Consistency text similarity on the example of the task of recognizing hallucinations of language models

Kseniia Petrushina

Moscow Institute of Physics and Technology,
Skolkovo Institute of Science and Technology

Scientific supervisor: Alexander Panchenko

2023

Introduction

Explanation

Hallucination of the language model is a grammatically correctly generated response, which, however, contains incorrect information.

Examples

Paraphrase generation & Machine translation – different meaning.
Definition modeling – deviation from the database.

Problem statement

Language model is a function

$$\mathbf{f} : \mathcal{P}({\mathbf{T_s}}^{L_s}) \to \mathcal{P}({\mathbf{T_h}}^{L_h}),$$

$\mathbf{s}_i$ is called *source sentence* and $\mathbf{h}_i$ is called *model hypothesis*.
We can define function

$$\mathbf{f}^{-1} : \mathcal{P}({\mathbf{T_h}}^{L_h}) \to \mathcal{P}({\mathbf{T_s}}^{L_s})$$

Then it is said that $\mathbf{h} = \mathbf{f}(\mathbf{s})$ is a *hallucination* of the language model $\mathbf{f}$ with the input $\mathbf{s}$ if

$$p(\mathbf{f}^{-1}(\mathbf{f}(\mathbf{s})) = \mathbf{s}) = 0.$$

Problem statement

The task of recognizing hallucinations is to find a function $sim : \mathbf{T_s}^{L_s} \times \mathbf{T_h}^{L_h} \to [0, 1]$, such that

$$\mathbb{E}_{\mathbf{s}_i \sim \mathbf{T_s}^{L_s}, \mathbf{h}_i \sim f(\mathbf{s}_i)} \{\mathbb{I}[sim(\mathbf{s}_i, \mathbf{h}_i) \geq \mathrm{thr}] = y_i\} \to \max_{sim, \mathrm{thr}},$$

where y_i denotes the presence of a hallucination.

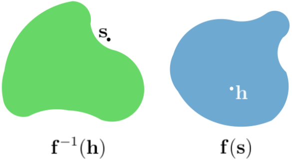

Figure: An illustration of a model's hallucination. $\mathbf{s}$ does not belong to the set of possible outputs of $\mathbf{f}^{-1}(\mathbf{h})$

Existing solutions

1. **Words or characters n-grams**

$$sim_{\text{BLEU}}(\mathbf{s}, \mathbf{h}) = \frac{|N_s \cap N_h|}{|N_h|}$$

2. **Similarity between static embeddings**

$$sim_{\cos}(\mathbf{s}, \mathbf{h}) = \cos(\mathbf{v_s}, \mathbf{v_h})$$

3. **Similarity between contextualized embeddings**

$$R = \frac{1}{L_s} \sum_{v_i \in \mathbf{v_s}} \max_{\hat{v}_j \in \mathbf{v_h}} v_i^T \hat{v}_j \ P = \frac{1}{L_h} \sum_{\hat{v}_j \in \mathbf{v_h}} \max_{v_i \in \mathbf{v_s}} v_i^T \hat{v}_j$$

$$\text{BERTScore} = 2 \frac{PR}{P + R}$$

Existing solutions

4. **Similarity between embeddings from bi-encoders**

$$sim_{\text{bi-enc}}(\mathbf{s}, \mathbf{h}) = \cos(\mathbf{enc_s}(\mathbf{s}), \mathbf{enc_h}(\mathbf{h}))$$

5. **Symmetric and asymmetric cross-encoders**

$$sim_{\text{cross-enc}}(\mathbf{s}, \mathbf{h}) = \mathbf{clf}(\mathbf{enc}(\mathbf{s}, \mathbf{h}))$$

In the general case, the similarity function should be defined for objects from different spaces ${\mathbf{T_s}}^{L_s}$ and ${\mathbf{T_h}}^{L_h}$.
The existing methods do not investigate whether there is enough information in $\mathbf{h}$ to restore $\mathbf{s}$.

Consistency similarity

Consistency similarity measure is defined by

$$sim_{\mathbf{C}}(\mathbf{s}_i, \mathbf{h}_i) = sim(\mathbf{s}_i, \mathbf{f}^{-1}(\mathbf{h}_i))).$$

1. The arguments of the function lie in the same space ${\mathbf{T_S}}^{L_s}$.
2. This similarity measure depends on how well the information about $\mathbf{s}_i$ is preserved in $\mathbf{h}_i$ and what the model $\mathbf{f}^{-1}$ can restore.

Hypothesis

1. $$\mathbb{E}_{\mathbf{f}^{-1}(\mathbf{h}_i)} sim_{\mathbf{C}}(\mathbf{s}_i, \mathbf{h}_i) \leq sim(\mathbf{s}_i, \mathbf{h}_i)$$
2. $$\mathbb{E}_{\mathbf{s}_i \sim {\mathbf{T_s}}^{L_s}, \mathbf{h}_i \sim f(\mathbf{s}_i)} \{\mathbb{I}[sim_C(\mathbf{s}_i, \mathbf{h}_i) \geq thr] = y_i\} \geq$$
$$\mathbb{E}_{\mathbf{s}_i \sim {\mathbf{T_s}}^{L_s}, \mathbf{h}_i \sim f(\mathbf{s}_i)} \{\mathbb{I}[sim(\mathbf{s}_i, \mathbf{h}_i) \geq thr] = y_i\}$$

Computational experiment

We are given the dataset

$$\mathcal{D} = \{(\mathbf{s}_i, \mathbf{h}_i, y_i)\}_{i=1}^{N}, \quad \mathbf{h}_i \in \mathbf{f}(\mathbf{s}_i), \quad y_i \in \{0, 1\}$$

The target variable y_i indicates the occurrence of a hallucination in the $\mathbf{f}$ model at the input of $\mathbf{s}_i$ and the output of $\mathbf{h}_i$.

1. The proportion of correct predictions:

$$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}[\hat{y}_i \geq \text{thr}] = y_i$$

2. Spearman's rank correlation coefficient:

$$r_s = \rho_{R(Y),R(\hat{Y})} = \frac{\text{cov}(R(Y), R(\hat{Y}))}{\sigma_{R(Y)}\sigma_{R(\hat{Y})}}$$

Results

- Paraphrase generation task

Method	Accuracy ↑	r_s ↑
$sim_{\text{bi-enc}}$	0.808	0.153
sim_{C}	**0.824**	**0.186**

Table: Hallucination recognition results in the PG task

- Machine translation task

Method	Accuracy ↑	r_s ↑
sim_{LaBSE}	0.786	0.592
$sim_{\text{BLASER-QE}}$	**0.802**	**0.605**

Table: Hallucination recognition results in the MT task

Conclusion

- Analysis of the existing measures of textual similarity.
- New method that corrects the disadvantages of the previous ones.
- Hypotheses about properties of the *consistency similarity measure*.
- Computational experiments with different measures.

Future work:

- Theoretical justification of hypotheses.
- Extend the method to work with an external database
- Conduct comprehensive ablation study.

Literature

1. David Dale et al. 2023. HalOmi: A Manually Annotated Benchmark for Multilingual Hallucination and Omission Detection in Machine Translation.
   URL: https://arxiv.org/pdf/2305.11746.pdf
2. Nikolay Babakov et al. 2022. A large-scale computational study of content preservation measures for text style transfer and paraphrase generation.
   URL: https://aclanthology.org/2022.acl-srw.23.pdf
3. Ashish Vaswani et al. 2017. Attention is All you Need
   URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf