Optimal Transport for Unsupervised Hallucination Detection in Neural Machine Translation

Nuno M. Guerreiro[1,2] **Pierre Colombo**[4] **Pablo Piantanida**[5] **André F. T. Martins**[1,2,3]
[1]Instituto de Telecomunicações, Lisbon, Portugal
[2]Instituto Superior Técnico, University of Lisbon, Portugal [3]Unbabel, Lisbon, Portugal
[4]MICS, CentraleSupélec, Université Paris-Saclay
[5]ILLS - Université Paris-Saclay - CentraleSupélec
miguelguerreironuno@gmail.com andre.t.martins@tecnico.ulisboa.pt
{pierre.colombo, pablo.piantanida}@centralesupelec.fr

Abstract

Neural machine translation (NMT) has become the de-facto standard in real-world machine translation applications. However, NMT models can unpredictably produce severely pathological translations, known as hallucinations, that seriously undermine user trust. It becomes thus crucial to implement effective preventive strategies to guarantee their proper functioning. In this paper, we address the problem of hallucination detection in NMT by following a simple intuition: as hallucinations are detached from the source content, they exhibit encoder-decoder attention patterns that are statistically different from those of good quality translations. We frame this problem with an optimal transport formulation and propose a fully unsupervised, plug-in detector that can be used with any attention-based NMT model. Experimental results show that our detector not only outperforms all previous model-based detectors, but is also competitive with detectors that employ large models trained on millions of samples.

1 Introduction

Neural machine translation (NMT) has achieved tremendous success (Vaswani et al., 2017; Akhbardeh et al., 2021), becoming the mainstream method in real-world applications and production systems for automatic translation. Although these models are becoming evermore accurate, especially in high-resource settings, they may unpredictably produce *hallucinations*. These are severely pathological translations that are detached from the source sequence content (Lee et al., 2018; Müller et al., 2020; Raunak et al., 2021; Guerreiro et al., 2022). Crucially, these errors have the potential to seriously harm user trust in hard-to-predict ways (Perez et al., 2022), hence the evergrowing need to develop security mechanisms. One appealing strategy to address this issue is to develop effective on-the-fly detection systems.

In this work, we focus on leveraging the encoder-decoder attention mechanism to develop an on-the-fly hallucination detector. This mechanism is responsible for selecting and combining the information contained in the source sequence that is relevant to retain during translation. Thus, as hallucinations are translations which content is detached from the source sequence, it is no surprise that connections between *anomalous* attention patterns and hallucinations have been drawn before in the literature (Lee et al., 2018; Raunak et al., 2021; Guerreiro et al., 2022; Ferrando et al., 2022). Those patterns usually exhibit very scattered source attention mass across the different tokens in the translation (e.g. most source attention mass is concentrated on a few irrelevant tokens such as punctuation and the end-of-sequence token; Berard et al. 2019). Inspired by such observations, previous work has designed *ad-hoc* heuristics to detect hallucinations that specifically target the anomalous maps. While such heuristics can be used to detect hallucinations to a satisfactory extent (Guerreiro et al., 2022), we argue that a more theoretically-founded way of using anomalous attention information for hallucination detection is lacking in the literature.

Rather than aiming at finding particular patterns, we go back to the main definition of hallucinations and draw the following hypothesis: as hallucinations—contrary to good translations—are not supported by the source content, they may exhibit encoder-decoder attention patterns that are different from those found in good quality translations. Based on this hypothesis, we approach the problem of hallucination detection as a problem of anomaly detection with an **optimal transport (OT) formulation** (Kantorovich, 2006; Peyré et al., 2019). Namely, we aim to find translations with source attention mass distributions that are highly distant from those of good translations. Intuitively, the more distant a translation's attention patterns are from those of good translations, the

more **anomalous** it is in light of that distribution.

Our key contributions can be summarized as follows:

- We propose an OT-inspired fully unsupervised hallucination detector that can be plugged into any attention-based NMT model;
- We find that it is possible to leverage the idea that attention maps for hallucinations are anomalous in light of a reference data distribution to effectively detect hallucinations;
- We show that our detector not only outperforms all previous unsupervised model-based detectors, but is also competitive with detectors that employ large models that have been trained on millions of samples.

2 Background

2.1 Neural Machine Translation with Transformer Models

An autoregressive NMT model $\mathcal{M}$ defines a probability distribution $p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x})$ over an output space of hypotheses $\mathcal{Y}$ conditioned on a source sequence $\boldsymbol{x}$ contained in an input space $\mathcal{X}$ by estimating word by word the conditional distribution $p_{\boldsymbol{\theta}}(y_i|\boldsymbol{x}, \boldsymbol{y}_{<i})$. In this work, we focus on models parameterized by an encoder-decoder transformer model (Vaswani et al., 2017) with a set of learned weights $\boldsymbol{\theta}$. In particular, we will look closely at the multi-head encoder-decoder attention mechanism. This mechanism is responsible for computing, for each generation step, a distribution over all source sentence words that informs the decoder on the relevance of each word in the source sentence for the word to be translated in that step. Concretely, for a source sequence of arbitrary length n and a target sequence of arbitrary length m, given as input a matrix $\boldsymbol{Q} \in \mathbb{R}^{m\times d}$ containing d-dimensional representations for m queries, and matrices $\boldsymbol{K}, \boldsymbol{V} \in \mathbb{R}^{n\times d}$ for n keys and values, the *scaled dot-product attention* at a single head is computed as:

$$\mathsf{attn}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \underbrace{\mathsf{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^\top}{\sqrt{d}}\right)}_{\boldsymbol{\Omega} \in [0,1]^{m\times n}} \boldsymbol{V}, \quad (1)$$

where $\boldsymbol{\Omega}$ is the encoder-decoder attention matrix. Transformer networks employ multi-head attention, in which a representation for each head h is computed by evoking Equation 1 in parallel. Moreover, this process is repeated in every transformer layer. As such, for each transformer layer $\ell \in \{1, \ldots, L\}$, we obtain a different set of encoder-decoder attention matrices $\boldsymbol{\Omega}_{\ell,h}$ for every head $h \in \{1, \ldots, H\}$, in that layer.

Following previous work in the literature (Berard et al., 2019; Raunak et al., 2021; Guerreiro et al., 2022), we will focus on the last layer encoder-decoder attention. We will designate as attention maps those that are obtained by averaging across all heads of the last layer of the decoder module $\boldsymbol{\Omega}_L(\boldsymbol{x}) := \frac{1}{H}\sum_{h=1}^{H} \boldsymbol{\Omega}_{L,h}(\boldsymbol{x})$. Moreover, given the NMT model $\mathcal{M}$ we will designate $\boldsymbol{\pi}_{\mathcal{M}}(\boldsymbol{x}) = \frac{1}{m}\left[\boldsymbol{\Omega}_L(\boldsymbol{x})\right]^\top \mathbf{1} \in \triangle_n$ as the source (attention) mass distribution computed by $\mathcal{M}$ when $\boldsymbol{x}$ is presented as input, where $\triangle_n = \{\boldsymbol{p} \in \mathbb{R}^n \,|\, \boldsymbol{p} \geq \mathbf{0}, \mathbf{1}^\top \boldsymbol{p} = 1\}$ is the $(n-1)$-dimensional probability simplex.

