RAINPROOF: AN UMBRELLA TO SHIELD TEXT GENERATORS FROM OUT-OF-DISTRIBUTION DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

As more and more conversational and translation systems are deployed in production, it is essential to implement and to develop effective control mechanisms guaranteeing their proper functioning and security. An essential component to ensure safe system behavior is out-of-distribution (OOD) detection, which aims at detecting whether an input sample is statistically far from the training distribution. Although OOD detection is a widely covered topic in classification tasks, it has received much less attention in text generation. This paper addresses the problem of OOD detection for machine translation and dialog generation from an operational perspective. Our contributions include: (i) `RAINPROOF` a Relative informAItioN Projection ODD detection framework; and (ii) a more operational evaluation setting for OOD detection. Surprisingly, we find that OOD detection is not necessarily aligned with task-specific measures. The OOD detector may filter out samples that are well processed by the model and keep samples that are not, leading to weaker performance. Our results show that `RAINPROOF` breaks this curse and achieve good results in OOD detection while increasing performance.

1 INTRODUCTION

Significant progress have been made in Natural Language Generation (NLG) in recent years with the development of powerful generic (*e.g.,* GPT (Radford et al., 2018; 2019; Brown et al., 2020)) and task-specific (*e.g.,* Grover (Zellers et al., 2019), Pegasus (Zhang et al., 2020) and DialogGPT (Zhang et al., 2019)) text generators. Text generators power machine translation systems or chat bots that are by definition exposed to the public and whose reliability is therefore a prerequisite for adoption. Text generators are trained in the context of a so-called closed world (Antonucci et al., 2021; Fei & Liu, 2016), where training and test data are assumed to be drawn i.i.d. from a single distribution, known as the in-distribution. However, when deployed, these models operate in an open world (Parmar et al., 2021; Zhou, 2022) where the i.i.d. assumption is often violated. This change in data distribution is detrimental and induces a drop in performance as illustrated in Tab. 3 and Tab. 4. Thus, to ensure the trustworthiness and adoption, it is necessary to develop tools to protect them from harmful distribution shifts. For example, a trained translation model is not expected to be reliable when presented with another language (*e.g.* a Spanish model exposed to Catalan, or a Dutch model exposed to Afrikaans) or unexpected technical language (*e.g.*, a colloquial translation model exposed to rare technical terms from the medical field).

Most of the existing research, which aims to protect models from Out-Of-Distribution (OOD) data, focuses on classification. Despite their importance, (conditional) text generation has received much less attention even though it is among the most exposed applications. Existing solutions fall into two categories. The first one called training-aware methods (Zhu et al., 2022; Vernekar et al., 2019a;b) modifies the classifier training by exposing the neural network to OOD samples during training. The second one called plug-in methods aims at distinguishing regular samples in the in distribution (IN) from OOD samples based on the behavior of the model on a new input. Plug-in methods include Maximum Softmax Prediction (MSP) (Hendrycks & Gimpel, 2016) or Energy (Lee et al., 2018a) or feature-based anomaly detectors that compute a per-class anomaly score (Ming et al., 2022; Ryu et al., 2017; Huang et al., 2020; Ren et al., 2021a). Although plug-in methods seem attractive, their adaptation to text generation may not be straightforward. The sheer number of words present in the vocabulary prevents it to be used directly within the classification framework.

In this work, we aim at developing new tools to build more reliable text generators, which can be used in practical systems. First, we work in the *unsupervised* detection setting where we do not assume that we have access to OOD samples as they are often not available. Second, we work in the *black-box scenario*, which is the most common in the Software as a Service framework Rudin & Radin (2019). In the black-box setting detection methods only have access to the output of the DNN architecture. Third, we want an *easy-to-use and effective method to ensure adoptability*. Last, we argue that OOD detection impacts on *tasks specific performance* of the whole system should be taken into account when choosing OOD detectors in an operational setting.

Our contributions. Our main contributions can be summarized as follows:

1. *A more operational benchmark for text generation OOD detection.* We present `LOFTER` the **L**anguage **O**ut o**F** dis**T**ribution p**E**rformance benchma**R**k. Existing works on OOD detection for language modeling (Arora et al., 2021) focus on (i) english language only, (ii) the GLUE benchmark and (iii) measure performance solely in terms of OOD detection. `LOFTER` is, in our view, a more operational setting with a strong focus on neural machine translation (NMT) and dialog generation. First, it introduces more realistic data shifts that go beyond English Fan et al. (2021): language shifts induced by closely related language pairs (*e.g.*, Spanish and Catalan or Dutch and Afrikaans[1]) and domain change (*e.g.*, medical vs news data or different types of dialogs). In addition, `LOFTER` comes with an updated evaluation setting: detectors' performance are jointly evaluated w.r.t the overall system's performance on the end task.

2. *Novel information theoretic-based detectors*. We present `RAINPROOF`: a **R**elative inform**AI**tio**N** **P**rojection **O**ut **OF** distribution detector. `RAINPROOF` is fully **unsupervised**. It is flexible and can be applied both when no reference samples (IN) are available (corresponding to scenario $\mathbb{s}_0$) and when they are (corresponding to scenario $\mathbb{s}_1$). `RAINPROOF` tackles $\mathbb{s}_0$ by computing the models' predictions negentropy (Brillouin, 1953). For $\mathbb{s}_1$, it relies its natural extension: the Information Projection (Kullback, 1954; Csiszár, 1967), an information-theoretic tool that remains overlooked by the machine learning community.

3. *New insights on the operational value of OOD detectors* Our extensive experiments on `LOFTER` show that *OOD detectors may filter out samples that are well processed by the model and keep samples that are not, leading to weaker performance*. Our results show that `RAINPROOF` breaks this curse and achieve good results in OOD detection while increasing performance.

4. *Code and reproductibility.* After acceptance, we will publish the open-source code on `github.com` and the data to facilitate future research, ensure reproducibility and reduce computational costs.

2 Problem Statement & Related Works

2.1 Notations & Conditional Text Generation

Let us denote Ω a vocabulary of size $|\Omega|$ and Ω^* its Kleene closure (Fletcher et al., 1990)[2]. We denote $\mathcal{P}(\Omega) = \left\{\mathbf{p} \in [0,1]^{|\Omega|} : \sum_{i=1}^{|\Omega|} \mathbf{p}_i = 1\right\}$ the set of probability distributions defined over Ω. Let $\mathcal{D}_{train}$ be the training set, composed of $N \geqslant 1$ i.i.d. samples $\{(\mathbf{x}^i,\ \mathbf{y}^i)\}_{i=1}^N \in (\mathcal{X} \times \mathcal{Y})^N$ with probability law $\mathbf{p}_{XY}$. We denote $\mathbf{p}_X$ and $\mathbf{p}_Y$ the associated marginal laws of $\mathbf{p}_{XY}$. Each $\mathbf{x}^i$ is a sequence of tokens and we denote $x^i_j \in \Omega$ the jth token of the ith sequence. $\mathbf{x}^i_{\leqslant t} = \{x^i_1, \cdots, x^i_t\} \in \Omega^*$ denotes the prefix of length t. The same notations hold for $\mathbf{y}$.

