Detecting Hallucinations using SemAnte

Tanmay Garg
CS20BTECH11063

Tanmay Goyal
AI20BTECH11021

Tanay Yadav
AI20BTECH11026

Abstract

In this work we propose a novel attention-based metric called SemAnte to detect hallucinations in natural language generation systems. Hallucinations, which are questionable or contradictory outputs, remain a challenge to detect as models are designed to understand, generate and interact with the human language. SemAnte utilizes attention scores from a transformer model to compute the cross-attention between words in the reference and predicted sentences and apply a similarity score between their embeddings to achieve our SemAnte Score. While preliminary, these findings indicate the potential of SemAnte for this task, the experimental results on multiple examples demonstrate SemAnte can better capture semantic similarity compared to traditional ROUGE metrics and is more effective at identifying hallucinatory predictions.

1. Introduction

Natural language systems have made significant progress in producing human-like text that can be easily misunderstood as being written by a person. However, these large language models (LLMs) still struggle with generating information that may be factually incorrect or might not be related to the context, known as hallucinations [12]. Such ungrounded text generations could have real-world consequences if generated by conversational agents or summarization models.

In this work we present, SemAnte, a novel attention-based approach for detecting hallucinations. SemAnte utilizes intrinsic information from the generation process itself to compute semantic similarity between generated and source text. By leveraging attention scores from transformer-based models, it aims to capture the underlying relationships between words better than existing surface-level metrics such as ROUGE [5].

In the following sections, we first outline the problem of hallucinations and existing evaluation methods. We then introduce SemAnte's technical approach and methodology. Experimental results on multi-sentence examples demonstrate its ability to better identify implausible predictions compared to ROUGE scores. Lastly, we discuss promising directions to advance this work, like analyzing longer text spans. Overall, this work contributes a new attention-based metric for an important open challenge in natural language generation evaluation.

2. SemAnte - Semantic Attention Metric

We are proposing - SemAnte - an attention-based metric for detecting semantic hallucinations. It utilizes information directly from the model's attention weights instead of relying on perfect reference texts. SemAnte looks at cross-attention scores that a transformer model assigns between the reference sentence and the generated sentence. It then identifies the word in the generated sentence with the highest attention for a word in the reference sentence and vice-versa.

This gives us a pair of words, where one word directly relates to the other based on their semantic understanding rather than just the lexical structure. SemAnte then calculates the sum of similarity scores of these pairs with respect to their word embeddings and averages them.

By utilizing the inherent semantic meaning of the words and their relation through attention weights, our goal is for SemAnte to offer an effective automatic approach for detecting hallucinations that are applicable to a wide range of NLP tasks.

2.1. Motivation

The motivation behind introducing a new metric was to capture the underlying semantic similarity between a pair of sentences unlike the ROUGE metric, which only takes the lexical appearance of words into account. The ROUGE-1 metric counts the number of common unigrams between the two sentences while the ROUGE-2 metric counts the number of common bigrams between the two sentences.

Another existing method to do a similar task is produce sentence embedding for the pair of sentences using a model such as All-MPNet-Base-v2 [11], and then calculate the similarity between the two sentences. For example, consider the pair of sentences:

Reference r : *He is travelling to Mumbai by Aeroplane*

Generated 1 g_1 : *He is coming to Mumbai by aeroplane*

Generated 2 g_2 : *He is coming from Mumbai by aeroplane*

The similarity score using the sentence embeddings produced by BERT [1] between r and g_1 is 0.887, while the score between r and g_2 is 0.901, which tells us that the second generated sentence is more similar to the reference sentence as compared to the first one. However, we know that the second sentence is completely contradictory to the reference sentence.

Keeping these points in mind, we came up with our metric SemAnte.

2.2. Method

In this section, we will introduce the method and mathematical notions for SemAnte.

