucination from E2E dataset

Hallucination can limit the application of many NLP related models in real-world and raises safety-related issues. There are domains such as finance and medical-related applications where information is crucial to making any decision, adding hallucinatory content to the doctor's report and financial document may hinder the potential use of these applications.

Figurative language is widely used in natural discourse, and it is frequently reflected in content generated on social media networks (Abulaish et al., 2020). Figurative languages are used to establish some communicative goals like- establishing a negative emotion, drawing emphasis to part of the text, adding interest to a subject, etc. (Roberts and Kreuz, 1994). The understanding of figurative languages like sarcasm, metaphor, simile, irony and hyperbole is important in various different NLP tasks like building accurate sentiment analysis systems, developing conversational AI systems that can hold meaningful conversations, etc. The example for the same was shown in Figure 2. This has lead to a great interest and value in understanding these figurative languages. The figurative languages like metaphor (Rai and Chakraverty, 2020)

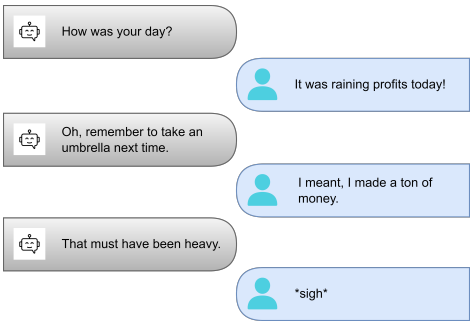

Figure 2: An example of the need for detecting hyperbolic and metaphoric sentences for AI systems.

and sarcasm (Joshi et al., 2017) are studied extensively while hyperbole remains less explored.

2 Motivation

Detecting hallucinated content helps in the improvement of hallucination mitigation models. It also serves as an important base to understand in which scenarios the NLG model is hallucinating depending on the task or context. Even for humans, it is hard to detect where the hallucination happens given the generated text. Since hallucinated content can hinder the performance and applications of NLP tasks to deploy and use in real-time. There are methods proposed to detect and mitigate the hallucination task, but they are specific to the task and vary with a different tasks. Hence, a general hallucination detection model can help to generalize the task of effective detection and mitigating hallucinated content in the generated text.

Metaphor is the most common choice of figurative language while hyperbole is the second most adopted rhetorical device in communication (Roger J., 1996) and hence it becomes important to study them to process them automatically. Hyperbole and metaphor are figurative languages that express an idea in contrast to the literal meaning of the sentence. Metaphors use comparison of objects or ideas to indicate the likeliness between them. Hyperbole is an exaggerated version of a statement often used for emphasis.

Relevance theorists had long treated both metaphors and hyperboles as not genuinely distinct categories as they are very closely related to each other (Sperber and Wilson, 2008). Recent works on hyperbole highlight the distinctive features of hyperbole over metaphors (Carston and Wearing, 2015). However, on the computational side, the existing works on hyperbole and metaphor detection treat them as isolated problems.

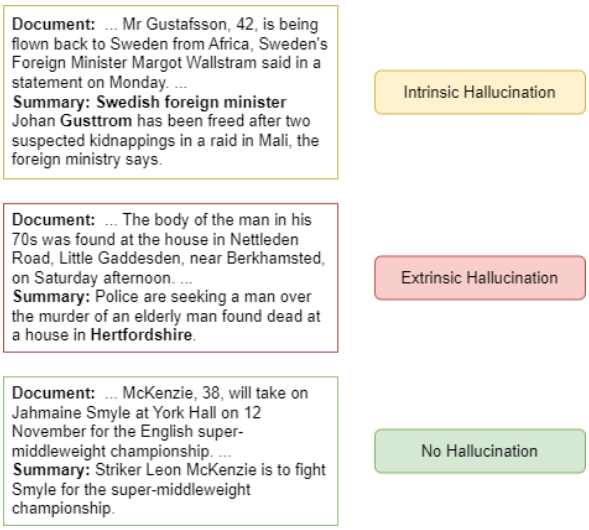

Figure 3: Examples of intrinsic, extrinsic, and non-hallucinated instances from the XSum dataset.

3 Background and Definitions

The term hallucination was inspired by psychology. In the medical context, hallucinations refer to the particular type of perception realized by an individual, without any external stimulus (Blom, 2010). Hallucination, as a psychological term, refers to an unreal perception that looks real on the surface. In the same way, in NLG, the generated text may contain information that might look correct but if we verify the information present, it might contain unfaithful or illogical text. Hallucination is further divided into **intrinsic** and **extrinsic hallucination**.