Conditional textual generation. In conditional textual generation, the goal is to model a probability distribution $\mathbf{p}_\star(\mathbf{x}, \mathbf{y})$ over variable-length text sequences $(\mathbf{x}, \mathbf{y})$ by finding $\mathbf{p}_\theta \approx \mathbf{p}_\star(\mathbf{x}, \mathbf{y})$ for any $(\mathbf{x}, \mathbf{y})$. In this work, we assume to have access to a pretrained conditional language model $f_\theta : \mathcal{X} \times \mathcal{Y} \to \mathbf{R}^{|\Omega|}$ where the output is the (unormalized) logits scores. f_θ parameterized $\mathbf{p}_\theta$, *i.e.,* for any $(\mathbf{x}, \mathbf{y})$, $\mathbf{p}_\theta(\mathbf{x}, \mathbf{y}) = \mathrm{softmax}(f_\theta(\mathbf{x}, \mathbf{y})/T)$ where $T \in \mathbf{R}$ denotes the temperature. Given an input sequence $\mathbf{x}$, the pretrained language f_θ can recursively generate an output sequence $\hat{\mathbf{y}}$ by

[1] Afrikaans is a daughter language of Dutch (Jansen et al., 2007). The Dutch sentence: Appelen zijn gewoonlijk groen, geel of rood can be translated in "Appels is gewoonlik groen, geel of rooi."

[2] The Kleene closure corresponds to sequences of arbitrary size written with words in Ω. Formally: $\Omega^* = \bigcup_{i=0}^{\infty} \Omega^i$.

sampling $y_{t+1} \sim \mathbf{p}_\theta^T(\cdot|\mathbf{x}, \hat{\mathbf{y}}_{\leqslant t})$, for $t \in [1, |\mathbf{y}|]$. Note that $\hat{y}_0$ is the start of sentence ($<$ SOS $>$ token). We denote by $\mathcal{S}(\mathbf{x})$, the set of normalized logits scores generated by the model when the initial input is $\mathbf{x}$ *i.e.,* $\mathcal{S}(\mathbf{x}) = \{\text{softmax}(f_\theta(\mathbf{x}, \hat{\mathbf{y}}_{\leqslant t}))\}_{t=1}^{|\hat{\mathbf{y}}|}$. Note that elements of $\mathcal{S}(\mathbf{x})$ are discrete probability distributions on Ω.

2.2 Problem Statement

In OOD detection the goal is to find an anomaly score $a : \mathcal{X} \to \mathbf{R}_+$ that quantifies how much a sample is far from the IN distribution. $\mathbf{x}$ is classified as IN or OUT according to the score $a(\mathbf{x})$. Following previous work (Hendrycks & Gimpel, 2016), one fixes a threshold γ and classifies the test sample IN if $a(\mathbf{x}) \leqslant \gamma$ or OOT if $a(\mathbf{x}) > \gamma$. Formally, let us denote $g(\cdot, \gamma)$ the decision function, we take: $g(\mathbf{x}, \gamma) = \begin{cases} 1 & \text{if } a(\mathbf{x}) > \gamma \\ 0 & \text{if } a(\mathbf{x}) \leqslant \gamma \end{cases}$

Remark 1. *In our setting, OOD examples are not available*. In our experiments, we take γ such that at least 80% of the train set is classified as IN data. This assumption is reasonable since, in practice, even a well tailored dataset might contains significant shares of outliers (Mishra et al., 2020).

2.3 Review of OOD detectors

OOD detection for classification. Most works on OOD detection have focused on detectors for classifiers and relies either on internal representations (*features-based detectors*) or on the final soft probabilities produced by the classifier (*softmax based detectors*).

Features-based detectors. They leverage latent representations to derive anomaly scores (Kirichenko et al., 2020; Zisselman & Tamar, 2020). The most well-known is the Mahanalobis distance (Lee et al., 2018b; Ren et al., 2021b) but there are other methods employing Grams matrices (Sastry & Oore, 2020), Fisher Rao distance (Gomes et al., 2022) or other statistical tests (Haroush et al., 2021). Other methods rely on the gradient space (Huang et al., 2021) or the moment of the features (Quintanilha et al., 2019; Sun et al., 2021). These methods require access to the latent representations of the models, which does not fit the black-box scenario. Moreover, they often rely on a per-class decision, which is fine for classifiers but the sheer number of words in Ω makes it impossible to use for text generation.

Softmax-based detectors. These detectors rely on the soft probabilities produced by the model. The maximum softmax probability (Hendrycks & Gimpel, 2017; Hein et al., 2019; Liang et al., 2018; Hsu et al., 2020) uses the probability of the mode while others take into account the entire distribution, such as the Energy-based OOD detection scores (Liu et al., 2020). Due to the large vocabulary size, it is unclear how these methods generalize to sequence generation tasks.

OOD detection for text generation. Little work has been done on OOD detection for text generation. Therefore, we will follow Arora et al. (2021) and will rely on their baselines but also generalize common OOD scores such as MSP or Energy to the context of text generation.

Generalization to sequence generation. We generalize common OOD detectors for classification tasks by computing the average OOD score along the sequence at each step of the text generation. We refer the reader to Sec. A.6 for more details.

Remark 2. Note that *features-based* detectors assume a white-box framework where the internal representations of an input are accessible. By contrast to *softmax-based* detectors which only rely on the final output. Following Arora et al. (2021), we work in a black-box framework (Chen et al., 2020). We also compare our results to the Mahalanobis distance (Lee et al., 2018b), as it is known to be a strong baseline.

3 RAINPROOF an Information Theoretic OOD detectors

3.1 Information Theoretical Background

An information measure $\mathcal{I} : \mathcal{P}(\Omega) \times \mathcal{P}(\Omega) \to \mathbf{R}$ quantifies the similarity between any pair of discrete distributions $\mathbf{p}, \mathbf{q} \in \mathcal{P}(\Omega)$. Since Ω is a finite set, we will adopt the following notations $\mathbf{p} = [\mathbf{p}_1, \cdots, \mathbf{p}_{|\Omega|}]$ and $\mathbf{q} = [\mathbf{q}_1, \cdots, \mathbf{q}_{|\Omega|}]$. The development of new information measures for

specific applications has received much attention over the years (Fujisawa & Eguchi, 2008; Cichocki et al., 2011) (we refer the reader to Basseville (2013) for a complete review). While there exist information distances, it is, in general, difficult to build metrics that satisfy all the properties of a distance, thus we often rely on divergences which drop the symmetry property and the triangular inequality. In what follows, we motivate the information measures we will use in this work.