- Given a reference sentence s and a predicted sentence p, we obtain the cross-attention scores from s to p as $T(s, p)$
- For each word $w_s \in s$ and $w_p \in p$, let $T(s, p)[w_s, w_p]$ be the attention w_s gives to the word w_p.
- Define

$$v(w_s) = \arg\max_{w_p \in p} T(s, p)[w_s, w_p]$$

- Our score from s to p is then given by:

$$S_{s \to p} = \frac{1}{|s|} \sum_{w \in s} CS(E(w), E(v(w))) \quad (1)$$

where $CS(\cdot, \cdot)$ stands for the cosine similarity and $E(\cdot)$ stands for the embedding of a word.

- We similarly calculate the scores for p to s:

$$S_{p \to s} = \frac{1}{|p|} \sum_{w \in p} CS(E(w), E(v(w))) \quad (2)$$

- We return our final score as the maximum of the two:

$$S = \max\{S_{s \to p}, S_{p \to s}\}$$

For SemAnte, we have decided to use two kinds of embeddings: GloVe [9] and embeddings produced from an uncased-BERT Model with the inputs as the sentence individually. We initially experimented with Word2Vec [7] as well, but decided to drop the embeddings because of the removal of stop words and other words that would otherwise provide context information for semantic similiarity. Moreover, Word2Vec required the entire sentence to be preprocessed heavily and would result in a loss of meaning and other semantic relations across words and the pair of sentences.

SemAnte has been tested across multiple sentence pairs that are discussed in Section 3. We also assume that the LLM would generate 3 sentences and then SemAnte would help choose the least hallucinated sentence out of the three and compare it with the least hallunicated sentence from ROUGE.

3. Experiments

3.1. Ablation Study

We noticed that while creating the attention maps, a lot of the attention would flow into the [SEP] tokens for reasons unknown to us. We used the following convention while considering the [SEP] tokens: for a word $w \in s$, if $v(w) =$ [SEP], then we simply ignore w while calculating the score in Eq. 1 and Eq. 2. However, this resulted in extremely low scores which were not very helpful.

On the other hand, if we were to take into account [SEP], for a word $w \in s$, if $v(w) =$ [SEP], then we simply take the word with the next highest attention score and use it for calculating the similarity.

For example, consider the following two sentences:

Reference: *Mumbai is the best city on earth*

Generated: *The best city on earth is Mumbai*

The attention maps for these sentences is given in Figure 1

Embeddings	[SEP] (Included)	[SEP] (Not Included)
GloVE	0.014	0.459
BERT	0.035	0.538

Table 1. Comparing the scores obtained by including and omitting [SEP]

We find that the scores generated while including [SEP] is lower by atleast a factor of 10 than the scores generated by excluding the token. Also, in Fig. 1, both the sentences are alike in meaning, and hence, we would wish for "Mumbai" to attend to the words "Mumbai" and "city", which is exactly what happens once we exclude the [SEP] token. Thus, we made the conscious decision to ignore the [SEP] token.

3.2. Testing the metric

To test the metric, we decided to evaluate a few examples and draw comparisons between SemAnte and the ROUGE