Intrinsic Hallucination

Intrinsic Hallucinations occur when the output generated by any NLG model contradicts the source text. For example, in a machine translation task, intrinsic hallucinations are defined as a span of the word(s) in the generated sequence containing incorrect information but they might represent the same entity type. Similarly, in the summarization task, if the generated summary contradicts the given source information or document, it is referred to as intrinsic hallucination.

Extrinsic Hallucination

Extrinsic Hallucinations - the output generated by any NLG model cannot be verified by the source information. In other words, the generated output neither contradicts nor is supported by the source information. It is important to note that extrinsic hallucinations are challenging to detect as they are not implied by the source text or information. one

Task	Sub-Task	Type	Source	Output
Abstractive Summarization [133]		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 [103]	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 [129]		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 [195]		Intrinsic	TEAM CITY WIN LOSS PTS FG_PCT BLK Rockets Houston 18 5 108 44 7	The Houston Rockets **(18-4)** defeated the Denver Nuggets (10-13) 108-96 on Saturday.
		Extrinsic	Nuggets Denver 10 13 96 38 7	**Houston has won two straight games and six of their last seven.**
Translation [224]		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.**

Figure 4: Examples of Intrinsic and Extrinsic hallucinations for different NLG tasks

interesting identity of extrinsic hallucinations is that it does not always contain factually incorrect data i.e. although the generated output text might not be validated whether it is true or false from the source information provided. But, the generated output can be factually correct considering the external or world knowledge. But, identifying extrinsic hallucination improves the consistency with the reference text and further identifies the content which is not required for the given specific task or context. The example of extrinsic hallucination in machine translation refers to the span of words consisting of additional information which can't be inferred from the given input or source text. In the context of the summarization task, extrinsic hallucinations refer to the output text neither supported nor contradicts by the given input article

The definitions of intrinsic and extrinsic slightly vary depending on the task, for example in machine translation intrinsic hallucination refers to the substitution of some other entity in place of the real or true entity while in abstractive summarization intrinsic hallucination refers to the contradiction to the source text. For other NLP tasks, a few examples are shown in Figure 4 (Ji et al., 2022)

In this section, we formally define the figurative languages that are used in our work.

Metaphors A metaphor is a figure of speech that is frequently used in everyday conversations. It makes a comparison in an implicit manner to something that is not true. Metaphors are formed by the intersection of the source and target domains, with the source domain features related to the target domain features via comparable properties. (Lakoff, 1993). For example, consider the sentence *"Life is a journey"*. Here, the source domain '*journey*' has a defined start and end as the property and it is mapped to the target domain '*life*', bringing out an implicit comparison between life and journey through the property of having a start and an end.

Simile A simile is another figure of speech where two different or unrelated things are compared explicitly (Israel et al., 2004). Similes are explicit about the comparison, whereas metaphors have a subtlety associated with them giving them more flexibility. For example, in the sentence *"He fought like a lion"*, a man is compared with a lion explicitly which could be expressed with a metaphor *"He is a lion"*. However, the metaphor *"His judgement*

Sentence	Hyperbole	Metaphor
Your plan is too risky, its a suicide	✓	✓
This kind of anger rages like a sea in a storm	✓	✗
Her strength awoke in poets an abiding love	✗	✓
My ex boyfriend! Treacherous person	✗	✗

Figure 5: Example sentences with Hyperbole and Metaphor labels.

is somewhat murky" cannot be explicitly expressed accurately with a simile as *"His judgement is something like murky"*.

Hyperbole Hyperbole is a figurative language in which the literal meaning is exaggerated intentionally. It exaggerates expressions and blows them up beyond the point they are perceived naturally to emphasize them (Claridge, 2010). In a hyperbolic statement, exaggeration can be brought about quantitatively by increasing or decreasing the quantity of the object or qualitatively by changing the subjective property of the object (Mora, 2009). It often makes use of similes and metaphors for bringing out exaggeration but it is not mandatory. Consider the following sentences:

- *I'm tired, I can't lift my hand.*
- *My heart is bleeding right now.*
- *Her anger radiated like a nuclear explosion.*

In the first sentence, the phrase, "can't lift my hand" is an exaggeration. The exaggeration here is brought about without any comparison. In the second example, the exaggeration is achieved with the help of a metaphor as we make an implicit comparison to bleeding, to drive home our point. In the third example, anger is explicitly compared to a nuclear explosion to underline exaggeration. In our work, we work with all these types of hyperboles and study the impact of understanding the metaphoricity of statements in identifying hyperboles.

4 Related Work on Hallucination Detection

Hallucination detection and mitigation have become important for any neural-based text generation model as the systems are often prone to hallucinate data in various scenarios which is undesirable for any type of task. To understand the hallucination in NLG, there are works to analyze the contributors to hallucination (Ji et al., 2022). These include hallucinations from data and hallucinations from language and inference.

Data-induced Hallucinations Hallucination from data can be the main cause for the models to deviate from the reference text and generation hallucinated content. This data issue was observed in many automatically created datasets like WikiBio (Dhingra et al., 2019). This can happen due to various data collection heuristics which might not be proven enough to collect only relevant and factual data. When an NLG model is trained on this type of data, it is obvious that the model hallucinates the data, since the reference text itself contains the such divergent text. For tasks like data-to-text, it is important for NLG models to maintain source information relevant and truthful to be able to effectively use them. Another problematic situation is the redundancy of the data present in the dataset. Since it is very difficult to filter the duplicate entries from the dataset but the model which uses this data to generalize the task and produce text, it is often observed that because of this duplicate data, models are favoring repeated or duplicate phrases more i.e. there exists some kind of bias to favor this type of behavior in models (Lee et al., 2021). There are tasks like automatic story generation and dialogue generation systems like chatbots etc. where the text generation should be diverse and interesting enough to use in text-based applications. Although it is required that these models should control their generation specific to a given prompt in the case of automatic story generation and appropriate response for a given query in chatbot applications. Even hallucination is important in these scenarios as well, because the generated text should not contradict existing prompts or already generated text in NLG tasks like automatic story generation. Even in dialogue generation systems like chatbots, along with interesting conversation styles, it is also important to maintain factual relevance, and generated text should not contradict or diverge from the earlier conversation data.

Training and Inference-induced Hallucinations Hallucination from training data can occur even if there is little divergence in the dataset (Parikh et al., 2020). The reason for the inclusion of hallucinations is due to the training and modeling decisions of neural models which are prone to generate hallucinated content. The encoder of these neural mod-

els can be a potential factor for the hallucinated content i.e. if encoders learn correlations in such a way that they are prone to make hallucinations across the given training data, this will generate factually incorrect output or divergent output text. In decoding strategies, the decoder receives encoded input and outputs the final target text. There are many existing approaches to achieving diverse text by modifying these decoding strategies to top-k-sampling (Dziri et al., 2021). There are two ways decoding strategies can lead to hallucinated results. First, the decoder can attend to irrelevant content of source text which might add hallucinated content to the output-generated text. Second, to include diversity property so that it will help in generating text that will be diversified enough for certain applications like automatic story generation, etc. but it will lead to adding hallucinated content as we increased the randomness of text generation by sampling top-k samples instead of selecting using the greedy approach i.e. selecting the most probable token. Apart from these decoding strategies, there is an exposure bias problem (Ranzato et al., 2015), where there is a difference in decoding between training time and inference time. when the target output text is longer, it is often observed that the generated text diverges from the text that is already generated and adds hallucinatory content due to this bias problem. Using large pre-trained models by fine-tuning to the corresponding resulting in fluent text, this fine-tuning approach has become important for any NLP tasks, but this approach may result in the model memorizing the various knowledge in the parameters (Madotto et al., 2020). Since these models rely on parametric knowledge more than the provided input text i.e. the model using this large pre-trained language model uses that parametric knowledge than taking information from the given input source text which results in adding unwanted or hallucinatory content that diverges from the expected output text.

Hallucination Detection and Mitigation There are approaches proposed to identify hallucinated content at the token level. (Zhou et al., 2021) propose a task of identifying each token in the generated output sentence with respect to source input that is hallucinated or not-hallucinated. To validate, they created synthetic data of machine translation and summarization i.e which includes automatically inserted hallucinations. An example of this task for machine translation is shown in Figure 6. This problem of token-level hallucination detection is formulated as the task where a label of binary i.e. either 0 or 1 is predicted for each word in the output text with respect to the source text.