First, we rely on the *Rényi divergences* (Csiszár, 1967). Rényi divergences belong to the f-divergences family and are parametrized by a parameter $\alpha \in \mathbf{R}_+ - \{1\}$. They are flexible and include well-known divergences such as the Kullback-Leiber divergence (KL) Kullback (1959) (when $\alpha \to 1$) or the Hellinger distance (Hellinger, 1909) (when $\alpha = 0.5$). The Rényi divergence between $\mathbf{p}$ and $\mathbf{q}$ is defined as follows:

$$D_\alpha(\mathbf{p}\|\mathbf{q}) = \frac{1}{\alpha - 1} \log \left(\sum_{i=1}^{|\Omega|} \frac{\mathbf{p}_i^\alpha}{\mathbf{q}_i^{\alpha-1}} \right). \tag{1}$$

The Renyi divergence is widely used in machine learning (Peters et al., 2019) because α allows weighting the relative influence of the distributions' tail.

Second, we investigate the *Fisher-Rao distance* (FR). FR is a distance on the Riemannian space formed by the parametric distributions, using Fisher information matrix as its metric (Amari, 2012). It computes the geodesic distance between two discrete distributions (Rao, 1992; Pinele et al., 2020) and is defined as follows:

$$\mathrm{FR}(\mathbf{p}\|\mathbf{q}) = \frac{2}{\pi} \arccos \sum_{i=1}^{|\Omega|} \sqrt{\mathbf{p}_i \times \mathbf{q}_i}. \tag{2}$$

It has recently found many applications (Picot et al., 2022; Colombo et al., 2022b;a) and is known to be more accurate than popular divergence measures (Costa et al., 2015).

3.2 RAINPROOF FOR THE NO-REFERENCE SCENARIO ($\mathbb{s}_0$)

At inference time, the no-reference scenario ($\mathbb{s}_0$) does not assume the existence of a reference set of IN samples to decide whether a new input sample is OOD. Softmax-based detectors such as MSP (Hendrycks & Gimpel, 2016), Energy (Liu et al., 2020) or the sequence likelihood (Arora et al., 2021) are examples of OOD scores operating under $\mathbb{s}_0$.

Under these assumptions, our OOD detector RAINPROOF is composed of three steps. For a given input $\mathbf{x}$ with generated sentence $\hat{\mathbf{y}}$:

1. We first use f_θ to extract the step-by-step sequence of soft distributions $\mathcal{S}(\mathbf{x})$.

2. We then compute an anomaly score ($a_\mathcal{I}(\mathbf{x})$) by averaging a step-by-step score provided by $\mathcal{I}$. This step-by-step score is obtained by measuring the similarity between a reference distribution $\mathbf{u} \in \mathcal{P}(\Omega)$ and one element of $\mathcal{S}(\mathbf{x})$. Formally:

$$a_\mathcal{I}(\mathbf{x}) = \frac{1}{|\mathcal{S}(\mathbf{x})|} \sum_{\mathbf{p} \in \mathcal{S}(\mathbf{x})} \mathcal{I}\left(\mathbf{p}\|\mathbf{u}\right), \tag{3}$$

where $|\mathcal{S}(\mathbf{x})| = |\hat{\mathbf{y}}|$.

3. The last step consists in thresholding the previous anomaly score $a_\mathcal{I}(\mathbf{x})$. If $a_\mathcal{I}(\mathbf{x})$ is over a given threshold γ, we classify $\mathbf{x}$ as an OOD example.

Interpretation of Eq. 3. $a_\mathcal{I}(\mathbf{x})$ measures the average dissimilarity of the probability distribution of the next token to normality (as defined by $\mathbf{u}$). $a_\mathcal{I}(\mathbf{x})$ also corresponds to the token average uncertainty of the model f_θ to generate $\hat{\mathbf{y}}$ when the input is $\mathbf{x}$. The intuition behind Eq. 3 is that the distributions produced by f_θ, when exposed to an OOD sample, should be far from normality and thus should have a high score.

Choice of $\mathbf{u}$ and $\mathcal{I}$. The uncertainty definition of Eq. 3 depends on the choice of both the reference distribution $\mathbf{u}$ and the information measure $\mathcal{I}$. A natural choice for $\mathbf{u}$ is the uniform distribution, *i.e.,* $\mathbf{u} = [\frac{1}{|\Omega|}, \cdots, \frac{1}{|\Omega|}]$ which we will use in this work. It is worth pointing out that $\mathcal{I}(\cdot||\mathbf{u})$ yields the negentropy of a distribution. Other possible choices for $\mathbf{u}$ include one hot or tf-idf distribution (Colombo et al., 2022b). For $\mathcal{I}$, we rely on the Rényi divergence to obtain $a_{\mathcal{D}_\alpha}$ and the Fisher-Rao distance to obtain a_{FR}.

3.3 RAINPROOF FOR THE REFERENCE SCENARIO ($\mathbb{s}_1$)

In the with reference scenario ($\mathbb{s}_1$), we assume that one has access to a reference set of **IN samples** $\mathcal{R} = \{\mathbf{x}^i : (\mathbf{x}^i, \mathbf{y}^i) \in \mathcal{D}_{train}\}_{i=1}^{|\mathcal{R}|}$ where $|\mathcal{R}|$ is the size of the reference set. For example, the Mahalanobis distance works under this assumption. One of the weakness of Eq. 3 is to imposes an ad-hoc choice when using $\mathbf{u}$ (the uniform distribution). In $\mathbb{s}_1$, we can leverage $\mathcal{R}$, to obtain a *data-driven notion normality*.

Under $\mathbb{s}_1$, our OOD detecor RAINPROOF follows these four steps:

1. (*Offline*) For each $\mathbf{x}^i \in \mathcal{R}$, we generate $\hat{\mathbf{y}}^i$ and the associated sequence of probability distributions ($\mathcal{S}(\mathbf{x}^i)$). Overall we thus generate $\sum_{\mathbf{x}\in\mathcal{R}} |\hat{\mathbf{y}}^{\mathbf{i}}|$ probability distributions which could explode for long sequences. To overcome this limitation, we rely on the bag of distributions of each sequence (Colombo et al., 2022b). We form the set of these bags of distributions

$$\bar{\mathcal{S}}^* = \bigcup_{\mathbf{x}^i\in\mathcal{R}} \left\{ \frac{1}{|\mathcal{S}(\mathbf{x}^i)|} \sum_{\mathbf{p}\in\mathcal{S}(\mathbf{x}^i)} \mathbf{p} \right\}. \tag{4}$$

2. (*Online*) For a given input $\mathbf{x}$ with generated sentence $\hat{\mathbf{y}}$, we compute its bag of distributions representation

$$\bar{\mathbf{p}}(\mathbf{x}) = \frac{1}{|\mathcal{S}(\mathbf{x})|} \sum_{\mathbf{p}\in\mathcal{S}(\mathbf{x})} \mathbf{p}. \tag{5}$$

3. (*Online*) For $\mathbf{x}$, we then compute an anomaly score $a^\star_{\mathcal{I}}(\mathbf{x})$ by projecting $\bar{\mathbf{p}}(\mathbf{x})$ on the set $\bar{\mathcal{S}}^*$. Formally, $a^\star_{\mathcal{I}}(\mathbf{x})$ is defined as:

$$a^\star_{\mathcal{I}}(\mathbf{x}) = \min_{\mathbf{p}\in\bar{\mathcal{S}}^\star} \mathcal{I}(\mathbf{p}\|\bar{\mathbf{p}}(\mathbf{x})). \tag{6}$$

We denote $\mathbf{p}^\star(\mathbf{x}) = \arg\min_{\mathbf{p}\in\bar{\mathcal{S}}^*} \mathcal{I}(\mathbf{p}\|\bar{\mathbf{p}}(\mathbf{x}))$.