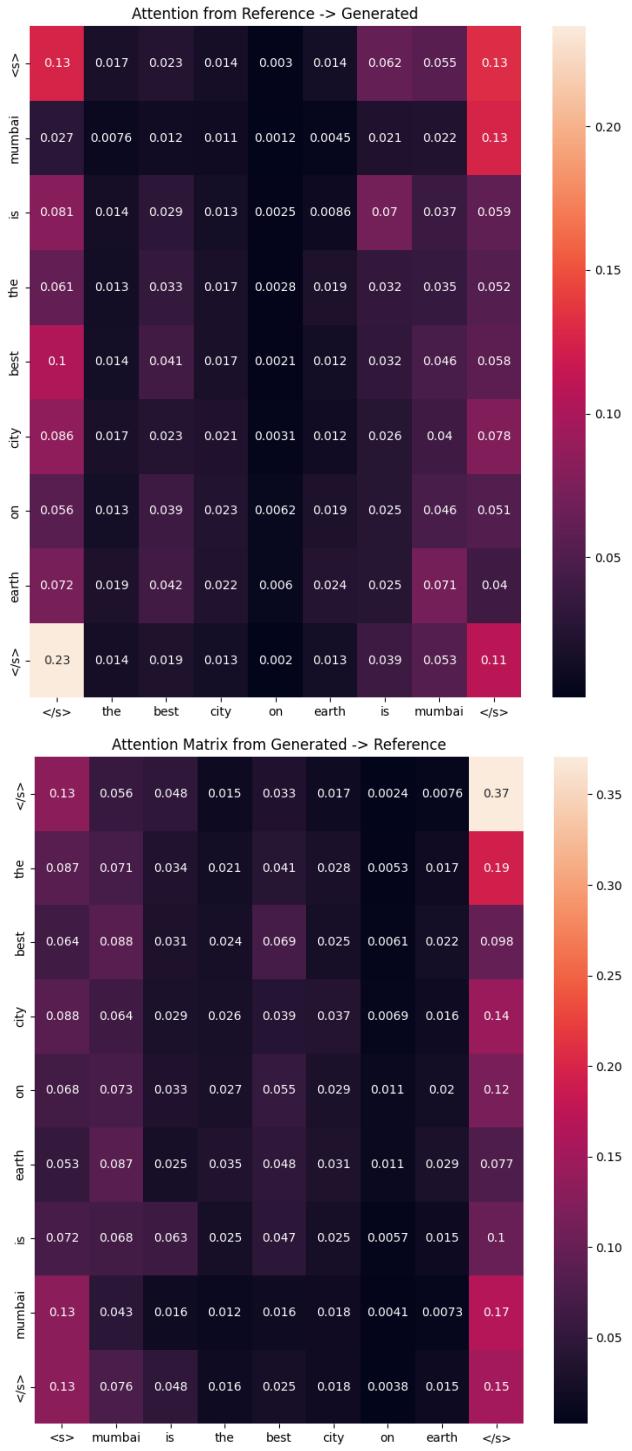

Figure 1. Attention scores for the sentences: ”Mumbai is the best city on earth” and ”The best city on earth is Mumbai”

metric. These examples are drawn in a way where the meaning of the reference and generated sentences is exactly the same, or is tweaked in a manner that ROUGE would not be able to pick up upon.

1.

Reference sentence r : *He is a good man*

Predicted sentence s_1 : *He is a great guy*

The results for these pair of sentences are tabulated in Table 2

Sentence	Ours (GloVe)	Ours (BERT)	ROUGE-1	ROUGE-2
s	0.818	0.825	0.6	0.5

Table 2. Results for the pair of sentences ”He is a good man” and ”He is a great guy”

2.

Reference sentence r : *he is taking a flight to mumbai*

Predicted sentence 1 s_1 : *he is arriving in mumbai by airplane*

Predicted sentence 2 s_2 : *he is coming from mumbai by airplane*

Predicted sentence 3 s_3 : *he is travelling by air to mumbai*

The results for these sentences are tabulated in Table 3

Sentence	Ours (GloVe)	Ours (BERT)	ROUGE-1	ROUGE-2
s_1	0.508	0.592	0.429	0.167
s_2	0.439	0.537	0.429	0.167
s_3	0.474	0.517	0.571	0.333

Table 3. Results for reference sentence ”he is taking a flight to mumbai” and the three generated sentences ”he is arriving in mumbai by aeroplane”, ”he is coming from mumbai by airplane”, and ”he is travelling by air to mumbai” respectively.

3.

Reference sentence r : *i will go see a movie*

Predicted sentence 1 s_1 : *i am going to watch a movie*

Predicted sentence 2 s_2 : *i will watch a film now*

Predicted sentence 3 s_3 : *i am watching a movie*

The results for these sentences are tabulated in Table 4

4.