2.2 Optimal Transport Problem and Wasserstein Distance

The first-order Wasserstein distance between two arbitrary probability distributions $\boldsymbol{\mu} \in \triangle_n$ and $\boldsymbol{\nu} \in \triangle_m$ is defined as

$$W(\boldsymbol{\mu}, \boldsymbol{\nu}) = \inf_{\boldsymbol{\gamma} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})} \mathbb{E}_{(u,v)\sim\boldsymbol{\gamma}}\left[c(u, v)\right], \quad (2)$$

where $c : [n] \times [m] \to \mathbb{R}_0^+$ is a cost function[1], and $\Pi(\boldsymbol{\mu}, \boldsymbol{\nu}) = \{\boldsymbol{\gamma} \in \triangle_{n\times m} : \boldsymbol{\gamma}\mathbf{1} = \boldsymbol{\mu}; \boldsymbol{\gamma}^\top \mathbf{1} = \boldsymbol{\nu}\}$[2] is the set of all joint probability distributions whose marginals are $\boldsymbol{\mu}, \boldsymbol{\nu}$. The Wasserstein distance arises from the method of optimal transport (OT) (Kantorovich, 2006; Peyré et al., 2019): OT measures distances between distributions in a way that depends on the geometry of the sample space. Intuitively, this distance indicates how much probability mass must be transferred from $\boldsymbol{\mu}$ to $\boldsymbol{\nu}$ in order to transform $\boldsymbol{\mu}$ into $\boldsymbol{\nu}$ while minimizing the transportation cost defined by c.

A notable example is the Wasserstein-1 distance, W_1, also known as Earth Mover's Distance (EMD), obtained with the choice $c(u, v) = \|u - v\|_1$. The name follows from the simple intuition: if the distributions are interpreted as "two piles of mass" that can be moved around, the EMD represents the minimum amount of "work" required to transform one pile into the other, where the work is defined as the amount of mass moved multiplied by the distance it is moved.

[1] We denote the set of indices $\{1, \ldots, n\}$ by $[n]$.

[2] We extend the simplex notation for matrices representing joint distributions, $\triangle_{n\times m} = \{\boldsymbol{P} \in \mathbb{R}^{n\times m} : \boldsymbol{P} \geq \mathbf{0}, \mathbf{1}^\top \boldsymbol{P}\mathbf{1} = 1\}$.

2.3 Hallucinations in NMT

Hallucinations are translations that lie at the extreme end of NMT pathologies (Raunak et al., 2021). By definition, these translations contain content that is detached from the source sentence (Raunak et al., 2021; Guerreiro et al., 2022). To disentangle the different types of hallucinations, they can be categorized as: *largely fluent detached hallucinations* or *oscillatory hallucinations* (Raunak et al., 2021; Guerreiro et al., 2022). The former are translations that bear *little or no relation at all* to the source content and may be further split according to the severity of the detachment (e.g. strong or full detachment) while the latter are inadequate translations that contain erroneous repetitions of words and phrases. We illustrate in Appendix A the categories described above through examples of hallucinated outputs found in the hallucinations dataset introduced in Guerreiro et al. (2022).

3 Detection of Hallucinations in NMT

3.1 Categorization of On-the-Fly Detectors

On-the-fly hallucinations detectors are systems that can detect hallucinations without access to reference translations. These detectors are particularly relevant for real-world applications where references are not readily available. As such, in this work, we will not consider metrics that depend on a reference sentence (e.g. chrF (Popović, 2016), COMET (Rei et al., 2020a)). For an analysis on the performance of such metrics, please refer to Guerreiro et al. (2022).

Previous work on on-the-fly detection of hallucinations in NMT has primarily focused on two categories of detectors: external detectors and model-based detectors. External detectors employ large language models trained on millions, or even billions, of samples in a (self-)supervised way for related tasks such as quality estimation (QE) and cross-lingual embedding similarity (Raunak et al., 2021, 2022; Guerreiro et al., 2022). On the other hand, model-based detectors only require access to the NMT model that is generating the translations, and work by leveraging effectively relevant internal features such as model confidence and encoder-decoder attention maps (Guerreiro et al., 2022). These detectors are attractive due to their flexibility and low memory footprint, as they can very easily be plugged in on a vast range of NMT models without the need for additional training data or computing infrastructure. Moreover, Guerreiro et al. (2022) show that model-based detectors can be particularly predictive of hallucinations, outperforming quality estimation models and even performing on par with state-of-the-art reference-based metrics. In this work, we build upon this previous research to propose a new model-based detector that achieves even greater improvements over all previously proposed detectors.

3.2 Problem Statement

In this problem statement, we will focus specifically on model-based detectors that require obtaining internal features from a model $\mathcal{M}$. Building a hallucination detector generally falls down to finding a scoring function $s_{\mathcal{M}} : \mathcal{X} \rightarrow \mathbb{R}$ and a threshold $\tau \in \mathbb{R}$ to build a binary rule $g_{\mathcal{M}} : \mathcal{X} \rightarrow \{0, 1\}$. For a given test sample $\boldsymbol{x} \in \mathcal{X}$,

$$g_{\mathcal{M}}(\boldsymbol{x}) = \mathbb{1}\{s_{\mathcal{M}}(\boldsymbol{x}) > \tau\}. \tag{3}$$

If $s_{\mathcal{M}}$ is an anomaly score, $g_{\mathcal{M}}(\boldsymbol{x}) = 0$ implies that the model $\mathcal{M}$ generates a 'normal' translation for the source sequence $\boldsymbol{x}$, and $g_{\mathcal{M}}(\boldsymbol{x}) = 1$ implies that the model $\mathcal{M}$ generates a 'hallucination' for the source sequence $\boldsymbol{x}$.[3]

4 Unsupervised Hallucination Detection with Optimal Transport

Anomalous attention maps have been connected to the hallucinatory mode in several works (Lee et al., 2018; Raunak et al., 2021; Guerreiro et al., 2022). Our method builds on this idea and uses the Wasserstein distance (see Section 2.2) to estimate the cost of transforming a translation source mass distribution into a reference distribution. Intuitively, the higher the cost of such transformation, the more distant—and hence the more anomalous—the attention of the translation is with respect to that of the reference translation.

4.1 Wass-to-Unif: A data independent scenario

In this scenario, we only rely on the generated translation and its source mass distribution to decide whether the translation is an hallucination or not. For a given test sample $\boldsymbol{x} \in \mathcal{X}$:

1. We first obtain the source mass attention distribution $\boldsymbol{\pi}_{\mathcal{M}}(\boldsymbol{x}) \in \triangle_{|\boldsymbol{x}|}$;

[3] Hereinafter, we will omit the subscript $\mathcal{M}$ from all model-based scoring functions to ease notation effort.

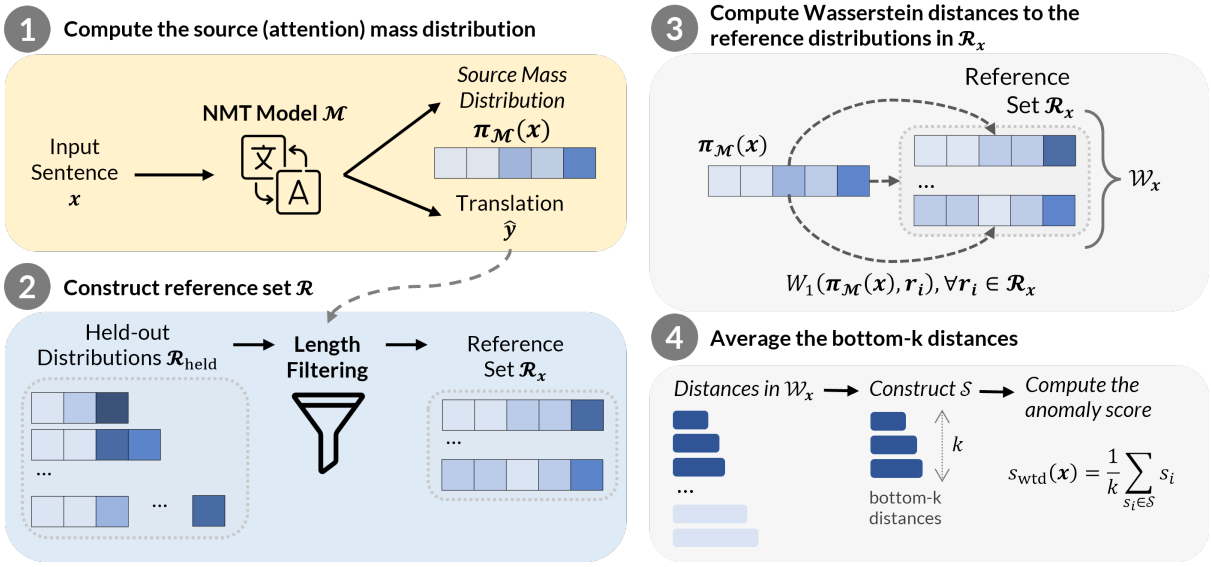

Figure 1: Procedure diagram for computation of the detection scores for the data-driven method Wass-to-Data.

2. We then compute an anomaly score, $s_{\text{wtu}}(\boldsymbol{x})$, by measuring the Wasserstein distance between $\boldsymbol{\pi}_{\mathcal{M}}(\boldsymbol{x})$ and a reference distribution $\boldsymbol{u}$:

$$s_{\text{wtu}}(\boldsymbol{x}) = W(\boldsymbol{\pi}_{\mathcal{M}}(\boldsymbol{x}), \boldsymbol{u}). \tag{4}$$

Choice of reference translation. A natural choice for $\boldsymbol{u}$ is the uniform distribution, $\boldsymbol{u} = \frac{1}{n} \cdot \mathbf{1}$, where $\mathbf{1}$ is a vector of ones of size n. In the context of our problem, a uniform source mass distribution means that all tokens are equally attended.

Choice of cost function. We consider the 0/1 cost function, $c(i, j) = \mathbf{1}[i \neq j]$, as it guarantees that the cost of transporting a unit mass from any token i to any token $j \neq i$ is constant. For this distance function, the problem in Equation 2 has the following closed-form solution (Villani, 2009):

$$W(\boldsymbol{\pi}_{\mathcal{M}}(\boldsymbol{x}), \boldsymbol{u}) = \frac{1}{2}\|\boldsymbol{\pi}_{\mathcal{M}}(\boldsymbol{x}) - \boldsymbol{u}\|_1. \tag{5}$$

This is a well-known result in optimal transport: the Wasserstein distance under the $0/1$ cost function is equivalent to the total variation distance between the two distributions. On this metric space, the Wasserstein distance depends solely on the probability mass that is transported to transform $\boldsymbol{\pi}_{\mathcal{M}}(\boldsymbol{x})$ to $\boldsymbol{u}$. Importantly, *this formulation ignores the starting locations and destinations of that probability mass* as the cost of transporting a unit mass from any token i to any token $j \neq i$ is constant.

Interpretation of Wass-to-Unif. Our method is inspired by observations made in Raunak et al. (2021); Guerreiro et al. (2022): attention maps for which the source attention mass is highly concentrated on a very sparse set of tokens (regardless of their location in the source sentence) can be predictive of hallucinations. Thus, the bigger the distance between the attention distribution of a test sample and the uniform distribution, the more peaked the former is, and the closer it is to maps that have been shown to be predictive of hallucinations.

4.2 Wass-to-Data: A data-driven scenario

In this scenario, instead of using a single reference distribution, we use a set of reference source mass distributions, $\mathcal{R}_{\boldsymbol{x}}$, obtained from the same model. This allows for a more accurate evaluation of how anomalous a given translation is compared to the data distribution, rather than relying on an arbitrary choice of reference distribution.

First, we use a held-out dataset $\mathcal{D}_{\text{held}}$ that contains samples for which the model $\mathcal{M}$ generates good quality translations according to an automatic evaluation metric (in this work, we use COMET (Rei et al., 2020a)). We use this dataset to construct (offline) a set of held-out source attention distributions $\mathcal{R}_{\text{held}} = \{\boldsymbol{\pi}_{\mathcal{M}}(\boldsymbol{x}) \in \triangle_{|\boldsymbol{x}|} : \boldsymbol{x} \in \mathcal{D}_{\text{held}}\}$. Then, for a given test sample $\boldsymbol{x} \in \mathcal{X}$:

1. We generate a translation $\hat{\boldsymbol{y}} = (y_1, \ldots, y_m)$ and obtain the source mass attention distribution $\boldsymbol{\pi}_{\mathcal{M}}(\boldsymbol{x}) \in \triangle_{|\boldsymbol{x}|}$;

2. We apply a length filter to construct the sample reference set $\mathcal{R}_{\boldsymbol{x}}$, by restricting $\mathcal{R}_{\boldsymbol{x}}$ to contain

source mass distributions of $\mathcal{R}_{\text{held}}$ correspondent to translations of size $[(1-\delta)m, (1+\delta)m]$ for a predefined $\delta \in]0, 1[$;[4]

3. We compute pairwise Wasserstein-1 distances between $\boldsymbol{\pi}_{\mathcal{M}}(\boldsymbol{x})$ and each element $\boldsymbol{r}_i$ of $\mathcal{R}_{\boldsymbol{x}}$:

$$\mathcal{W}_{\boldsymbol{x}} = \big(W_1(\boldsymbol{\pi}_{\mathcal{M}}(\boldsymbol{x}), \boldsymbol{r}_1), \ldots, W_1(\boldsymbol{\pi}_{\mathcal{M}}(\boldsymbol{x}), \boldsymbol{r}_{|\mathcal{R}_{\boldsymbol{x}}|})\big)\,. \tag{6}$$

4. We obtain the anomaly score $s_{\mathsf{wtd}}(\boldsymbol{x})$ by averaging the bottom-k distances in $\mathcal{W}_{\boldsymbol{x}}$:

$$s_{\mathsf{wtd}}(\boldsymbol{x}) = \frac{1}{k}\sum_{s_i \in \mathcal{S}} s_i, \tag{7}$$

where $\mathcal{S}$ is a sequence containing the k smallest elements of $\mathcal{W}_{\boldsymbol{x}}$.

This procedure is illustrated in Figure 1.

Interpretation of Wass-to-Data. Hallucinations, unlike good translations, are not fully supported by the source content. Wass-to-Data evaluates how anomalous a translation is by comparing the source attention mass distribution of that translation to those of good translations. The higher the Wass-to-Data score, the more anomalous the source attention mass distribution of that translation is in comparison to those of good translations, and the more likely it is to be an hallucination.

Relation to Wass-to-Unif. The Wasserstein-1 distance (see Section 2.2) between two distributions is equivalent to the ℓ_1-norm of the difference between their *cumulative distribution functions* (Peyré and Cuturi, 2018):

$$W_1(\boldsymbol{\pi}_{\mathcal{M}}(\boldsymbol{x}), \boldsymbol{r}) = \|\boldsymbol{A} \cdot (\boldsymbol{\pi}_{\mathcal{M}}(\boldsymbol{x}) - \boldsymbol{r})\|_1, \tag{8}$$

where $\boldsymbol{A} = [a_{ij}]$ where $a_{ij} = \mathbf{1}[i \leq j]$.[5] Note that this is different from the result in Equation 5, as the Wasserstein distance under $c(i,j) = \mathbf{1}[i \neq j]$ as the cost function is proportional to the norm of the difference between their *probability mass functions*. Thus, Wass-to-Unif will be more sensitive to the overall structure of the distributions (e.g. sharp probability peaks around some points), whereas Wass-to-Data will be more sensitive to the specific values of the points in the two distributions.

[4] For efficiency reasons, we set the maximum cardinality of $\mathcal{R}_{\boldsymbol{x}}$ to $|\mathcal{R}|_{\max}$. If $\mathcal{R}_{\boldsymbol{x}}$ contains more than $|\mathcal{R}|_{\max}$ samples, we restrict its cardinality by randomly sampling $|\mathcal{R}|_{\max}$ samples from $\mathcal{R}_{\boldsymbol{x}}$.

[5] When the distributions have different dimensions, we append a zero vector to the smallest one and compute Equation 8 using the modified distribution.

4.3 Wass-Combo: The best of both worlds

With this scoring function, we aim at combining Wass-to-Unif and Wass-to-Data into a single detector. To do so, we propose using a two-staged process that exploits the computational benefits of Wass-to-Unif over Wass-to-Data. Put simply, (i) we start by assessing whether a test sample is deemed an hallucination according to Wass-to-Unif, and if not (ii) we compute the Wass-to-Data score. Formally,

$$s_{\mathsf{wc}}(\boldsymbol{x}) = \mathbb{1}\left[s_{\mathsf{wtu}}(\boldsymbol{x}) > \tau_{\mathsf{wtu}}\right] \times \tilde{s}_{\mathsf{wtu}}(\boldsymbol{x}) + \mathbb{1}\left[s_{\mathsf{wtu}}(\boldsymbol{x}) \leq \tau_{\mathsf{wtu}}\right] \times s_{\mathsf{wtd}}(\boldsymbol{x}) \tag{9}$$

for a predefined scalar threshold τ_{wtu}. To set that threshold, we compute $\mathcal{W}_{\mathsf{wtu}} = \{s_{\mathsf{wtu}}(\boldsymbol{x}) : \boldsymbol{x} \in \mathcal{D}_{\text{held}}\}$ and set $\tau_{\mathsf{wtu}} = P_K$, i.e τ_{wtu} is the K^{th} percentile of $\mathcal{W}_{\mathsf{wtu}}$ with $K \in]98, 100[$ (in line with hallucinatory rates reported in (Müller et al., 2020; Xiao and Wang, 2021; Raunak et al., 2022)).[6]

5 Experimental Setup

5.1 Model and Data

We follow the setup in Guerreiro et al. (2022). In that work, the authors released a dataset of 3415 translations for WMT'18 DE-EN news translation data (Bojar et al., 2018) with structured annotations on critical errors and hallucinations. We use this dataset as it is the only available dataset that contains real hallucinations produced naturally by a NMT model. For all our experiments, we obtain all model-based information required to build the detectors using the same Transformer model that generated the translations in the dataset. We provide full details about the dataset and the model in Appendix A.

5.2 Baseline detectors

5.2.1 Model-based detectors

We compare our methods to the two best performing model-based methods in Guerreiro et al. (2022).

Attn-ign-SRC. This method consists in computing the proportion of source words with a total incoming attention mass lower than a threshold λ:

$$s_{\mathsf{ais}}(\boldsymbol{x}) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}\left[(\boldsymbol{\Omega}_L^{\top}(\boldsymbol{x})\mathbf{1})_j < \lambda\right]. \tag{10}$$

[6] In order to make the scales of s_{wtu} and s_{wtd} compatible, we use a scaled $\tilde{s}_{\mathsf{wtu}}$ value instead of s_{wtu} in Equation 9. We obtain $\tilde{s}_{\mathsf{wtu}}$ by min-max scaling s_{wtu} such that $\tilde{s}_{\mathsf{wtu}}$ is within the range of s_{wtd} values obtained for a held-out set.

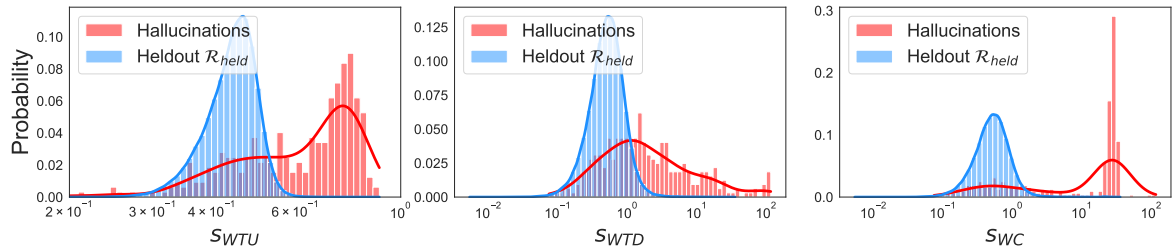

Figure 2: Histogram scores for our methods – Wass-to-Unif (left), Wass-to-Data (center) and Wass-Combo (right). We display Wass-to-Data and Wass-Combo scores on log-scale.

This method was initially proposed in Berard et al. (2019) and is effective as a detector of fully detached hallucinations (Guerreiro et al., 2022). We follow their work and use $\lambda = 0.2$.

Seq-Logprob. This method consists in computing the length-normalised sequence log-probability of the generated translation:

$$s_{\text{slp}}(\boldsymbol{x}) = \frac{1}{m}\sum_{k=1}^{m} \log p_{\boldsymbol{\theta}}(y_k \mid \boldsymbol{y}_{<k}, \boldsymbol{x}). \qquad (11)$$

This method has been shown to be surprisingly effective for detection of fully and strongly detached hallucinations (Guerreiro et al., 2022).

5.2.2 External detectors

Although the main focus of our work is plug-in methods that only rely on model-based features, we provide a comparison to external detectors. This comparison provides an estimate of the upper-bound in performance that is helpful to monitor the development of new model-based detectors.

CometKiwi. We compute the sentence-level quality assessments provided by the reference-free model CometKiwi (Rei et al., 2022), the winner of all tracks on the WMT'22 shared-task on Quality Estimation (Zerva et al., 2022). This model has been trained on nearly one million direct assessment annotations. Importantly, contrary to previous reference-free COMET models, the training data of CometKiwi includes human annotations from the MLQE-PE dataset (Fomicheva et al., 2022) that has been shown to include several low-quality translations and hallucinations (Specia et al., 2021; Tang et al., 2022). In Appendix C, we show that CometKiwi is significantly more predictive of hallucinations than the previous reference-free COMET versions (Rei et al., 2020b) studied in (Raunak et al., 2022; Guerreiro et al., 2022).

