DRAFT – What is Hallucination? – DON'T CITE

Kees van Deemter.
Utrecht University

Despite impressive recent advances in Natural Language Generation, researchers are still fundamentally unclear about what it means for a text to be factually accurate. To substantiate this claim, I examine current ideas about hallucination *and related concepts in Data-To-Text Natural Language Generation, subjecting these ideas to an informal logical analysis. I conclude that much about these concepts is still unclear, and that the NLP community has much to learn about them from colleagues in other research communities.*

1. Introduction: evaluating the veracity of a text

When computers produce text, the quality of the texts that are generated is of paramount concern. Accordingly, a substantial body of work across Natural Language Generation (NLG) focusses on evaluation of generated text REF REF. Evaluation, when performed well, can tell us how well a given NLG technique works, what its potential flaws are, what risks flow from these flaws , and how these risks may be mitigated.

A key quality criterion, for many text *genres*, is the veracity of the text. I will take the view that an assessment of the veracity of a text will include checking whether the text is factually correct and whether it includes all the information that it should. If a text is lacking in veracity, then any other virtues that it may possess are insignificant; far from cancelling out its lack of veracity, virtues such as fluency will only amplify the risks that the text poses when it is read (see e.g. Crothers et al. 2023).

In this short piece, I will argue that, when we assess the veracity of a text, we do not quite know what we are doing. I will argue, first, that most existing analyses of veracity hinge explicitly or implicitly on the notion of logical consequence (i.e., the "follows from" relation that is modelled in mathematical logic and Natural Language Inference). Second, I will point out some flaws in these analyses and offer a synthesis that does not suffer from these flaws. Third, I will argue that all analyses that I am aware of still suffer from some important limitations.

2. Existing analyses of hallucination and omission

When rule-based NLG systems are assessed, veracity is sometimes taken for granted (though see Van Deemter and Reiter 2018 for discussion); with the ascent of neural methods, researchers have become more widely aware that this is not justified (REF Vinyals & Le 2015, Kuhn & Knowles 2017, Rohrbach et al. 2018, Maynez et al. 2020, Dusek & Kasner 2020, Raunak et al. 2021 REF). Let us look briefly at three representative attempts to say more precisely what the problem is and what forms it can take. Since veracity comes into focus most sharply in relation to Data-To-Text NLG – where the input to the generator is data, not text – we focus primarily on that type of NLG.

A clear-headed, but coarse-grained, attempt came from Dusek & Kasner (2020) [REF], who discussed Data-text NLG systems whose input was a set of atomic statements and whose output was a verbalization of each of these statements. For instance, the input could say $Type(x) = Restaurant \wedge Food(x) = Italian \wedge Price(x) = low$, and the output could be a sentence of the form "x is an affordable Italian restaurant". The

authors noted two possible problems, namely hallucination and omission. They took an essentially logic-based approach, in which *hallucination* occurs when the output does not follow from the input; *omission* happens when the input does not follow from the output. Thus, if the generator produces "x is an affordable *vegetarian* Italian restaurant", then the information that $Style(x) = vegetarian$ is hallucinated; if it produces "x is an affordable restaurant", then that involves an *omission*, because $Food(x) = Italian$ is omitted. Note that an analysis along "logical" lines assumes that we are able to somehow handle all the peculiarities of natural language inference, including such thorny problems as the vagueness and ambiguity that can occur in generated texts (e.g., what combination of prices can still count as affordable?), and including the treatment of infelicities, typos and incorrect rendering of names (REF Faille et al. 2021).

A more finegrained analysis, based on a similar outlook, was offered in Ji et al (2023), who distinguished between *intrinsic hallucination* and *extrinsic hallucination*: the former is "output that contradicts the source"; the latter is "output that can neither be supported nor contradicted by the source". This analysis was applied to a number of NLG areas, including data-text NLG; omissions were not discussed.

Finally, in their work on categorizing errors in sports reports, Thomson and Reiter 2022 offered a categorisation into six categories. Although they did not present this as a generic ontology of hallucination and omission, it will be instructive to look at it through that lens. In a nutshell, these authors distinguished between generations that contain an incorrect number, an incorrect named entity, an incorrect word (when it is neither an incorrect number nor an incorrect named entity), "non-checkable" information (much like Ji et al.'s extrinsic hallucination), context errors , and "other" types of mistakes. Most other catagorisations are even more strongly domain dependent, for example by focussing on the different kinds of errors that are made in summarisations from doctor-patient dialogues [REF Moramarco et al. 2022].