Reference sentence r : *i am on leave today*

Predicted sentence 1 s_1 : *this is my day off*

Predicted sentence 2 s_2 : *i am not going to work today*

The results for these sentences are tabulated in Table 5

Sentence	(GloVe)	(BERT)	ROUGE-1	ROUGE-2
s_1	0.758	0.686	0.462	0.182
s_2	0.739	0.573	0.500	0.200
s_3	0.654	0.559	0.545	0.222

Table 4. Results for reference sentence "i will go see a movie" and the three generated sentences "i am going to watch a movie", "i will watch a film now", and "i am watching a movie" respectively.

Sentence	(GloVe)	(BERT)	ROUGE-1	ROUGE-2
s_1	0.610	0.384	0.000	0,000
s_2	0.650	0.516	0.500	0.200

Table 5. Results for reference sentence "i am on leave today" and the three generated sentences "this is my day off", "i am not going to work today" respectively.

4. Results

SemAnte works best as a relative metric rather than an absolute metric. During inference time, an LLM would may generate 3 draft outputs and use an internal scoring method to give the user the best output. SemAnte could be used in the process of choosing the best answer by helping select the least hallucinated answer.

In Table 2, we can observe that SemAnte can understand the semantic similarity between the sentences and can catch the pair of words that should have high attention scores such as "*great, good*" and "*man, guy*", compared to ROUGE metric that only judges the similarity based on the n-gram method. We can observe a similar trend in Table 3.

In Table 4, the generated sentences have different tenses such as present and future tense. In all the generated sentences, SemAnte with GloVe and BERT embeddings score s_1 as the least hallucinated sentence whereas ROUGE-1 and ROUGE-2 score s_3 as the least hallucinated sentence. The reference sentence is a sentence in the future tense, while s_3 is a present tense sentence and s_1 is a future tense sentence, and thus, s_1 is the sentence that should have been scored higher.

5. Conclusion

In this work, we have presented our metric and also discussed preliminary results that indicate that SemAnte performs better than ROUGE metric in detecting hallucinations from a sentence with the semantic meaning of the understand as its foundation. We also observe that SemAnte works better as a relative metric than an absolute metric as seen in Tables 3, 4, and5.

One of the future directions for our work is testing SemAnte on larger datasets with various models and scenarios could yield better analysis on the applications of SemAnte. We can also explore a more complex transformer model architecture in our method to see if it gives better results is a promising direction of research. Finaly, we wanted to check if more sophisticated mathematical models could be used to model our score. For example, some of the ideas we had in mind were using a weighted averaging or adding some kind of penalty for longer sentences.

Overall the contributions of SemAnte could open up ways to detect hallucinations between the generated passage and reference passage on a semantic level providing a step to detect hallucination after the lexical analysis.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, 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 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 2

[2] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions, 2023.

[3] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12), mar 2023.

[4] Haoqiang Kang, Terra Blevins, and Luke Zettlemoyer. Comparing hallucination detection metrics for multilingual generation, 2024.

[5] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. 1

[6] Tianyu Liu, Yizhe Zhang, Chris Brockett, Yi Mao, Zhifang Sui, Weizhu Chen, and Bill Dolan. 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)*. Association for Computational Linguistics, May 2022.

[7] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013. 2

[8] Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cicero Nogueira dos Santos, Henghui Zhu, Dejiao Zhang, Kathleen McKeown, and Bing Xiang. Entity-level factual consistency of abstractive text summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, Apr. 2021.

[9] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Oct. 2014. 2

[10] Thibault Sellam, Dipanjan Das, and Ankur Parikh. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, July 2020.

[11] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding, 2020. 1

[12] S. M Towhidul Islam Tonmoy, S M Mehedi Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. A comprehensive survey of hallucination mitigation techniques in large language models, 2024. 1

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017.