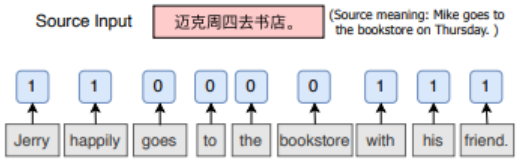

Figure 6: Token-level detection of hallucination in MT task

For dataset creation, used training data to create examples in a synthetic manner by substituting the new hallucinated tokens in the target side automatically.BERT (Devlin et al., 2018) is suitable for predicting the missing tokens or masked tokens independently by using the context of both sides of the masked token and is not suitable for generation easily. Autoregressive decoders like GPT-2 (Radford et al., 2019) prove enough to generate text auto-regressively and are suitable for generation tasks. Since the generated words are conditioned on left context only and generated masked token which is similar to the process of how humans writhe the text, due to this property they are able to maintain fluency. As shown in Figure 8, used the BART model (Lewis et al., 2020) to add the hallucinated tokens for the masked token of input text. The key challenge is that generated synthetic hallucinated text should be a sentence that is fluent and it doesn't vary considerably in comparison to the given input text. BART is a type of denoising autoencoder that can be used to generate text effectively.Its training procedure entails providing text that has been corrupted with an arbitrary noise function and expecting the training model to learn to reconstruct or generate the initially provided text.

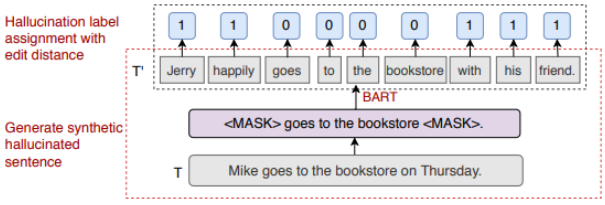

Figure 7: Synthetic Hallucinated Data Preparation

After generating the hallucinated sentences, labeled tokens as 1 or 0 depending on whether it

is hallucinated or not. This labeling strategy was based on Levenshtein distance, which gives appropriate labels for each token of the generated sentences. In this process, backtraced the deletion and substitution operations using dynamic programming, and all the positions involving these operations in generated hallucinated sentences are marked with the label of hallucinations, and the rest are considered not-hallucinated. The process of using a large pre-trained language model (LM) on existing data with respect to a specific task and eventually fine-tuning the pre-trained model is a common practice in natural language understanding. For fine-tuning, the input, original target, and hallucinated target (S, T, T^{l}) are combined and provided as input to the fine-tuned model. The final loss is defined as $L = L_{pred} + \alpha\ L_{mlm}$, where L_{pred} is the binary label classification loss, L_{mlm} is the masked LM Loss and α is the hyperparameter.

(Dhingra et al., 2019) and (Scialom et al., 2019) discuss the problem of hallucination based on sentence-level with involving tasks having reference text. (Maynez et al., 2020) deal with the hallucination problem in the task of abstractive summarization and provided large-scale human annotation to prove that summarization models hallucinate the content in the generated summaries. (Wang and Sennrich, 2020) showed how the hallucination problem exists in the machine translation tasks as well. (Rebuffel et al., 2022) discuss the word-level hallucination problem in the data-to-text generation task.

(Liu et al., 2022) proposed a task of hallucination detection which is at the granularity of the token level and specific to the free-form text generation systems. This approach takes generated text with marked tokens as input and outputs whether the marked tokens are hallucinated or not.

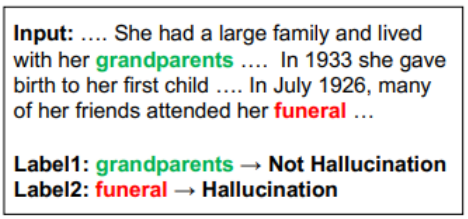

Figure 8: Overview of token-level reference-free hallucination detection

A hallucination dataset named hades was also created. For the creation of this dataset, they initially perform the operation of perturbation which converts the existing raw text information to format of perturbed text. After that, it was given to annotators to identify the given spans of text which was perturbed in earlier process contain hallucinations with respect to the given original text. It also proposed two sub-tasks based on real-world NLG applications i.e. offline setting and online setting. It is considered that in an offline setting, generation of text is completed and can use the bidirectional context. In an online setting, unidirectional proceeding context is considered and it can be used in applications where on-the-fly generation is required. To implement the token level hallucination detection task they created baselines of feature-based models like logistic regression and SVM etc. using features like statistical/model-based features and parsing features. Transformer-based models are employed by using a pre-trained model and perform fine-tuning to predict the binary hallucination labels for a given text span. These pre-trained models include BERT, GPT-2, XLNet, RoBERTa.