The literature documents a wide range of types of information that can be hallucinated or added. To obtain the broadest possible understanding of what error types can occur, I will give examples from a well-understood domain and use an input that is simple yet complex enough to permit a diverse range of possible output error types. Our discussion will not be limited to errors that current NLG models are known to commit frequently, because tomorrow's models may commit different errors than today's. So, for concreteness, suppose our domain is a localized weather description, regarding a place somewhere in a desert, at noontime. The input of the generator is the information that **the temperature is between 20 and 25 degrees Celsius**: $input = 20 < t < 25$. Now consider a number of possible outputs:

0. "It's between 20 and 25 degrees"
   D&K and Ji-et-al and T&R: Input and output are well matched.
1. "It's between 10 and 40 degrees"
   D&K: Omission. Ji-et-al: (Not discussed)] . T & R: Incorrect number.
2. "It's between 24.5 and 24.6 degrees"
   D& K: Hallucination. Ji-et-al: Extrinsic hallucination? T&R: Incorrect number.
3. "It's between 23 and 30 degrees".
   D&K: Hallucination and Omission. Ji: Extrinsic hallucination? T&R: Incorrect number.
4. "It's between 40 and 45 degrees"
   D&K: Hallucination and Omission. Ji: Intrinsic hallucination? T&R: Incorrect number.

The above represents my best guess at how the definitions should be applied to the examples at hand. For example, although the input $20 < t < 25$ does not follow from output (1) – because (1) is consistent with temperatures that are lower or higher than what the input specifies – it is hard to isolate the precise part of the statement that is wrong. Several things are clear though. For example, Dusek & Krasner's analysis does not distinguish between (3) and (4), because all they can say about both situations is that omission *and* hallucination occur (i.e., the input does not imply the output and the output does not imply the input). Ji et al. do distinguish between (3) and (4), but they conflate (2) and (3), because both these outputs "can neither be supported nor contradicted by the source". Thomson and Reiter 2022 [REF] treat all four errors as incorrect numbers.

In other words, the three analyses differ from each other, with each analysis conflating two or more types of misinformation, each of which would pose very different kinds of risks if they occurred in an actual weather report. It seems to me that if we want to use the concepts of Hallucination and Omission as tools for recognizing what kinds of errors a generator can make, or as the bedrock of a metric for assessing problems with the veracity of generated text, or as a starting point for mitigating errors (see Ji et al. REF for discussion), then we need to go back to the drawing board.

3. A synthesis of existing analyses

The good news is that a synthesis that combines the core of all three analyses is not hard to imagine. The key is to ask what Logical Consequence (i.e., "follows from", denoted by the symbol "$\models$") relations can exist between input and output. Let's assume that the output to a Data-text generator is a clear, unambiguous proposition, and that neither the input nor the output is internally inconsistent. Now, it can be true or false that $input \models output$; likewise, it can be true or false that $output \models input$; furthermore, if $input \not\models output$ and $output \not\models input$ (cases (d) and (e)), it is important to know whether $input \models \neg output$ (as in Ji et al.'s intrinsic hallucination, which hinges on input and out being inconsistent with each other):[1] The result is a clear analysis that is able to separate between each of (0)-(5):

a. $input \models output$ and $output \models input$.
   An example is (0) above. (Input and output are perfectly matched.)
b. $input \models output$ but $output \not\models input$. (Output is too weak.)
   An example is (1) above.
c. $input \not\models output$ but $output \models input$. (Output is too strong.)
   An example is (2) above.
d. $input \not\models output$ and $output \not\models input$, and $input \not\models \neg output$.
   (Input and output are logically independent of each other.)
   An example is (3) above.
e. $input \not\models output$ and $output \not\models input$, and $input \models \neg output$.
   (Input and output contradict each other.)
   An example is (4) above.

1 In case (c), $input \models \neg output$ would imply that the $output$ is internally inconsistent. Likewise, in case (b), $output \models \neg input$ would imply that the $input$ is internally inconsistent.

4. Limitations of these analyses

The good news, I think, is that the above synthesis is straightforward enough that it might be used in error annotation. It can also be the basis of a computational metric, because logical consequence can be computationally approximated through Natural Language Inference (see e.g. REF Dusek and Kasner). The bad news is that this synthesis throws some limitations of current thinking about hallucination and omission into sharp relief.

The first limitation is a lack of clarity concerning the distinction between semantics and pragmatics. The issues here are ones that are highlighted in Thomson and Reiter 2022. When a multi-sentence text is generated, then the analyses above can be applied *per* sentence, or to the text as a whole. In both cases, each sentence should be given an interpretation appropriate for its context (e.g., with anaphoric pronouns resolved). Likewise, if a sentence has a conversational or conventional implicature (e.g. REF GRice) then this implicature should be judged in the same way as the literal meaning of the text, in terms of the categories (a)-(e) of section 3.[2] Similar remarks could be made about presupposition, irony, and metaphor, which go beyond what is stated literally in a text, yet all of which can cause hallucination and omission. Note that, conversely, a text can also be misclassified as hallucinating if it fails to apply pragmatic reasoning. For example, when the weather report for a sunny day speaks metaphorically of “wall-to-wall sunshine", then a narrowly semantic version of the above analyses would classify this statement as hallucination, but a version that understands metaphor might consider it to be truthful. This is not an unsurmountable problem, but it means that the “follows from" relation should be pragmatically (as well as semantically) aware.

