# SemAnte - Semantic Attention Metric

Tanmay Garg, Tanmay Goyal, Tanay Yadav

CS20BTECH11063, AI20BTECH11021, AI20BTECH11026

CS5803 - NLP

# Problem Statement: Hallucination Detection

- NLP systems are designed to understand, generate or interact with human language through various training paradigms
- Such models can produce implausible or contradictory outputs termed as *hallucinations*
- The two types of hallucinations are:
  - **Intrinsic**: Output contradicts the source information
  - **Extrinsic**: Output cannot be verified from source information
- The common causes of hallucinations include:
  - Datasets might contain contradictory or incorrect information
  - Improper training due to modeling choice and training paradigm
- Existing hallucination detection methods rely on ground truth summaries, that are not available for most free-form text generation applications

# SemAnte - Semantic Attention Metric

- Propose a attention-based semantic hallucination detection metric - SemAnte
- Uses attention scores from a transformer model to retrieve cross attention scores
- Attention scores help choose the highest cross attention pairs
- Distance between the two embedded words are calculated and averaged across all the words

# SemAnte - Method

- 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 $w$ in $s$, we obtain the corresponding word $v$ which gets the highest attention from $w$.
- Our score from $s$ to $p$ is then given by:

$$S_{s \to p} = \frac{1}{|s|} \sum_{w \in s} cosine\_similarity(embedding(w), embedding(v))$$

- We similarly calculate the scores for $p$ to $s$:

$$S_{p \to s} = \frac{1}{|p|} \sum_{w \in p} cosine\_similarity(embedding(w), embedding(v))$$

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

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

# Choosing the Embeddings

- For our purpose, we decided to use three kinds of embeddings: Word2Vec, GloVe, and embeddings produced from a uncased-BERT model.
- After some experiments, we decided to drop the embeddings from Word2Vec because of the removal of stop words which may contain relevant information regarding the sentences.
- We have tested SemAnte across multiple sentence pairs. Each reference sentence, the LLM would generate 3 predictions and then it chooses the best result.
- SemAnte helps in choosing the best sentence with least hallucination compared to ROGUE.

# Experiments: I

- Reference sentence $r$: *He is a good man*
- Predicted sentence $s_1$: *He is a great guy*

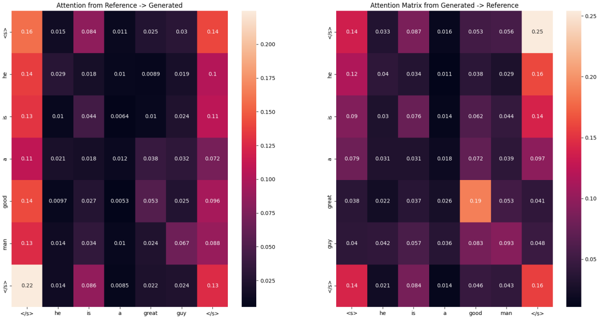


Figure: Attention Scores

The score using the GloVe embeddings is 0.818 while the score using BERT embeddings is 0.825. The ROGUE-1 and ROGUE-2 scores are 0.6 and 0.5 respectively.

# Experiments: II

- Reference sentence $r$: *he is taking a flight to mumbai*
- Predicted sentence $s_1$: *he is arriving in mumbai by airplane*
- Predicted sentence $s_2$: *he is coming from mumbai by airplane*
- Predicted sentence $s_3$: *he is travelling by air to mumbai*

| Sentence | Ours (GloVe) | Ours (BERT) | ROGUE-1 | ROGUE-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 |

# Experiments: III

- Reference sentence $r$: *i will go see a movie*
- Predicted sentence $s_1$: *i am going to watch a movie*
- Predicted sentence $s_2$: *i will watch a film now*
- Predicted sentence $s_3$: *i am watching a movie*

| Sentence | Ours (GloVe) | Ours (BERT) | ROGUE-1 | ROGUE-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 |

# Experiments: IV

- Reference sentence $r$: *i am on leave today*
- Predicted sentence $s_1$: *this is my day off*
- Predicted sentence $s_2$: *i am not going to work today*

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

# Results

- Our metric SemAnte captures the similarity scores based on semantic meaning of words better than the ROGUE metric, which only takes the lexical appearance into account
- Time complexity for comparing each word $w$ from reference sentence $s$ and each word $v$ from predicted sentence $p$, is reduced from $\mathcal{O}(N^2) \rightarrow \mathcal{O}(N)$.
- According to our preliminary results, SemAnte performs better in detecting hallucinations as compared to ROGUE metric.

# Conclusion and Future Work

- SemAnte captures the underlying semantic similarity to get a score better than ROGUE metric.
- For the task of detecting hallucinations, our metric performs better in most cases than the ROGUE metric.
- To extend this further, we wish to understand the performance of our metric on paragraphs rather than single sentences. We could achieve this using public datasets.
- Another possible future direction would be use a more complex BERT model to test SemAnte.
- For our metric, we have neglected the attention that flows into the [SEP] token. For our report, we wish to include an ablation study that compares the result of considering those attentions as well.

# References


Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," *ACM Comput. Surv.*, vol. 55, mar 2023.

H. Kang, T. Blevins, and L. Zettlemoyer, "Comparing hallucination detection metrics for multilingual generation," 2024.

L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. Liu, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," 2023.

T. Liu, Y. Zhang, C. Brockett, Y. Mao, Z. Sui, W. Chen, and B. 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.

T. Sellam, D. Das, and A. 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.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017.

F. Nan, R. Nallapati, Z. Wang, C. Nogueira dos Santos, H. Zhu, D. Zhang, K. McKeown, and B. 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.

T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013.

J. Pennington, R. Socher, and C. 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.