5 Related Work on Hyperbole and Metaphor Detection

Metaphors and hyperboles are the most used figures of speech in everyday utterances (Roger J., 1996). In the recent years, there are efforts made to understand metaphors and hyperboles computationally, giving rise to interesting techniques to automatically detect and generate them.

The existing works on metaphor and hyperbole detection deal with hyperboles and metaphors separately even though they have some common properties.

Hyperbole Detection Troiano et al. (2018) introduced hyperbole detection as task of identifying given sentence is hyperbole or not. They also released a 'Hypo' dataset for hyperbole detection. They used a feature set composed of imageability, unexpectedness, polarity, subjectivity, and emotional intensity. The classification was done with traditional ML based classification algorithms like logistic regression, svm, etc. Kong et al. (2020) introduced 'Hypo-cn', a Chinese dataset for hyperbole detection to complement Hypo. They showed that deep learning models can perform better at hyperbole detection with increased data. There is another work Biddle et al. (2021) using a BERT based detection system to extract the literal sentences of the hyperbolic counterparts in order to identify the hyperbolic phrases effectively. They also released

a test suite to detect the quality of hyperbole detection models. Tian et al. (2021) proposed a hyperbole generation task to generate clause or sentence level hyperbolic statements by completing an input prompt. Zhang and Wan (2022) introduced an unsupervised approach for generating hyperbolic sentences from literal sentences. They introduced two new datasets 'Hypo-XL' and 'HYPO-L' for their experiments.

Metaphor Detection Metaphors have been extensively studied even before hyperbole detection was introduced. Tsvetkov et al. (2014) introduced the TSV dataset with 884 metaphorical and non-metaphorical adjective-noun (AN) phrases collected from the web for metaphor detection. They showed that conceptual mapping learnt between literal and metaphorical use of words is transferable across languages. Mohler et al. (2016) released the dataset named LCC which contains sentence-level annotations for metaphors in four languages-English, Spanish, Russian, and Fars accounting to 188,741 data instances in total. Steen (2010) studied metaphor at word-level and was the first to include function words for metaphor detection with the new VUA dataset.

Metaphor detection has been studied in recent years using pre-trained large language models. Choi et al. (2021) used the contextual embeddings from large language models like BERT (Devlin et al., 2018) for the effective classification of metaphoric sentences. Aghazadeh et al. (2022) probed and analysed the metaphorical knowledge gained by large language models by testing them on metaphor datasets across languages.

6 Datasets and Analysis

There are no standard hallucination detection benchmark datasets available for the sentence-level hallucination detection tasks. Hence, existing datasets like the E2E NLG dataset (Novikova et al., 2017) are used to create hallucinated sentences and existing hallucination annotation datasets like XSum hallucination annotations (Maynez et al., 2020) are used.

E2E NLG Dataset E2E dataset consists of different pairs of meaning representations (mr) which are key-value pairs of restaurant-related data and a fluent sentence called reference which describes the information of restaurants based on given key-value pairs data. The example for the instance of this dataset is given as follows: for a given mr key-value pairs as, name [The Cotto], eatType [restaurant], food [Chinese], priceRange [high], customerRating [4/5], Area[countryside], familyFriendly[no], near[Lake view] and the associated reference fluent sentence is given as - The four-star restaurant, The Cotto offers a high-priced dining experience with a selection of wines and cheeses. The Cotto can be found near Lake View.

The details related to this dataset include a training data size of 42061 pairs of meaning representations (mr) and reference text. Validation or development set includes 4672 meaning representation pairs (mr) and associated text of reference. The test data set includes 4693 pairs of mr and reference.

To create the hallucinated sentences for the corresponding mr and reference, used a sequence2sequence model which takes meaning representations which are a series of key-value pairs that encode information about a restaurant as input and output are the fluent sentences that describe the restaurant based on the given input meaning representations which are in key-value format. The model is a simple sequence2sequence model and returns the output sequences based on the beam search evaluator i.e. it returns all the output sequences say if the beam size is k, it should return k sequences and their associated probabilities. The model is trained using the E2E dataset training data and generate hallucinated data, collected the output generated by model on unseen instance of data i.e. development and test data. In total there are 9365 instances of data that contain meaning representations (mr) source information, original reference text, and hallucinated reference text. The instance of this data i.e. source data, reference text, and hallucinated text are shown in Figure 9 This created data of hallucinated sentences are used in further sentence-level hallucination detection approaches.