LaBSE. This methods consists in leveraging LaBSE (Feng et al., 2020) to compute cross-lingual sentence representations for the source sequence and translation. We use the cosine similarity of these representations as the detection score. This model was trained on billions of samples in a self-supervised way. LaBSE makes for a very good baseline, as it was optimized with a translate matching training objective that is very much aligned with the task of hallucination detection: during training, LaBSE is given a source sequence and a set of translations including the true translation and multiple negative alternatives, and the model is optimized to specifically discriminate the true translation from the other negative alternatives by assigning a higher similarity score to the former.

5.3 Evaluation metrics

We will use the Area Under the Receiver Operating Characteristic curve (AUROC) and the False Positive Rate at 90% True Positive Rate (FPR@90TPR) to evaluate the performance of different detectors. As these are threshold-independent evaluation metrics, they provide a more comprehensive view of the performance of an anomaly detector without being influenced by the choice of threshold.

5.4 Implementation Details

We use the library `POT: Python Optimal Transport` (Flamary et al., 2021) to compute our detector's scores. We use WMT'18 German-English data samples from the held-out set used in Guerreiro et al. (2022), and construct $\mathcal{D}_{\text{held}}$ to contain the 250k samples with highest COMET score. To obtain Wass-to-Data scores, we set $\delta = 0.1$, $|\mathcal{R}|_{\max} = 1000$ and $k = 4$. To obtain Wass-to-Combo scores, we set $\tau_{\text{wtu}} = P_{99.9}$. We perform ablations on $\mathcal{R}_{\text{held}}$ in Section 6.3, and on all other hyperparameters in Appendix D.

DETECTOR	AUROC ↑	FPR@90TPR ↓
External Detectors		
CometKiwi	86.96	53.61
LaBSE	**91.72**	**26.91**
Model-based Detectors		
Attn-ign-SRC	79.36	72.83
Seq-Logprob	83.40	59.02
OURS		
Wass-to-Unif	80.37	72.22
Wass-to-Data	84.20 ± 0.15	**48.15** ± 0.54
Wass-Combo	**87.17** ± 0.07	**47.56** ± 1.30

Table 1: Performance of all hallucination detectors. For Wass-to-Data and Wass-Combo we present the mean and standard deviation scores across five random seeds.

6 Results

6.1 Performance on on-the-fly detection

We start by analyzing the performance of our proposed detectors on a real world on-the-fly detection scenario. In this scenario, the detector must be able to flag hallucinations *regardless of their specific type* as those are unknown at the time of detection. Overall, we show that Wass-Combo is the best model-based detector. However, it still lags behind LaBSE, hinting that more powerful detectors can be built using model-based features.

Wass-Combo is the best model-based detector. Table 1 shows that Wass-Combo outperforms most other methods both in terms of AUROC and FPR. When compared to the previous best-performing unsupervised method (Seq-Logprob), Wass-Combo obtains boosts of approximately 4 and 10 points in AUROC and FPR, respectively. These performance boosts are further evidence that model-based features can be leveraged, in an unsupervised manner, to build effective plug-in detectors. Nevertheless, the high values of FPR suggest that *(i) the task is difficult for model-based detectors, and (ii) there is a significant performance margin to further reduce in future research.*

The notion of data proximity is helpful to detect hallucinations. Table 1 shows that Wass-to-Data outperforms the previous best-performing model-based method (Seq-Logprob) both in AUROC and FPR (by more than 10%). This supports the idea that encoder-decoder attention patterns for hallucinations are anomalous with respect to those of good model-generated translations, and that our method can effectively measure this level of anomalousness. On the other hand, compared to Wass-to-Uni, Wass-to-Data shows a significant improvement of 30 FPR points. This highlights the effectiveness of leveraging the data-driven distribution of good translations instead of the ad-hoc uniform distribution. Nevertheless, Table 1 and Figure 2 show that combining both methods brings further performance improvements. This suggests that these methods may specialize in different types of hallucinations, and that combining them allows for detecting a broader range of anomalies. We will analyze this further in Section 6.2.

Our model-based method achieves comparable performance to external models. Table 1 shows that Wass-Combo outperforms CometKiwi, with significant improvements on FPR. However, there still exists a gap to LaBSE, *the best overall detector*. This performance gap indicates that *more powerful detectors can be built, paving the way for future work in model-based hallucination detection.* Nevertheless, while relying on external models seems appealing, deploying them in practice usually comes with additional infrastructure costs, while our detector relies on information that can be obtained when generating the translation.

Translation quality assessments are less predictive than similarity of cross-lingual sentence representations. Table 1 shows that LaBSE outperforms the state-of-the-art quality estimation system CometKiwi, with vast improvements in terms of FPR. This shows that for the task of hallucination detection, quality assessments obtained with a QE model are less predictive than the similarity between cross-lingual sentence representations. This may be explained through their training objectives (see Section 5.2.2). While CometKiwi employs a more general regression objective in which the model is trained to match human quality assessments, LaBSE is trained with a translate matching training objective that is very closely related to the task of hallucination detection.

6.2 Do detectors specialize in different types of hallucinations?

In this section, we present an analysis on the performance of different detectors for different types of hallucinations (see Section 2.3). We report both a quantitative analysis to understand whether a detector can distinguish a specific hallucination type from other translations (Table 2), and a qualitative

DETECTOR	Fully Detached	Oscillatory	Strongly Detached
External Detectors			
CometKiwi	87.75	**93.04**	81.78
LaBSE	**98.91**	84.62	**89.72**
Model-based Detectors			
Attn-ign-SRC	95.76	59.53	77.42
Seq-Logprob	95.64	71.10	**80.15**
OURS			
Wass-to-Unif	96.35	69.75	72.19
Wass-to-Data	88.24 ± 0.29	**87.80** ± 0.10	77.60 ± 0.18
Wass-Combo	**96.57** ± 0.10	85.74 ± 0.10	78.89 ± 0.15

(a) AUROC – the higher the better.

DETECTOR	Fully Detached	Oscillatory	Strongly Detached
External Detectors			
CometKiwi	33.70	**23.80**	42.98
LaBSE	**0.52**	50.26	**28.88**
Model-based Detectors			
Attn-ign-SRC	8.51	81.24	76.68
Seq-Logprob	4.62	72.99	**65.39**
OURS			
Wass-to-Unif	**3.27**	78.78	88.32
Wass-to-Data	36.60 ± 1.92	**40.04** ± 1.57	**63.96** ± 2.04
Wass-Combo	3.56 ± 0.00	**41.38** ± 1.59	**64.55** ± 1.93

(b) FPR@90TPR (%) – the lower the better.

Table 2: Performance of all hallucination detectors for each hallucination type. For Wass-to-Data and Wass-Combo, we present the mean and standard deviation across five random seeds.

