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

**Kseniia Petrushina**
petrushina.ke@phystech.edu

## Abstract

This work explores the recognition of hallucinations of language models in the tasks of machine translation, paraphrasing and modeling definitions. Approaches to solving this problem are based on measures of semantic similarity between sentences. We compare different approaches to the solution, and also propose a new measure of proximity - *consistency similarity*.

## 1 Introduction

Currently, there is a real boom in language models that are able to perform various tasks of Natural Language Generation (NLG).

A hallucination of the language model is a grammatically correctly generated response, which, however, contains incorrect information. And currently, language models apt to give a fluent answer, in particular due to the fact that they adjusted to metrics that pay attention to the sonority of the answer, rather than its factual component. In the tasks of paraphrasing and machine translation, hallucination is different meanings in the generated and original sentences. Also, hallucinations may be discrepancies between the model's response and the actual data from the external knowledge base.

## 2 Problem statement

*Language model* is a function

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

where $\mathbf{s} \in {\mathbf{T_s}}^{L_s}$ is a sequence of $L_s$ tokens from the overall set of source tokens $\mathbf{T_s}$, $\mathbf{h} \in {\mathbf{T_h}}^{L_h}$ is a sequence of $L_h$ tokens from the overall set of hypothesis tokens $\mathbf{T_h}$. And $\mathcal{P}(\mathbf{T}^L)$ is the probability distribution over $\mathbf{T}^L$. $\mathbf{s}_i$ is called *source sentence* and $\mathbf{h}_i$ is called *model hypothesis*.

Assuming $\mathbf{f}$ has the following output probabilities

$$\mathcal{P}(\mathbf{f}(\mathbf{s}) = \mathbf{h}) = p_{\mathbf{sh}}$$

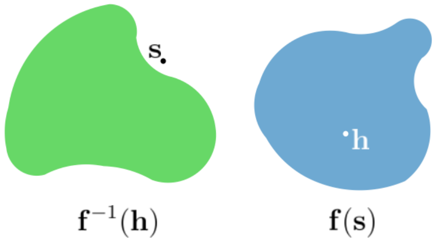

Figure 1: An illustration of a model's hallucination. $\mathbf{s}$ does not belong to the set of possible outputs of $\mathbf{f}^{-1}(\mathbf{h})$

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

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

This is illustrated in the Figure 1.

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

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

where $y_i$ denotes the presence of a hallucination.

The function $sim(\mathbf{s}, \mathbf{h})$ is perceived as similarity measure between source sentence $\mathbf{s}$ and model hypothesis $\mathbf{h}$.

## 3 Theory

### 3.1 Existing solutions

The tasks of paraphrasing and textual style transfer rely on measuring similarity between sentences (Babakov et al., 2022). This is done using different functions $sim : {\mathbf{T_s}}^{L_s} \times {\mathbf{T_h}}^{L_h} \to [0, 1]$.

These content preserving metrics can be divided into the following groups:

**Words or characters n-grams** The similarity of the sentences can be calculated based on n-gramms, i.e. all possible subsequences of s and $\mathbf{h}$ of length $n$.

$$N_s \subset \mathbf{T_s}^n, \quad N_h \subset \mathbf{T_h}^n$$

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

**Similarity between static embeddings** A vector space $\mathbb{R}^d$ can be a more convenient representation of the text for calculating distances. So, the **emb** : $\mathbf{T_s} \cup \mathbf{T_h} \to \mathbb{R}^d$ function is introduced. Vector from the sentence s will be obtained according to the following formula:

$$\mathbf{v_s} = \frac{1}{L_s} \sum_{i=1}^{L_s} \mathbf{emb}(\mathbf{s}[i])$$

Then the similarity is:

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

**Similarity between contextualized embeddings** Vector representation of a token can depend on the context, i.e. surrounding tokens. Contextualized embeddings are obtained using model $\mathbf{enc} : \mathbf{T_s}^{L_s} \to \mathbb{R}^{L_s \times d}$. $\mathbf{v_s} = \{v_1, \dots, v_{L_s}\}$, $\mathbf{v_h} = \{\hat{v}_1, \dots, \hat{v}_{L_h}\}$. BERTScore (Zhang et al., 2020) is calculated as

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

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

**Similarity between embeddings from bi-encoders** Encoder models $\mathbf{enc_s} : \mathbf{T_s}^{L_s} \to \mathbb{R}^d$, $\mathbf{enc_h} : \mathbf{T_h}^{L_h} \to \mathbb{R}^d$ implicitly get a vector representations of the entire sentences. The resulting similarity can be calculated as

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

An example of such encoder model is LaBSE (Feng et al., 2022).