XSum Hallucination Annotations XSum hallucination annotations are large-scale human level annotations performed on various abstractive summarizer models using the XSum dataset (Narayan et al., 2018). These annotations have labels for both factuality and faithfulness with respect to world knowledge and reference source input text respectively. This helps in the analysis of what type of hallucinations these systems often produce. Their analysis shows that pre-trained models are perfoming better in generating abstractive summaries fluently as well as interms of metrics like ROUGE

source data	reference text	hallucinated text
name [Alimentum], area [city center], familyFriendly [no]	There is a place in the city center, Alimentum, that is not family-friendly.	Alimentum is located in the city center. It is not family friendly and serves English food. It has a price range of 20-25 and is not child friendly.
name [The Punter], customer-rating [average], area [riverside], familyFriendly [yes]	The Punter at Riverside is Children friendly and has an average rating	In Riverside, The Waterman is a family-friendly restaurant with an average customer rating.
name [The Wrestlers], customer-rating [5 out of 5], familyFriendly [yes]	The Wrestlers has a customer rating of 5 out of 5. It is a family friendly place.	There is a family friendly restaurant called The Waterman. It has a customer rating of 5 of out of 5 and is 5
name [The Waterman], eatType [pub], food [Indian], priceRange [moderate], area [riverside], familyFriendly[no], near [Raja Indian Cuisine]	The Waterman is a moderate priced Indian pub. It is not kid friendly and is in the riverside area near Raja Indian Cuisine.	There is a moderately priced Indian pub called The Wrestlers. It is located in the riverside area. It is not kid friendly.
name [Wildwood], eatType [pub], food [French], area [riverside], familyFriendly [yes], near [Raja Indian Cuisine]	Wildwood is a child friendly French pub in riverside located near the Raja Indian Cuisine	There is a family friendly French restaurant called The Wrestlers. It is located in the riverside area. It has a price range of £20-25.

Figure 9: Examples of E2E dataset along with hallucinated text

etc. But, they are prone to add hallucinated content to the generated summaries. An example of this abstractive summarization of data hallucinations is shown in Figure 10. The faithfulness annotations are collected by showing news articles and system summaries to the evaluators and assigned task of finding the hallucinated content spans that are not implied from the given input article. Following this, the resultant data involves the information as follows: bbcid: corresponding doc id in the XSum dataset, system: neural summarizer model name, summary: genereated system summary, hallucination-type: hallucination type (intrinsic or extrinsic), hallucinated-span: part of summary where hallucination is present in 'summary', hallucinated-span-start: starting index of the hallucinated span, hallucinated-span-end: ending index of the hallucinated span, worker-id: id of the annotator. The factuality annotations are also collected by human evaluation by showing the related article along with the hallucinated system summary to the evaluators who are asked to assess the summary whether it is factually correct or not. The result of this process contains information as follows: bbcid: doc id in the XSum dataset, system: neural summarizer model name, summary: generated summary by abstractive summarizer system used. is-factual: yes/no, worker-id: id of the annotator.