4. The last step consists in thresholding the previous anomaly score $a_{\mathcal{I}}(\mathbf{x})$. If $a_{\mathcal{I}}(\mathbf{x})$ is over a given threshold γ, we classify $\mathbf{x}$ as an OOD example.

Interpretation of Eq. 6. $a_{\mathcal{I}}(\mathbf{x})$ relies on a Generalized Information Projection (Kullback, 1954; Csiszár, 1975; 1984)[3] which measures the similarity between $\bar{\mathbf{p}}(\mathbf{x})$ and the set $\bar{\mathcal{S}}^*$. Note that the closest element of $\bar{\mathcal{S}}^*$ in the sens of $\mathcal{I}$ can give insights on the decision of the detector. It allows to interpret the decision of the detector as we will see in Tab. 5.

Choice of $\mathcal{I}$. Similarly to Sec. 3.2, we will rely on the Rényi divergence to define $a^\star_{\mathcal{R}_\alpha}(\mathbf{x})$ and the Fisher-Rao distance $a^\star_{\mathrm{FR}}(\mathbf{x})$.

4 RESULTS ON LOFTER

4.1 LOFTER: LANGUAGE OUT OF DISTRIBUTION PERFORMANCE BENCHMARK

LOFTER for NMT. We consider two main types of changes: language changes and domain changes, which both can occur in real-world situations. For each shift, we rely on pretrained generators from the HuggingFace Hub. Further experiemental details are relegated to Ap. A. *Language shifts* can appear when a translation system is exposed to a language that is extremely similar to the language the system has been trained on (*e.g.*, Afrikaans for a system trained on Dutch) and, therefore, can lead to significant translation errors (see Tab. 7)). For language shifts, we focus on closely related language pairs coming from the Tatoeba dataset (Tiedemann, 2012b) (see Tab. 6). We study the shifts induced by Catalan-Spanish, Portugese-Spanish and Afrikaans-Dutch. *Domain shifts*, which occur when the model is exposed to a specific topic that was not seen during training, can also affect the quality of the translation (see Tab. 4). To simulate domain shifts, we use the language Tatoeba MT dataset (Tiedemann, 2020) and the news commentary dataset (Tiedemann, 2012b) as base datasets

[3]The minimization problem of Eq. 6 finds numerous connections in the theory of large deviation (Sanov, 1958) or in statistical physics (Jaynes, 1957).

Table 1: Summary of the performance and computational cost of every detector.

(a) Summary of the performance of our detectors (Ours) compared to commonly used strong baselines (Bas.). We report in bold the best detector for each scenario and we underline the best overall.

			Language shifts		Domain shifts		Dialog shifts	
			AUROC ↑	FPR ↓	AUROC ↑	FPR ↓	AUROC ↑	FPR ↓
$\mathbb{S}_0$	Ours	a_{D_α}	**<u>0.95</u>**	**<u>0.25</u>**	**0.85**	**0.62**	**0.79**	**0.64**
		a_{FR}	0.93	0.28	0.81	0.67	0.76	0.68
	Bas.	a_E	0.89	0.44	0.79	0.78	0.65	0.76
		a_{MSP}	0.87	0.44	0.79	0.77	0.66	0.72
		a_L	0.78	0.79	0.73	0.88	0.65	0.95
$\mathbb{S}_1$	Ours	$a_{D^*_\alpha}$	0.88	0.34	**<u>0.86</u>**	**<u>0.50</u>**	**<u>0.86</u>**	**<u>0.52</u>**
		a_{FR^*}	0.88	0.35	0.81	0.69	0.76	0.75
	Bas.	a_M	**0.92**	**0.26**	0.78	0.59	0.84	0.55

(b) Computation time (in seconds) for the different detectors. Off. (Onl.) stands for offline (resp. online) time.

Score	Off.	Onl.
a_{D_α}	✗	2.10^{-3} s
a_{MSP}	✗	1.10^{-4} s
a_M	40s	3.10^{-3} s
$a_{D^*_\alpha}$	✗	9.10^{-2} s

and the shifts are induced by the EuroParl dataset (Tiedemann, 2012a) and EMEA (Tiedemann, 2012b) dataset.

LOFTER for dialogs. For conversational agents, an interesting scenario is when a goal-oriented agent designed to handle a specific type of conversations (*e.g.*, customer conversations, daily dialogue) is exposed to an unexpected conversation. In this case, it is crucial to interrupt the agent so it does not damage the user's trust with misplaced responses (Perez et al., 2022). We rely on the Multi WOZ dataset (Zang et al., 2020), a human to human dataset collected in the Wizard-of-Oz set-up (Kelley, 1984), for IN distribution data. This choice is mostly motivated by the availability of pretrained models on Multi WOZ. For dialog shifts, we use spoken datasets coming from various sources which are part of the SILICONE benchmark (Chapuis et al., 2020). Specifically, we use a goal-oriented dataset (*i.e.*, Switchboard Dialog Act Corpus (SwDA) (Stolcke et al., 2000)), a multi-party meetings dataset (*i.e.*, MRDA (Shriberg et al., 2004) and Multimodal EmotionLines Dataset MELD (Poria et al., 2018)), daily communication dialogs (*i.e.*, DailyDialog DyDA Li et al. (2017)), and scripted scenarii (*i.e.*, IEMOCAP Tripathi et al. (2018)). We refer the curious reader to Sec. A.4 for more details on each dataset.

Metrics. OOD detection is usually framed as an unbalanced binary classification problem where the class of interest is OUT. We can assess the performance of our OOD detectors focusing on the **False alarm rate** (FPR) and on the **True detection rate** (TPR). To evaluate the performance on the OOD task we report the AUROC ↑ and the FPR ↓ .
Area Under the Receiver Operating Characteristic curve (AUROC ↑) (Bradley, 1997). The AUROC ↑ can be interpreted as the probability that an IN-distribution example has an higher anomaly score than an OOD sample. For this metric, higher is better.
False Positive Rate at $r\%$ True Positive Rate (FPR ↓). In many practical application, we have to detect at least $r\%$ of the the OOD samples. This corresponds to pre-defined safety level. FPR ↓ quantifies the share of IN samples we wrongly detect under this constraint. It leads to select a threshold γ_r such that the corresponding TPR equals r. In our work r is set to 95%. Additional details on these metrics can be found in Sec. A.1.