**Symmetric and asymmetric cross-encoders** Encoder model $\mathbf{enc} : \mathbf{T_s}^{L_s} \times \mathbf{T_h}^{L_h} \to \mathbb{R}^d$, uses cross-attention mechanism (Vaswani et al., 2017) when processing two texts at the same time. It is symmetrical if $\mathbf{enc}(\mathbf{s}, \mathbf{h}) = \mathbf{enc}(\mathbf{h}, \mathbf{s})$, otherwise it is asymmetrical, this property is determined by the NLG problem being solved. The $\mathbf{clf} : \mathbb{R}^d \to [0, 1]$ function is also introduced, which determines the value of the similarity measure.

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

### 3.2 Proposed method

First, in the general case, the similarity function should be defined for objects from different spaces $\mathbf{T_s}^{L_s}$ and $\mathbf{T_h}^{L_h}$. Secondly, the existing methods do not investigate whether there is enough information in $\mathbf{h}$ to restore s. Therefore, we suggest using a new metric - *consistency similarity measure*.

Consistency similarity measure is defined by

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

where $sim$ can be arbitrary method from the existing solutions.

The benefits of the proposed formula are:

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**

   $$\mathbb{E}_{\mathbf{f}^{-1}(\mathbf{h}_i)} sim_{\mathbf{C}}(\mathbf{s}_i, \mathbf{h}_i) \leq sim(\mathbf{s}_i, \mathbf{h}_i)$$

**Hypothesis 2** *Consistency similarity measure $sim_C$ will be more stable and will give higher accuracy values.*

## 4 Computational experiments

### 4.1 Data

We are given the dataset

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

The target variable $y_i \in \{0, 1\}$ defines a binary relation on the subset of the set $\mathbf{T_s}^{L_s} \times \mathbf{T_h}^{L_h}$ represented in $\mathcal{D}$ and, in our formulation, it indicates the occurrence of a hallucination in the $\mathbf{f}$ model at the input of $\mathbf{s}_i$ and the output of $\mathbf{h}_i$.

### 4.2 Metrics

Given similarity predictions $\hat{y}_i = sim(\mathbf{s}_i, \mathbf{h}_i)$ the quality metrics in the hallucination recognition task are

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:

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

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

Table 1: Hallucination recognition results in the PG task

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

Table 2: Hallucination recognition results in the MT task

### 4.3 Paraphrase generation

The detection of hallucinations in the paraphrase generation (PG) task assumes that the sentences $\mathbf{s}_i$, $\mathbf{h}_i$ are in the same space $\mathbf{T}^N$.

The results of the experiments are presented in the Table 1. $sim_{\text{bi-enc}}$ uses the Sentence Transformer (Reimers and Gurevych, 2019) as a text encoder. Consistency similarity $sim_{\text{C}}$ is calculated like $sim_{\text{bi-enc}}(\mathbf{s}_i, \mathbf{f}^{-1}(\mathbf{h}_i))$, since in the task of paraphrasing $\mathbf{f}^{-1} \equiv \mathbf{f}$.

### 4.4 Machine translation

In the machine translation (MT) task, texts are written in various languages. In this problem, $\mathbf{s}_i$ is in Russian and $\mathbf{h}_i$ is in English.

The results of the experiments are presented in the Table 2. The considered methods also use bi-encoders, vector representations are obtained using LaBSE and SONAR (Duquenne et al., 2023) models. SONAR embeddings subsequently are processed with BLASER model.

$$sim_{\text{LaBSE}} = cos(\mathbf{enc}_L(\mathbf{s}_i), \mathbf{enc}_L(\mathbf{h}_i))$$

$$sim_{\text{BLASER-QE}} = \mathbf{clf}_{\text{B}}(\mathbf{enc}_S(\mathbf{s}_i), \mathbf{enc}_S(\mathbf{h}_i))$$

## 5 Conclusion

We conducted a review of existing methods for investigating the proximity between two texts, and also proposed a new method - consistency similarity, which draws attention to the correspondence of the semantic content of model hypothesis to information in source sentence.

## References

Nikolay Babakov, David Dale, Varvara Logacheva, and Alexander Panchenko. 2022. A large-scale computational study of content preservation measures for text style transfer and paraphrase generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 300–321, Dublin, Ireland. Association for Computational Linguistics.

Paul-Ambroise Duquenne, Holger Schwenk, and Benoît Sagot. 2023. Sonar: Sentence-level multimodal and language-agnostic representations. *ArXiv*, abs/2308.11466.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.