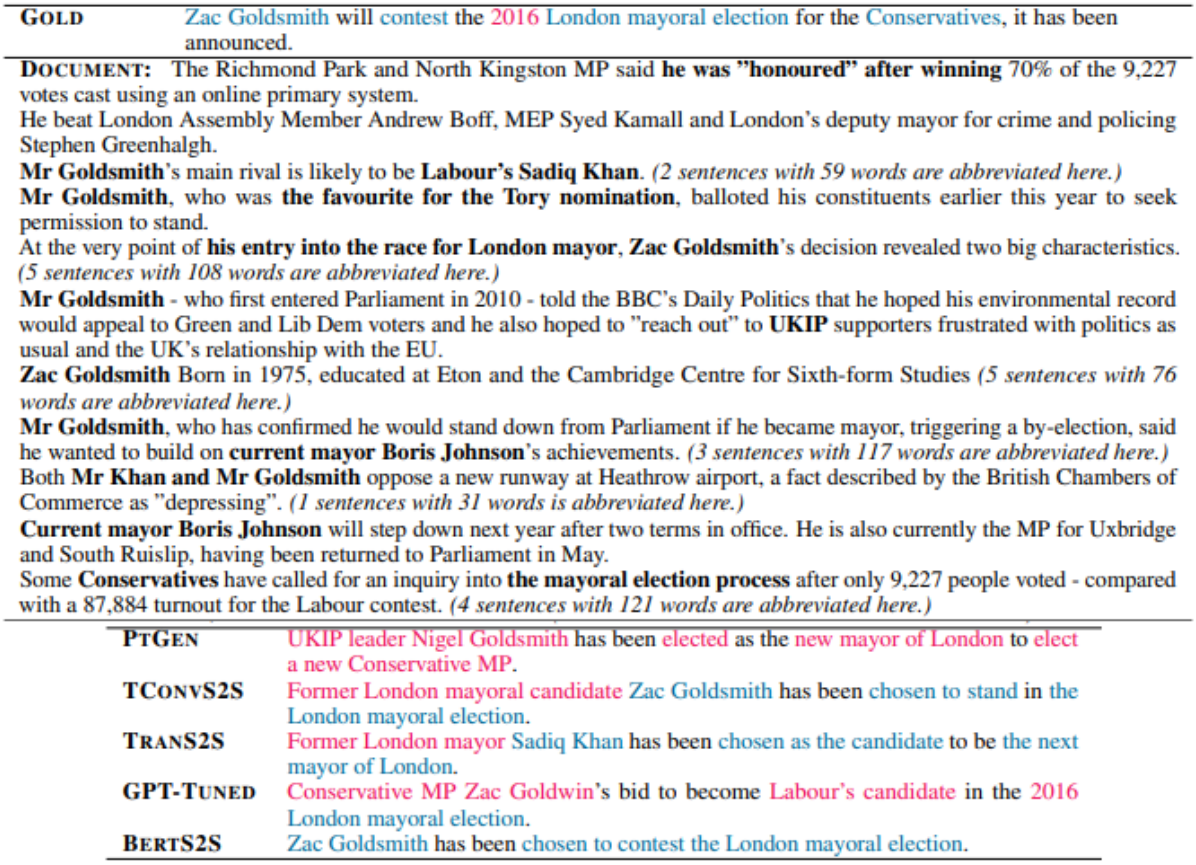

Figure 10: Examples of Intrinsic and Extrinsic hallucinations in XSumFaith annotations dataset.

Hyperbole and Metaphor Datasets There are existing hyperbole datasets like Hypo (Troiano et al., 2018) which consists of 709 hyperbolic sentences. Each sentence is accompanied by the paraphrased literal sentence and a sentence that contains the hyperbolic words or phrases in a literal sense. The hyperbolic and paraphrased sentences from the dataset amount to 1418 sentences. The Hypo-L dataset (Zhang and Wan, 2022). consists of 1007 hyperbolic sentences and 2219 paraphrased sentences. The statistics of the hyperbole and metaphor sentences in both the datasets are shown in Table 1 .

Dataset (# sent.)	Hyperbole	# sent.
HYPO (1,418)	✓	709
	✗	709
HYPO-L (3,326)	✓	1007
	✗	2219

Table 1: Statistics of hyperbole datasets.

The two metaphor datasets- LCC (Mohler et al., 2016) and VUA (Steen, 2010) statistics are shown in Table 2.

Dataset (# sent.)	Metaphor	# sent.
TroFi (3,838)	✓	1919
	✗	1919
LCC (7,542)	✓	3802
	✗	3740

Table 2: Statistics of metaphor datasets.

7 Summary

This survey paper provides an overview, and detailed background definitions with examples and related work existing for the rare language phenomena in natural language processing specifically hallucination, hyperbole, and metaphor. It also examines the existing work's limitations and areas to improve along with the availability of datasets for both the tasks i.e. hallucination detection and hyperbole & metaphor detection. This work serves as a study to understand the existing literature on rare language phenomena like hallucination, hyperbole, and metaphor in order to propose effective solutions for their identification.

References

Muhammad Abulaish, Ashraf Kamal, and Mohammed J. Zaki. 2020. A survey of figurative language and its computational detection in online social networks. *ACM Trans. Web*, 14(1).

Ehsan Aghazadeh, Mohsen Fayyaz, and Yadollah Yaghoobzadeh. 2022. Metaphors in pre-trained language models: Probing and generalization across datasets and languages.

Rhys Biddle, Maciek Rybinski, Qian Li, Cecile Paris, and Guandong Xu. 2021. Harnessing privileged information for hyperbole detection. In *Proceedings of the The 19th Annual Workshop of the Australasian Language Technology Association*, pages 58–67, Online. Australasian Language Technology Association.

Jan Dirk Blom. 2010. *A dictionary of hallucinations*. Springer.

Robyn Carston and Catherine Wearing. 2015. Hyperbolic language and its relation to metaphor and irony. *Journal of Pragmatics*, 79:79–92.