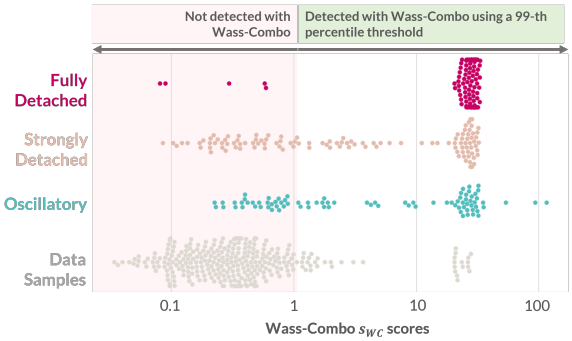

Figure 3: Distribution of Wass-Combo scores (on log-scale) for hallucinations and data samples, and performance on a fixed-threshold scenario.

analysis on a fixed-threshold scenario[7] (Figure 3). This analysis is particularly relevant to better understand how different detectors specialize in different types of hallucinations.

Fully detached hallucinations. Detecting fully detached hallucinations is remarkably easy for most detectors. Interestingly, Wass-to-Unif significantly outperforms Wass-to-Data on this type of hallucinations. This highlights how combining both methods can be helpful. In fact, Wass-Combo performs similarly to Wass-to-Unif, and can very easily separate most fully detached hallucinations from other translations on a fixed-threshold scenario (Figure 3). Note that the performance of Wass-to-Unif for fully detached hallucinations mirrors closely that of Attn-ign-SRC. This is not surprising, since both methods, at their core, are trying to capture similar patterns: translations for which the source attention mass distribution is highly concentrated on a small set of source tokens.

Strongly detached hallucinations. These are the hardest hallucinations to detect with our methods. Nevertheless, Wass-Combo performs very competitively with the previous best-performing model-based method for this type of hallucinations (Seq-Logprob). We hypothesize that the difficulty in detecting these hallucinations is due to the varying level of detachment from the source sequence. Indeed, Figure 3 shows that the scores of Wass-Combo exhibit some variance, spanning from a cluster of hallucinations with scores similar to other data samples (less detachment), to those with scores similar to most fully detached hallucinations (very high detachment).

Oscillatory hallucinations. Wass-to-Data and Wass-Combo significantly outperform all previous model-based detectors on detecting oscillatory hallucinations. This is of relevance in the context of model-based detectors, as previously best-performing detectors struggle with detecting these hallucinations. Moreover, Wass-to-Data also manages to outperform LaBSE with significant improvements in FPR. This hints that the repetition of words or phrases may not be enough to create sentence-level representations that are highly dissimilar from the non-oscillatory source sequence. In contrast, we find that CometKiwi appropriately penalize oscillatory hallucinations, which aligns with observations made in (Guerreiro et al., 2022).

We further analyze the performance of Wass-Combo scores for detection of oscillatory hallucinations. Figure 3 shows that, similarly to the scores

[7]We set the threshold by finding the 99-th percentile of Wass-Combo scores obtained for 100k samples from the clean WMT18 German-English held-out set (see Section 5.4).

for strongly detached hallucinations, the detector scores for oscillatory hallucinations are also scattered along a broad range of scores. After close manual evaluation, we observed that this is highly related to the severity of the oscillation: almost all non-detected hallucinations are not severe oscillations (see Appendix E).

6.3 Tracing-back performance boosts to the construction of the reference set $\mathcal{R}_{\boldsymbol{x}}$

In Section 6.1, we showed that evaluating how distant a given translation is compared to a data-driven reference distribution–rather than to an *ad-hoc* reference distribution– led to increased performance. Therefore, we will now analyze the construction of the reference set $\mathcal{R}_{\boldsymbol{x}}$ to obtain Wass-to-Data scores (step 2 in Figure 1). We conduct experiments to investigate the importance of the two main operations in this process: defining and length-filtering the distributions in $\mathcal{R}_{\text{held}}$.

Construction of $\mathcal{R}_{\text{held}}$. To construct $\mathcal{R}_{\text{held}}$, we first need to obtain the source attention mass distributions for each sample in $\mathcal{D}_{\text{held}}$. If $\mathcal{D}_{\text{held}}$ is a parallel corpus, we can force-decode the reference translations to construct $\mathcal{R}_{\text{held}}$. As shown in Table 3, this construction produces results similar to using good-quality model-generated translations. Moreover, we also evaluate the scenario where $\mathcal{R}_{\text{held}}$ is constructed with translations of any quality. Table 3 shows that although filtering for quality improves performance, the gains are not substantial. This connects to findings by Guerreiro et al. (2022): hallucinations exhibit different properties from other translations, including other incorrect translations. We offer further evidence that properties of hallucinations—in this case, the source attention mass distributions—are not only different to those of good-quality translations but also to most other model-generated translations.

Length-filtering the distributions in $\mathcal{R}_{\text{held}}$. The results in Table 4 show that length-filtering boosts performance significantly. This is expected: our translation-based length-filtering penalizes translations whose length is anomalous for their respective source sequences. This is particularly useful for detecting oscillatory hallucinations.

ABLATION	AUROC ↑	FPR@90TPR ↓
Model-Generated Translations		
Any	83.27 ± 0.39	50.08 ± 1.65
Quality-filtered	**84.20** ± 0.15	**48.15** ± 0.54
Reference Translations		
Any	**83.95** ± 0.16	50.26 ± 0.60

Table 3: Ablations on Wass-to-Data by changing the construction of $\mathcal{R}_{\text{held}}$. We present the mean and standard deviation across five random seeds.

ABLATION	AUROC ↑	FPR@90TPR ↓
Random Sampling	80.65 ± 0.15	57.06 ± 2.04
Length Filtering	**84.20** ± 0.15	**48.15** ± 0.54

Table 4: Ablations on Wass-to-Data by changing the length-filtering window to construct $\mathcal{R}_{\boldsymbol{x}}$. We present the mean and standard deviation across five random seeds.

7 Conclusions

We propose a novel plug-in model-based detector for hallucinations in NMT. Unlike previous attempts at building an attention-based detector, we do not design *ad-hoc* heuristics to detect hallucinations, and instead pose the problem of hallucination detection as an optimal transport problem: our detector aims at finding translations with source attention mass distribution that are highly distant from those of good quality translations. Through a rigorous comparison between different detectors, we show that our detector not only outperforms all previous model-based detectors, but is also competitive with detectors that use large, pre-trained models that have been trained on millions of samples. Importantly, our detector does not require any training data and, thanks to its flexibility, it can be easily deployed in real-world scenarios.

Acknowledgments

This work is partially supported by the European Research Council (ERC StG DeepSPIN 758969), by EU's Horizon Europe Research and Innovation Actions (UTTER, contract 101070631), by the FCT through contract UIDB/50008/2020, and by the P2020 programs MAIA and Unbabel4EU (LISBOA-01-0247-FEDER-045909 and LISBOA01-0247-FEDER-042671).

References

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Ro-

man Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Alexandre Berard, Ioan Calapodescu, and Claude Roux. 2019. Naver labs Europe's systems for the WMT19 machine translation robustness task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 526–532, Florence, Italy. Association for Computational Linguistics.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding.

Javier Ferrando, Gerard I. Gállego, Belen Alastruey, Carlos Escolano, and Marta R. Costa-jussà. 2022. Towards opening the black box of neural machine translation: Source and target interpretations of the transformer.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. 2021. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8.

Marina Fomicheva, Shuo Sun, Erick Fonseca, Chrysoula Zerva, Frédéric Blain, Vishrav Chaudhary, Francisco Guzmán, Nina Lopatina, Lucia Specia, and André F. T. Martins. 2022. MLQE-PE: A multilingual quality estimation and post-editing dataset. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4963–4974, Marseille, France. European Language Resources Association.