4.2 Experiments in machine translation and results

Results on language shifts. We assess, for each language pair, the OOD detection performance of RAINPROOF and report the average AUROC ↑ and FPR ↓ in Tab. 1a. We provide the detailed results in Tab. 8. We find that our no-reference methods (a_{D_α} and a_{FR}) achieve better performance that common no-reference baselines but also outperform the reference-based baseline. In particular, a_{D_α}, by achieving an AUROC ↑ of 0.95 and FPR ↓ of 0.25, outperforms all considered methods. Moreover, while no-reference baselines only capture up to 62% of the OOD samples on average, ours detect up to 83.5%, achieving even better results than the with-reference baseline (75.3%).

Results on domain shifts. We evaluate the OOD detection performance of RAINPROOF on domain shifts in Spanish and German with technical medical data and parliamentary data. We report the average OOD detection performance in Tab. 1a. In $\mathbb{S}_0$, we observe that a_{D_α} and a_{FR} outperform the strongest baselines (*i.e.*, Energy, MSP and sequence likelihood) by several AUROC ↑ points.

Interestingly enough even our no-reference detectors outperform the reference-based baseline (*i.e.*, a_M). However, we find that relying on a reference set is a must-have in terms of FPR ↓. While a_{D_α} achieves similar AUROC ↑ performance to its information projection counterpart $a_{D^*_\alpha}$, the latter achieve much better FPR ↓.

4.3 Experiments in dialog generation and results

Results on Dialog shifts. The dialog shifts benchmark is more difficult than NMT benchmark as all detectors achieve lower performances. It is the only case where our no-reference detectors do not outperform the Mahalanobis baseline and achieve only 0.79 in AUROC ↑. The best baseline is the Mahalanobis distance and achieves better performance on dialog task than on NMT domain shifts reaching an AUROC ↑ of 0.84. However, our reference based detector based on the Rényi information projection secures better AUROC ↑ (0.86) and better FPR ↓ (0.52). Even though RAINPROOF outperforms all the baselines, shifts in dialog are hard to detect and will require further investigations. Non-aggregated results for dialog are provided in Ap. C. They show that RAINPROOF consistently outperforms baselines on all datasets.

Table 2: Correlation between OOD scores and translation metrics BLEU and BERT-S on domain shifts datasets.

			BERT-S			BLEU		
			ALL	IN	OUT	ALL	IN	OUT
$\mathcal{S}_0$	Ours	a_{D_α}	-0.30	-0.25	-0.18	-0.17	-0.22	-0.09
		a_{FR}	-0.36	-0.29	-0.26	-0.24	-0.25	-0.19
	Bas.	a_E	-0.19	-0.24	-0.33	-0.26	-0.18	-0.39
		a_L	-0.46	-0.51	-0.48	-0.49	-0.44	-0.50
		a_{MSP}	-0.16	-0.19	-0.29	-0.24	-0.16	-0.37
$\mathcal{S}_1$	Ours	$a_{D^*_\alpha}$	-0.14	0.00	-0.10	-0.19	0.00	-0.08
		a_{FR^*}	-0.12	0.00	-0.13	-0.17	0.00	-0.09
	Bas.	a_M	-0.04	0.00	0.05	-0.13	0.00	0.00

Importance of distribution tails. Our results show that, when it comes to domain shift (domain shifts in translation or dialog shifts), reference-based detectors are required to obtain good results. They also show that, the more these detectors take into account the tail of the distributions, the better they are, as displayed in Sec. B.1. We find that low values of α (near 0) yields better results with the Rényi Information projection $a_{D^*_\alpha}$. It suggests that the tail of the distributions used during text generation carries context information and insights on the processed texts. Such results are consistent with findings of recent works in the context of automatic evaluation of text generation (Colombo et al., 2022b).

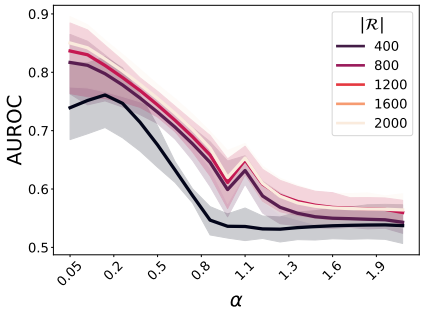

Figure 1: Impact of α on the performance of the Rényi information projection for dialog shifts detection. A smaller α increases the weight of the tail of the distribution. An α of 0 would consist in counting the number of the common non zero elements.

Comparison to the Mahalanobis distance. Our reference-based detector work with a small reference set. In our experiments, we use reference sets of size 10 to 2000. The Mahalanobis distance requires to approximate the covariance matrix of the reference set. In our simulations, the embeddings of dimension 512 make the estimation unreliable. On the contrary, RAINPROOF, which rely on information projections, remains numerically sound with small reference set.

5 Towards a practical evaluation of OOD detectors

Following previous work, we measure the performance of the detectors on the OOD detection task based on AUROC ↑ and FPR ↓. However, this evaluation framework neglects the impact of the detector on the overall system's performance. We identify three main evaluation criteria that are important in practice: execution time, overall system performance in terms of quality of the generated sentences, and interpretability of the decision. Our study is conducted on NMT because due to the existence of relevant and widely adopted metrics for assessing the quality of a generated sentence (*i.e.*, BLEU (Papineni et al., 2002) and BERTSCORE (BERT-S) (Unanue et al., 2021)).

Table 3: Average impact of different OOD detectors on the BLEU score for different type of dataset: IN data only, OOD data and the combination of both ALL. For each we report the absolute average BLEU score (Abs.), the average gains in BLEU (G.s) compared to a setting without OOD filtering (f_θ only) and the share of the subset removed by the detector (R.Sh.). These results are achieved by setting γ such that we remove 20% of the IN dataset.

			IN			OOD			ALL		
			Abs.	G.s.	R.Sh.	Abs.	G.s.	R.Sh.	Abs.	G.s.	R.Sh
		f_θ	53.8	+0.0	0.0%	30.8	+0.0	0.0%	44.6	+0.0	0.0%
$\mathbb{s}_0$	Ours	a_{D_α}	57.4	+3.6	19.8%	40.2	+9.4	57.7%	55.8	+11.2	35.4%
		a_{FR}	56.8	+3.0	19.6%	39.7	+8.9	52.5%	54.5	+10.0	33.0%
	Baselines	a_E	56.4	+2.7	20.0%	32.0	+1.2	31.5%	48.4	+3.9	24.0%
		a_L	58.1	+4.3	19.0%	34.4	+3.7	42.8%	52.3	+7.7	28.4%
		a_{MSP}	52.4	-1.3	19.5%	26.8	-4.0	38.3%	43.2	-1.3	28.7%
$\mathbb{s}_1$	Ours	$a_{D^*_\alpha}$	54.3	+0.5	19.9%	32.2	+1.4	60.9%	49.3	+4.8	38.0%
		a_{FR^*}	54.4	+0.6	19.3%	32.2	+1.4	60.6%	49.3	+4.8	37.6%
	Baselines	a_M	54.1	+0.3	20.0%	30.1	-0.7	65.0%	49.5	+4.9	39.9%

5.1 Complexity Study

Runtime and memory costs. We report in Tab. 1b the runtime of all methods. Detectors for $\mathbb{s}_0$ are faster than the ones for $\mathbb{s}_1$. Contrarily to detectors using references, the no-reference detectors do not require additional memory. They can be setup easily in a plug&play manner at the output of any model.