Minjin Choi, Sunkyung Lee, Eun-Kyu Choi, Heesoo Park, Junhyuk Lee, Dongwon Lee, and Jongwuk Lee. 2021. Melbert: Metaphor detection via contextualized late interaction using metaphorical identification theories. *ArXiv*, abs/2104.13615.

Claudia Claridge. 2010. *Hyperbole in English: A corpus-based study of exaggeration*. Cambridge University Press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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

Nouha Dziri, Andrea Madotto, Osmar R Zaiane, and Avishek Joey Bose. 2021. Neural path hunter: Reducing hallucination in dialogue systems via path grounding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2197–2214.

Michael Israel, Jennifer Riddle Harding, and Vera Tobin. 2004. On simile. *Language, culture, and mind*, 100.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2022. Survey of hallucination in natural language generation. *arXiv preprint arXiv:2202.03629*.

Aditya Joshi, Pushpak Bhattacharyya, and Mark J Carman. 2017. Automatic sarcasm detection: A survey. *ACM Computing Surveys (CSUR)*, 50(5):1–22.

Li Kong, Chuanyi Li, Jidong Ge, Bin Luo, and Vincent Ng. 2020. Identifying exaggerated language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7024–7034, Online. Association for Computational Linguistics.

George Lakoff. 1993. The contemporary theory of metaphor.

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*.

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*, pages 7871–7880.

Tianyu Liu, Yizhe Zhang, Chris Brockett, Yi Mao, Zhifang Sui, Weizhu Chen, and Bill Dolan. 2022. A token-level reference-free hallucination detection benchmark for free-form text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6723–6737, Dublin, Ireland. Association for Computational Linguistics.

Andrea Madotto, Zihan Liu, Zhaojiang Lin, and Pascale Fung. 2020. Language models as few-shot learner for task-oriented dialogue systems. *arXiv preprint arXiv:2008.06239*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan Thomas Mcdonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online.

Michael Mohler, Mary Brunson, Bryan Rink, and Marc Tomlinson. 2016. Introducing the LCC metaphor datasets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4221–4227, Portorož, Slovenia. European Language Resources Association (ELRA).

Laura Cano Mora. 2009. All or nothing: A semantic analysis of hyperbole. *Revista de Lingüística y Lenguas aplicadas*, 4(1):25–35.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *arXiv preprint arXiv:1808.08745*.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2017. The e2e dataset: New challenges for end-to-end generation. *arXiv preprint arXiv:1706.09254*.

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)*, pages 1173–1186.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Sunny Rai and Shampa Chakraverty. 2020. A survey on computational metaphor processing. *ACM Comput. Surv.*, 53(2).

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2015. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*.

Clément Rebuffel, Marco Roberti, Laure Soulier, Geoffrey Scoutheeten, Rossella Cancelliere, and Patrick Gallinari. 2022. Controlling hallucinations at word level in data-to-text generation. *Data Mining and Knowledge Discovery*, 36(1):318–354.

Richard M. Roberts and Roger J. Kreuz. 1994. Why do people use figurative language? *Psychological Science*, 5(3):159–163.

Kreuz Roger J. 1996. *Figurative language occurrence and co-occurrence in contemporary literature*.

Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2019. Answers unite! unsupervised metrics for reinforced summarization 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)*, pages 3246–3256.

D. Sperber and D. Wilson. 2008. Relevance: Communication and cognition. *A Deflationary Account of Metaphor*, page 84 – 108. Cited by: 1.

Gerard Steen. 2010. A method for linguistic metaphor identification : from mip to mipvu.

Yufei Tian, Arvind krishna Sridhar, and Nanyun Peng. 2021. HypoGen: Hyperbole generation with commonsense and counterfactual knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1583–1593, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Enrica Troiano, Carlo Strapparava, Gözde Özbal, and Serra Sinem Tekiroğlu. 2018. A computational exploration of exaggeration. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3296–3304.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 248–258, Baltimore, Maryland. Association for Computational Linguistics.

Chaojun Wang and Rico Sennrich. 2020. On exposure bias, hallucination and domain shift in neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3544–3552.

Yunxiang Zhang and Xiaojun Wan. 2022. MOVER: Mask, over-generate and rank for hyperbole generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 6018–6030, Seattle, United States. Association for Computational Linguistics.

Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Francisco Guzmán, Luke Zettlemoyer, and Marjan Ghazvininejad. 2021. Detecting hallucinated content in conditional neural sequence generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1393–1404.