The second limitation is that by relying on the “follows from" relation, one focusses on truth conditions only. This is perhaps most clear in relation to case (d) above.

d1 “It’s between 23 and 30 degrees" (Sentence (3) above).
d2 “The night temperature will be below 10 degrees."
d3 “The cat is on the mat".

All three outputs are logically independent of the input, which says that $20 < t < 25$, so all three fall into class (d). Yet (d1) gets everything right except the boundaries of the time interval; (d2) changes the topic but makes a statement that is highly probable (given knowledge about temperature differences between day and night in the desert); (d3), finally, is completely unrelated to the input. None of the analyses discussed seems able to distinguish between (d1)-(d3). More generally, logic-based analyses tend to be blind to what a sentence is *about*. In this respect they are inferior to domain-specific analyses such as Moramarco et al. 2022 [REF], with some highly specific error categories, such as “Symptom mentioned (by the patient) is reported as fact".

The third limitation relates to error severity. The severity of an error can vary enormously within each of the different classes (b-e). In sentence (3), for instance, if the output said “It is 24.9–25.1 degrees", it would still be a class-(d) type error, yet the error would have been far less grave. One thing that is needed here is a distance metric that quantifies the extent of an error. Despite some nascent ideas (Van Miltenburg REF, Moramarco 2022, Chen and Van Deemter 2023), this is still a little explored area.

2 Thomson and Reiter REF, devoted a separate category (“Context Errors") to statements that are correct when interpreted verbatim but erroneous in context.

Moreover, it is not evident that a *big* error is always a serious error, for instance because a big error may be more likely to be noticed by readers, and hence less risky.

Large Language Models. With the widespread use of LLM-driven Generative AI, our lack of clarity about veracity is becoming increasingly hazardous. The question is raised, not only by developers and readers but by the wider public interested in uses of AI, what is the quality of the texts that are generated by these systems.

Analyses along the lines of sections 2 and 3 are about the relation between an output and an input; for example, (1) was classified as a category (a) ("Output is too weak") error because an aspect of the input was not expressed in the output of the generator. These analyses are only applicable when the task of NLG is to express all information in its input, which is not the case with many typical uses of LLMs. In such cases, output texts will have to be assessed independently of any input, for example in terms of whether the text is true, and whether it is informative enough for the purpose at hand.

Although "informative enough" is a complex, and not purely logical, concept, logical analysis might once again help us on our way, for example *via* a Bratman-style logic of beliefs, desires and intentions (BDI) (REF Bratman book). BDI logics add to our arsenal because they allow us to reason about the inferences that a hearer will draw from an utterance.

Suppose, for example, today's weather forecast does not mention a hurricane, then listeners are likely to infer that no hurricane will take place; if a hurricane hits the country nonetheless, then the forecast was misleading, and hence arguably of low veracity.[3] Hallucinations of this type are, in the terminology of Sakama et al., "Withholding Information" ; they are important because, in politics and elsewhere, misinformation often hinges on strategic omissions REF. A BDI analysis does not by itself tell us how to *recognize* whether a given LLM-generated text "withholds information", but it might give us a starting point for understanding what a concept like omission can mean when LLMs are used for purposes other than classic Data-text NLG. Another type of misleading highlighted by Sakama and colleagues is called Half Truth: an output is a Half Truth if a certain false proposition r is not communicated directly, yet an output is generated of which the hearer believes that r follows from it.[4] Sakama et al's example is of someone bragging that he holds a permanent position at a company, without saying that the company is actually almost bankrupt.

5. Conclusion

NLG has made great strides, but there are no grounds for triumphalism. An important reason for practicing humility is that we do not yet know how to think properly about the veracity of generated text. I believe that, to start addressing these problems, the NLP community should be prepared to do two unfashionable things: first, it should liaise with the logic and formal argumentation communities, where concepts like truth and evidence have always taken center stage; and secondly, it should revisit insights into the pragmatics of natural language, with its distinction into what a sentence asserts and what an utterance of the sentence gives readers to understand in a given context, *via* such mechanisms as implicature, presupposition, irony, and metaphor (e.g. Levinson REF). NLP disregards such matters at their peril, and potentially at our collective peril.

3 Letting q abbreviate "There will be a hurricane", and using $C_{SH}q$ to say that the speaker (S) tells the hearer (H) that q , this can be expressed (simplifying REF Sakama et al's formalism) as $\not\models q$ *and* $B_H(q \rightarrow (C_{SH}q)$ *and* $\neg C_{SH}q$.

4 In BDI notation, with p in the role of the output, $C_{SH}p$ *and* $\not\models r$ *and* $\neg C_{SH}r$ *and* $B_H(p \rightarrow r)$.