Nuno M. Guerreiro, Elena Voita, and André F. T. Martins. 2022. Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation.

Leonid V Kantorovich. 2006. On the translocation of masses. *Journal of mathematical sciences*, 133(4):1381–1382.

Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. 2018. Hallucinations in neural machine translation.

Mathias Müller, Annette Rios, and Rico Sennrich. 2020. Domain robustness in neural machine translation. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 151–164, Virtual. Association for Machine Translation in the Americas.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*.

Gabriel Peyré, Marco Cuturi, et al. 2019. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607.

Gabriel Peyré and Marco Cuturi. 2018. Computational optimal transport.

Maja Popović. 2016. chrF deconstructed: beta parameters and n-gram weights. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 499–504, Berlin, Germany. Association for Computational Linguistics.

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. The curious case of hallucinations in neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1172–1183, Online. Association for Computational Linguistics.

Vikas Raunak, Matt Post, and Arul Menezes. 2022. Salted: A framework for salient long-tail translation error detection.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020a. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020b. Unbabel's participation in the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 911–920, Online. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C. Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte M. Alves, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. Cometkiwi: Ist-unbabel 2022 submission for the quality estimation shared task.

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*. Association for Computational Linguistics.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Chrysoula Zerva, Zhenhao Li, Vishrav Chaudhary, and André F. T. Martins. 2021. Findings of the WMT 2021 shared task on quality estimation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 684–725, Online. Association for Computational Linguistics.

Joël Tang, Marina Fomicheva, and Lucia Specia. 2022. Reducing hallucinations in neural machine translation with feature attribution.

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.

Cédric Villani. 2009. *Optimal transport: old and new*, volume 338. Springer.

Yijun Xiao and William Yang Wang. 2021. On hallucination and predictive uncertainty in conditional language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2734–2744, Online. Association for Computational Linguistics.

Chrysoula Zerva, Frédéric Blain, Ricardo Rei, Piyawat Lertvittayakumjorn, José G. C. de Souza, Steffen Eger, Diptesh Kanojia, Duarte Alves, Constantin Orăsan, Marina Fomicheva, André F. T. Martins, and Lucia Specia. 2022. Findings of the WMT 2022 shared task on quality estimation. In *Proceedings of the Seventh Conference on Machine Translation*, pages 69–99, Abu Dhabi. Association for Computational Linguistics.

A Model and Data Details

Model. The model used in Guerreiro et al. (2022) to create the hallucination dataset is a Transformer base model (Vaswani et al., 2017) (hidden size of 512, feedforward size of 2048, 6 encoder and 6 decoder layers, 8 attention heads). The model has approximately 77M parameters. It was trained with the `fairseq` toolkit (Ott et al., 2019) on WMT18 DE-EN data (excluding Paracrawl): the authors randomly choose 2/3 of the dataset for training and use the remaining 1/3 as a held-out set for analysis. We use that same held-out set in this work.

Dataset Stats. The dataset used in this paper was introduced in Guerreiro et al. (2022). It consists of 3415 translations from WMT18 German-English data with structured annotations on different types of hallucinations and pathologies. Overall, the dataset contains 118 translations annotated as fully detached hallucinations, 90 as strongly detached hallucinations, and 86 as oscillatory hallucinations.[8] The other translations are either incorrect (1073) or correct (2048). Details on annotation, a high-level overview and other statistics can be found in the original paper. We show examples of hallucinations for each category in Table 5.

B Details on External Detectors

COMET. We use models available in the official repository[9]: `wmt22-cometkiwi-da` for CometKiwi and `wmt20-comet-da` for COMET.

LaBSE. We use the version available in `sentence-transformers` (Reimers and Gurevych, 2019).[10]

C Performance of reference-free COMET-based models

Guerreiro et al. (2022) used the COMET-QE version `wmt20-comet-qe-da`, whereas we are using the latest iteration `wmt22-cometkiwi-da` (CometKiwi). CometKiwi was trained on human annotations from the MLQE-PE dataset (Fomicheva et al., 2022), which contains a high percentage of hallucinations for some language pairs (Specia et al., 2021; Tang et al., 2022). We show the performance of both these versions in Table 6. CometKiwi significantly outperforms the previous iteration of COMET-QE. This hints that training quality estimation models with more negative examples can improve their ability to adequately penalize hallucinations.

D Ablations

We perform ablations on Wass-to-Data and Wass-Combo for all relevant hyperparameters: the length-filtering parameter δ, the maximum cardinality of $\mathcal{R}$, $|\mathcal{R}|_{\max}$, the value of k to compute the Wass-to-Data scores (see step 4 in Figure 1), and the threshold for Wass-to-Unif scores to compute Wass-Combo scores. The results are shown in Table 7 to Table 10, respectively. We also report in Table 11 the performance of Wass-to-Data with a 0/1 cost function instead of the ℓ_1-distance function.

E Error Analysis of Wass-Combo

We show a qualitative analysis on the same fixed-threshold scenario described in Section 6.2 in Figure 4. Differently to Figure 3, we provide examples of translations that have not been detected by Wass-Combo for the chosen threshold.

Our detector is not able to detect fully detached hallucinations that come in the form of exact copies of the source sentence. For these pathological translations, the attention map is mostly diagonal and is thus not anomalous. Although these are severe errors, we argue that, in a real-world application, such translations can be easily detected with string matching heuristics.

We also find that our detector Wass-Combo struggles with oscillatory hallucinations that come in the form of mild repetitions of 1-grams or 2-grams (see example in Figure 4). To test this hypothesis, we implemented the binary heuristic top n-gram count (Raunak et al., 2021; Guerreiro et al., 2022) to verify whether a translation is a severe oscillation: given the entire $\mathcal{D}_{\text{held}}$, a translation is flagged as an oscillatory hallucination if (i) it is in the set of 1% lowest-quality translations according to CometKiwi and (ii) the count of the top repeated 4-gram in the translation is greater than the count of the top repeated source 4-gram by at least 2. Indeed, more than 90% of the oscillatory hallucinations not detected by Wass-Combo in Figure 4 were not flagged by this heuristic. We provide 8 examples sampled from the set of oscillatory hallucinations

[8] Some strongly detached hallucinations have also been annotated as oscillatory hallucinations. In these cases, we consider them to be oscillatory.

[9] https://github.com/Unbabel/COMET

[10] https://huggingface.co/sentence-transformers/LaBSE

Category	Source Sentence	Reference Translation	Hallucination
Oscillatory	Als Maß hierfür wird meist der sogenannte Pearl Index benutzt (so benannt nach einem Statistiker, der diese Berechnungsformel einführte).	As a measure of this, the so-called Pearl Index is usually used (so named after a statistician who introduced this calculation formula).	The term "Pearl Index" refers to the term "Pearl Index" (or "Pearl Index") used to refer to the term "Pearl Index" (or "Pearl Index").
Strongly Detached	Fraktion der Grünen / Freie Europäische Allianz	The Group of the Greens/European Free Alliance	Independence and Democracy Group (includes 10 UKIP MEPs and one independent MEP from Ireland)
Fully Detached	Die Zimmer beziehen, die Fenster mit Aussicht öffnen, tief durchatmen, staunen.	Head up to the rooms, open up the windows and savour the view, breathe deeply, marvel.	The staff were very friendly and helpful.

Table 5: Examples of hallucination types. Hallucinated content is shown shaded.

Model Version	AUROC ↑	FPR@90TPR ↓
`wmt20-comet-qe-da`	70.15	57.24
`wmt22-cometkiwi-da`	**86.96**	**53.61**

Table 6: Performance of COMET-QE (`wmt20-comet-qe-da`) and CometKiwi (`wmt22-cometkiwi-da`) on the on-the-fly detection scenario.

Ablation	AUROC ↑	FPR@90TPR ↓
Random Sampling	80.65 ± 0.15	57.06 ± 2.04
Length Filtering ($\delta > 0$)		
$\delta = 0.1$	84.20 ± 0.15	48.15 ± 0.54
$\delta = 0.2$	84.37 ± 0.17	47.12 ± 1.04
$\delta = 0.3$	83.93 ± 0.18	48.45 ± 2.32
$\delta = 0.4$	83.06 ± 0.16	50.12 ± 1.29
$\delta = 0.5$	82.78 ± 0.34	50.89 ± 0.71

Table 7: Ablation on Wass-to-Data by changing the length-filtering window to construct $\mathcal{R}$. We present the mean and standard deviation across five random seeds.

not detected with Wass-Combo in Table 12.

Ablation	AUROC ↑	FPR@90TPR ↓
$\lvert\mathcal{R}\rvert_{\max} = 100$	82.99 ± 0.19	50.86 ± 0.95
$\lvert\mathcal{R}\rvert_{\max} = 500$	83.93 ± 0.08	48.07 ± 1.37
$\lvert\mathcal{R}\rvert_{\max} = 1000$	84.20 ± 0.15	48.15 ± 0.54
$\lvert\mathcal{R}\rvert_{\max} = 2000$	84.40 ± 0.14	49.23 ± 1.08
$\lvert\mathcal{R}\rvert_{\max} = 5000$	84.43 ± 0.13	48.05 ± 0.59

Table 8: Ablation on Wass-to-Data by changing the maximum cardinality of $\mathcal{R}$, $|\mathcal{R}|_{\max}$. We present the mean and standard deviation across five random seeds.

Ablation	AUROC ↑	FPR@90TPR ↓
Minimum	84.00 ± 0.33	52.03 ± 1.28
Bottom-k ($k > 1$)		
$k = 2$	84.25 ± 0.23	50.07 ± 0.70
$k = 4$	84.20 ± 0.15	48.15 ± 0.54
$k = 8$	83.99 ± 0.08	48.38 ± 1.10
$k = 16$	83.64 ± 0.04	48.05 ± 1.10
$k = 32$	83.23 ± 0.07	47.34 ± 0.94

Table 9: Ablation on Wass-to-Data by obtaining the score s_{wtd} by averaging the bottom-k distances in $\mathcal{R}$ for different values of k. We present the mean and standard deviation across five random seeds.

Ablation	AUROC ↑	FPR@90TPR ↓
$\tau = P_{99}$	85.79 ± 0.08	51.09 ± 0.97
$\tau = P_{99.5}$	86.34 ± 0.07	49.64 ± 1.71
$\tau = P_{99.9}$	87.17 ± 0.07	47.56 ± 1.30
$\tau = P_{99.99}$	84.69 ± 0.15	48.15 ± 0.54

Table 10: Ablation on Wass-Combo by obtaining the score s_{wc} for different scalar thresholds $\tau = P_K$ (K-th percentile of $\mathbb{W}_{\mathsf{wtu}}$). We present the mean and standard deviation across five random seeds.

Cost Function	AUROC ↑	FPR@90TPR ↓
ℓ_1 (Wasserstein-1)	84.20 ± 0.15	48.15 ± 0.54
$0/1$ cost	81.78 ± 0.20	51.72 ± 1.17

Table 11: Ablation on Wass-to-Data by changing the cost function in the computation of the Wasserstein Distances in Equation 6.

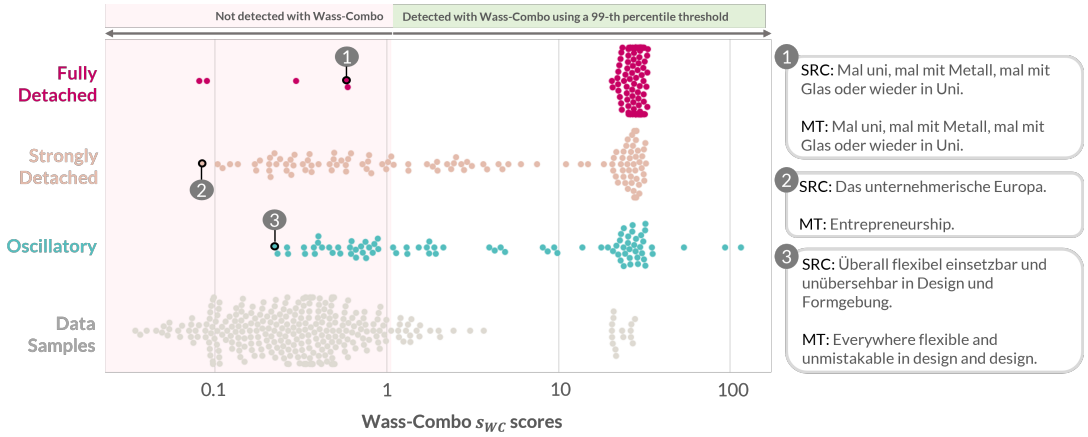

Figure 4: Distribution of Wass-Combo scores (on log-scale) for each type of hallucination and performance on a fixed-threshold scenario.

	OSCILLATORY HALLUCINATIONS NOT DETECTED WITH WASS-COMBO
SOURCE	Überall flexibel einsetzbar und unübersehbar in Design und Formgebung.
TRANSLATION	Everywhere flexible and unmistakable in design and design.
SOURCE	Um kahlen Stellen, wenn sie ohne Rüstung pg.
TRANSLATION	To dig dig digits if they have no armor pg.
SOURCE	Damit wird, wie die Wirtschaftswissenschaftler sagen, der Nennwert vorgezogen.
TRANSLATION	This, as economists say, puts the par value before the par value.
SOURCE	Besonders beim Reinigen des Verflüssigers kommt Ihnen dies zugute.
TRANSLATION	Especially when cleaning the liquefied liquefied liquefied.
SOURCE	Müssen die Verkehrsmittel aus- oder abgewählt werden ?
TRANSLATION	Do you need to opt-out or opt-out of transport?
SOURCE	Schnell drüberlesen - "Ja" auswählen und weiter gehts.
TRANSLATION	Simply press the "Yes" button and press the "Yes."
SOURCE	Auf den jeweiligen Dorfplätzen finden sich Alt und Jung zum Schwätzchen und zum Feiern zusammen.
TRANSLATION	Old and young people will find themselves together in the village's respective squares for fun and fun.
SOURCE	Zur Absicherung der E-Mail-Kommunikation auf Basis von PGP- als auch X.509-Schlüsseln hat die Schaeffler Gruppe eine Zertifizierungsinfrastruktur (Public Key Infrastructure PKI) aufgebaut.
TRANSLATION	The Schaeffler Group has set up a Public Key Infrastructure PKI (Public Key Infrastructure PKI) to secure e-mail communication based on PGP and X.509 keys.

Table 12: Examples of oscillatory hallucinations that have not been detected with Wass-Combo.