Numerical stability. The Mahalanobis distance requires to estimate both μ and $\mathbf{\Sigma}^{-1}$ (see Sec. A.6). The dimension of the latent space of the considered pre-trained model is either 768 or 512. In this setting, when the size of the reference set is small, the estimation of the Mahalanobis parameters is numerically unstable. For $\mathbb{s}_1$, RAINPROOF relies on information projection and does not involve numerically unstable computations but requires a larger memory footprint (0.5 GB) to store the reference set (2000 probability distributions of dimension 50K).

5.2 Impact of OOD filtering on translation quality

The main objective of OOD filtering is to remove samples that are far from the training distribution. On these samples, the user has no guarantee that the model will produce a good quality translation. In this experiment, we compare the performance of the system with and without the different detectors in terms of the quality of the generated sentence.

Global performance. In Tab. 3, we report the global performance of the systems (f_θ) without and with OOD detectors on IN samples, OOD samples and all samples (ALL). From the first row of Tab. 3, we notice that OOD samples are harmful to the model. We observe that, in most of the cases, adding detectors increases the model performance on IN, OOD and all samples. Exceptions include a_{MSP} (for OOD, IN and ALL) and a_M (for OOD). Results indicate that no-reference RAINPROOF outperforms the reference-based version of RAINPROOF. Thus, **OOD detector evaluation should consider the final task performance.** Overall, it is worth noting that directly adapting classical OOD detection methods (*e.g.*, MSP or Energy) to the sequence generation problem leads to poor results in terms of performance gains (*i.e.*, as measured by BLEU or BERT-S). **In others words, the final task does not benefit from adding classical OOD detectors.**

Finer performance analysis. In Tab. 4, we report the per-shift-types performance of f_θ with and without OOD detector. In Tab. 4, we observe a decrease in performance in the case of language and domain shifts, the latter being more harmful. On domain shifts, we observe that reference-based detectors decrease system's performance on OOD samples. This means that the detectors tend to filter out samples that are well-handled by the model and ignore sentences that are not. It is worth noting that reference-based detectors remove, in proportion, twice as many samples as their no-reference counterparts, while the threshold selection procedure remains the same. This observation also holds when removing less samples (*i.e.*, calibrating γ that we remove 10%, 5% or even 1% of the IN dataset) (Tab. 15).

Table 4: Detailed impacts on NMT performance results per tasks (Domain- or Language-shifts) of the different OOD detectors. We present results on the different part of the data: IN data, OOD data and the combination of both, ALL. For each we report the absolute average BLEU score (Abs.), the average gains in BLEU (G.s.) compared to a setting without OOD filtering (f_θ only) and the share of the subset removed by the detector (R.Sh.). We provide more detailed results on each dataset in Ap. D

			Domain shifts									Language shifts								
			IN			OOD			ALL			IN			OOD			ALL		
			Abs.	G.s.	R.Sh.	Abs.	G.s.	R.Sh.	Abs.	G.s.	R.Sh.	Abs.	G.s.	R.Sh.	Abs.	G.s.	R.Sh.	Abs.	G.s.	R.Sh.
		f_θ	47.1	+0.0	0.0%	43.4	+0.0	0.0%	45.3	+0.0	0.0%	60.5	+0.0	0.0%	18.1	+0.0	0.0%	43.9	+0.0	0.0%
$\mathcal{S}_0$	Ours	a_{D_α}	50.6	+3.6	19.8%	48.8	+5.4	31.9%	50.8	+5.6	25.9%	64.2	+3.7	19.7%	31.6	+13.4	83.5%	60.7	+16.8	44.9%
		a_{FR}	50.2	+3.1	20.0%	47.1	+3.7	24.7%	49.4	+4.1	22.3%	63.5	+3.0	19.2%	32.3	+14.1	80.3%	59.7	+15.8	43.6%
	Bs.	a_E	49.4	+2.4	20.0%	45.8	+2.4	17.9%	47.8	+2.6	18.9%	63.5	+2.9	20.0%	18.2	+0.1	45.1%	49.0	+5.1	29.1%
		a_L	50.8	+3.7	19.2%	47.6	+4.1	23.6%	49.9	+4.6	21.4%	65.4	+4.9	18.8%	21.3	+3.2	62.0%	54.6	+10.7	35.4%
		a_{MSP}	45.9	-1.2	19.6%	33.6	-9.8	45.0%	40.8	-4.4	32.3%	59.0	-1.5	19.4%	20.0	+1.9	31.6%	45.6	+1.8	25.1%
$\mathcal{S}_1$	Ours	$a_{D^*_\alpha}$	47.3	+0.2	19.9%	39.1	-4.4	61.0%	46.2	+1.0	40.5%	61.4	+0.8	19.8%	25.3	+7.2	60.8%	52.4	+8.5	35.6%
		a_{FR^*}	47.3	+0.3	19.0%	39.0	-4.4	60.4%	46.2	+1.0	39.7%	61.4	+0.9	19.7%	25.3	+7.2	60.7%	52.4	+8.5	35.5%
	Bs.	a_M	47.2	+0.1	20.0%	38.0	-5.5	54.7%	43.8	-1.5	37.4%	61.0	+0.5	20.0%	22.2	+4.1	75.3%	55.2	+11.4	42.4%

Threshold free analysis. In Tab. 2, we report the correlation between OOD scores and final task performance for the case of domain shifts. We refer the reader to Tab. 14 for the results on language shifts. We observe that the likelihood score is the most correlated with the final sentence quality, as measured by BLEU or BERT−S. This finding illustrates that higher correlation with sentence quality does not necessarily translate into higher performance gains when filtering OOD samples. *This result suggests that Quality Estimation* (Specia et al., 2010; Blatz et al., 2004)*, while closely related, is a different problem.*

5.3 Towards an interpretable decision

An important dimension fostering adoption is the ability to verify the decision taken by the automatic system (Montavon et al., 2018). RAINPROOF offers a step in this direction when used with references: for each input sample, RAINPROOF finds the closest sample (in the sens of the Information Projection) in the reference set to take its decision. We present in Tab. 5 some OOD samples along with their translation scores, projection scores, and their projection on the reference set. We notice that, in general, sentences that are close to the reference set, and whose projection has a close meaning, are better handled by f_θ. Therefore, one can visually interpret the prediction of RAINPROOF, and validate it. This observation further validate our method.

Table 5: OOD inputs, their translations and projections onto the reference set. The first 2 are far from the reference set and not well translated whereas the next 2 are very close to the reference set and well translated. We can, for that matter, notice that the projection is quite close to the input sentence grammatically speaking.

Source	Ahir a la nit vàrem treballar fins a les deu.	
Ground truth	Last night we worked until 10 p.m.	
Generated	Ahir a la nit vàrem treballar fins a les deu.	BLEU 3.75
$\mathbf{p}^*(\mathbf{x})$	Dar gato por liebre.	**Score** 1.23
Source	Aquesta cola s'ha esbravat i no té bon gust.	
Ground-truth	This cola has lost its fizz and doesn't taste any good.	
Generated	This tail s'ha esbravat i no tea bon gust.	BLEU 4.09
$\mathbf{p}^*(\mathbf{x})$	Esta cuchara es de té.	**Score** 1.14
source	Aquesta és una carta molt estranya.	
Ground-truth	This is a very strange letter.	
Generated	This is a molt estranya card.	BLEU 26.27
$\mathbf{p}^*(\mathbf{x})$	Este carro es chiquito.	**Score** 0.74
source	Austràlia no és Àustria.	
Ground-truth	Australia isn't Austria.	
Generated	Austràlia is not Austria.	BLEU 21.86
$\mathbf{p}^*(\mathbf{x})$	La vida no es fácil.	**Score** 0.82

6 Conclusions

In this work, we introduced both a detection framework called RAINPROOF as well as a new benchmark called LOFTER for detecting OOD samples when using textual generators in the black-box scenario. Our work adopts an operational perspective by not only considering OOD performance but also task-specific metrics. Our results show that, despite the good results obtained in pure OOD detection, *OOD filtering can harm the performance of the final system, as it is the case for MSP or Mahanalobis.* We found that, RAINPROOF breaks this curse and induces significant gains in translation performance both on OOD samples and in general. In conclusion, this work paves the way to the development of detectors tailored for text generators and calls for a global evaluation when benchmarking future OOD detectors.

REFERENCES

Shun-ichi Amari. *Differential-geometrical methods in statistics*, volume 28. Springer Science & Business Media, 2012.

Alessandro Antonucci, Alessandro Facchini, and Lilith Mattei. Structural learning of probabilistic sentential decision diagrams under partial closed-world assumption. 2021. doi: 10.48550/ARXIV. 2107.12130. URL https://arxiv.org/abs/2107.12130.

Udit Arora, William Huang, and He He. Types of out-of-distribution texts and how to detect them. *arXiv preprint arXiv:2109.06827*, 2021.

Michéle Basseville. Divergence measures for statistical data processing—an annotated bibliography. *Signal Processing*, 93(4):621–633, 2013.

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. Confidence estimation for machine translation. In *Coling 2004: Proceedings of the 20th international conference on computational linguistics*, pp. 315–321, 2004.

Andrew P Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159, 1997.

L. Brillouin. The negentropy principle of information. *Journal of Applied Physics*, 24(9):1152–1163, September 1953. doi: 10.1063/1.1721463. URL https://doi.org/10.1063/1.1721463.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Emile Chapuis, Pierre Colombo, Matteo Manica, Matthieu Labeau, and Chloé Clavel. Hierarchical pre-training for sequence labelling in spoken dialog. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 2636–2648, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.239. URL https://www.aclweb.org/anthology/2020.findings-emnlp.239.

Jianbo Chen, Michael I Jordan, and Martin J Wainwright. Hopskipjumpattack: A query-efficient decision-based attack. In *2020 ieee symposium on security and privacy (sp)*, pp. 1277–1294. IEEE, 2020.

Andrzej Cichocki, Sergio Cruces, and Shun-ichi Amari. Generalized alpha-beta divergences and their application to robust nonnegative matrix factorization. *Entropy*, 13(1):134–170, 2011.

Pierre Colombo, Guillaume Staerman, Nathan Noiry, and Pablo Piantanida. Learning disentangled textual representations via statistical measures of similarity. *arXiv preprint arXiv:2205.03589*, 2022a.

Pierre Jean A Colombo, Chloé Clavel, and Pablo Piantanida. Infolm: A new metric to evaluate summarization & data2text generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 10554–10562, 2022b.

Sueli IR Costa, Sandra A Santos, and Joao E Strapasson. Fisher information distance: A geometrical reading. *Discrete Applied Mathematics*, 197:59–69, 2015.

Imre Csiszár. Information-type measures of difference of probability distributions and indirect observation. *studia scientiarum Mathematicarum Hungarica*, 2:229–318, 1967.

Imre Csiszár. I-divergence geometry of probability distributions and minimization problems. *The annals of probability*, pp. 146–158, 1975.

Imre Csiszár. Sanov property, generalized i-projection and a conditional limit theorem. *The Annals of Probability*, pp. 768–793, 1984.

Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pp. 233–240, 2006.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22(107):1–48, 2021.

Geli Fei and Bing Liu. Breaking the closed world assumption in text classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 506–514, 2016.

Peter Fletcher, Hughes Hoyle, and C Wayne Patty. *Foundations of discrete mathematics*. Brooks/Cole, Florence, KY, November 1990.

Hironori Fujisawa and Shinto Eguchi. Robust parameter estimation with a small bias against heavy contamination. *Journal of Multivariate Analysis*, 99(9):2053–2081, 2008.

Eduardo Dadalto Camara Gomes, Florence Alberge, Pierre Duhamel, and Pablo Piantanida. Igeood: An information geometry approach to out-of-distribution detection. *arXiv preprint arXiv:2203.07798*, 2022.

Matan Haroush, Tzviel Frostig, Ruth Heller, and Daniel Soudry. A statistical framework for efficient out of distribution detection in deep neural networks, 2021.

Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 41–50, 2019.

Ernst Hellinger. Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen. *Journal für die reine und angewandte Mathematik*, 1909(136):210–271, 1909.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017.

Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10951–10960, 2020.

Haiwen Huang, Zhihan Li, Lulu Wang, Sishuo Chen, Bin Dong, and Xinyu Zhou. Feature space singularity for out-of-distribution detection. *arXiv preprint arXiv:2011.14654*, 2020.

Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. *ArXiv*, abs/2110.00218, 2021.

Carel Jansen, Robert Schreuder, and Anneke Neijt. The influence of spelling conventions on perceived plurality in compounds: A comparison of afrikaans and dutch. *Written language & literacy*, 10(2): 185–194, 2007.

Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.

John F Kelley. An iterative design methodology for user-friendly natural language office information applications. *ACM Transactions on Information Systems (TOIS)*, 2(1):26–41, 1984.

Polina Kirichenko, Pavel Izmailov, and Andrew G Wilson. Why normalizing flows fail to detect out-of-distribution data. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 20578–20589. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/ecb9fe2fbb99c31f567e9823e884dbec-Paper.pdf.

Solomon Kullback. *Information theory and statistics*. Courier Corporation, 1954.

Solomon Kullback. *Information Theory and Statistics*. John Wiley, 1959.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018a.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 31*, pp. 7167–7177. Curran Associates, Inc., 2018b.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks, 2018c. URL https://arxiv.org/abs/1807.03888.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of The 8th International Joint Conference on Natural Language Processing (IJCNLP 2017)*, 2017.

Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H1VGkIxRZ.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems*, 2020.

Yifei Ming, Yiyou Sun, Ousmane Dia, and Yixuan Li. Cider: Exploiting hyperspherical embeddings for out-of-distribution detection. *arXiv preprint arXiv:2203.04450*, 2022.

Swaroop Mishra, Anjana Arunkumar, Bhavdeep Sachdeva, Chris Bryan, and Chitta Baral. Dqi: Measuring data quality in nlp, 2020. URL https://arxiv.org/abs/2005.00816.

Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Methods for interpreting and understanding deep neural networks. *Digital signal processing*, 73:1–15, 2018.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040.

Jitendra Parmar, Satyendra Singh Chouhan, Vaskar Raychoudhury, and Santosh Singh Rathore. Open-world machine learning: Applications, challenges, and opportunities, 2021. URL https://arxiv.org/abs/2105.13448.

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

Ben Peters, Vlad Niculae, and André FT Martins. Sparse sequence-to-sequence models. *arXiv preprint arXiv:1905.05702*, 2019.

Marine Picot, Francisco Messina, Malik Boudiaf, Fabrice Labeau, Ismail Ben Ayed, and Pablo Piantanida. Adversarial robustness via fisher-rao regularization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Julianna Pinele, João E. Strapasson, and Sueli I. R. Costa. The fisher–rao distance between multivariate normal distributions: Special cases, bounds and applications. *Entropy*, 22(4), 2020. ISSN 1099-4300. doi: 10.3390/e22040404. URL https://www.mdpi.com/1099-4300/22/4/404.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*, 2018.

Igor M. Quintanilha, Roberto de M. E. Filho, José Lezama, Mauricio Delbracio, and Leonardo O. Nunes. Detecting out-of-distribution samples using low-order deep features statistics, 2019. URL https://openreview.net/forum?id=rkgpCoRctm.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

C Radhakrishna Rao. Information and the accuracy attainable in the estimation of statistical parameters. In *Breakthroughs in statistics*, pp. 235–247. Springer, 1992.

Jie Ren, Stanislav Fort, Jeremiah Liu, Abhijit Guha Roy, Shreyas Padhy, and Balaji Lakshminarayanan. A simple fix to mahalanobis distance for improving near-ood detection. *arXiv preprint arXiv:2106.09022*, 2021a.

Jie Ren, Stanislav Fort, Jeremiah Liu, Abhijit Guha Roy, Shreyas Padhy, and Balaji Lakshminarayanan. A simple fix to mahalanobis distance for improving near-ood detection, 2021b.

Cynthia Rudin and Joanna Radin. Why are we using black box models in ai when we don't need to? a lesson from an explainable ai competition. 2019.

Seonghan Ryu, Seokhwan Kim, Junhwi Choi, Hwanjo Yu, and Gary Geunbae Lee. Neural sentence embedding using only in-domain sentences for out-of-domain sentence detection in dialog systems. *Pattern Recognition Letters*, 88:26–32, 2017.

Ivan N Sanov. *On the probability of large deviations of random variables*. United States Air Force, Office of Scientific Research, 1958.

Chandramouli Shama Sastry and Sageev Oore. Detecting out-of-distribution examples with Gram matrices. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8491–8501. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/sastry20a.html.

Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. The icsi meeting recorder dialog act (mrda) corpus. Technical report, INTERNATIONAL COMPUTER SCIENCE INST BERKELEY CA, 2004.

Lucia Specia, Dhwaj Raj, and Marco Turchi. Machine translation evaluation versus quality estimation. *Machine translation*, 24(1):39–50, 2010.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Marie Meteer, and Carol Van Ess-Dykema. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3): 339–371, 2000.

Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. *ArXiv*, abs/2111.12797, 2021.

Jorg Tiedemann. Parallel data, tools and interfaces in opus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012a. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7.

Jörg Tiedemann. The Tatoeba Translation Challenge – Realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pp. 1174–1182, Online, November 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.wmt-1.139.

Jörg Tiedemann and Santhosh Thottingal. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal, 2020.

Jörg Tiedemann. Parallel data, tools and interfaces in opus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012b. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7.

Samarth Tripathi, Sarthak Tripathi, and Homayoon Beigi. Multi-modal emotion recognition on iemocap dataset using deep learning. *arXiv preprint arXiv:1804.05788*, 2018.

Inigo Jauregi Unanue, Jacob Parnell, and Massimo Piccardi. Berttune: Fine-tuning neural machine translation with bertscore, 2021. URL https://arxiv.org/abs/2106.02208.

Sachin Vernekar, Ashish Gaurav, Vahdat Abdelzad, Taylor Denouden, Rick Salay, and Krzysztof Czarnecki. Out-of-distribution detection in classifiers via generation, 2019a. URL https://arxiv.org/abs/1910.04241.

Sachin Vernekar, Ashish Gaurav, Taylor Denouden, Buu Phan, Vahdat Abdelzad, Rick Salay, and Krzysztof Czarnecki. Analysis of confident-classifiers for out-of-distribution detection. *arXiv preprint arXiv:1904.12220*, 2019b.

Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. Multiwoz 2.2: A dialogue dataset with additional annotation corrections and state tracking baselines. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI, ACL 2020*, pp. 109–117, 2020.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. Defending against neural fake news. *Advances in neural information processing systems*, 32, 2019.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pp. 11328–11339. PMLR, 2020.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*, 2019.

Zhi-Hua Zhou. Open-environment machine learning. *National Science Review*, 9(8):nwac123, 2022.

Qiuyu Zhu, Guohui Zheng, and Yingying Yan. Effective out-of-distribution detection in classifier based on pedcc-loss, 2022. URL https://arxiv.org/abs/2204.04665.

Ev Zisselman and Aviv Tamar. Deep residual flow for out of distribution detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

7 APPENDIX

A Experimental setting **16**

A.1 Additionnal Details on Metrics 16

A.2 Language pairs 16

A.3 Samples 17

A.4 Dialog datasets 17

A.5 Choices of models 17

A.6 Generalization of existing OOD detectors to Sequence Generation 17

B Parameters tuning **18**

B.1 Impact of α 18

C Performance of our detectors in OOD detection **18**

C.1 Summary of our results 18

C.2 Detailed results of OOD detection performances 19

C.3 ROC AUC curves 21

D NTM performance **23**

D.1 Absolute performances 23

D.2 Gains 26

D.3 Effect of a Larger threshold on NMT performance 26

E Negative results **26**

E.1 Different aggregation of OOD metrics 26

E.2 Negentropy of bag of distributions 26

A EXPERIMENTAL SETTING

In this section we dive into the details and definitions of our experimental setting. First we present our OOD detection performance metrics (Sec. A.1), then we provide a couple samples for one of the small language shifts (Sec. A.3). We also discusse the choices of pretrained model (Sec. A.5) and how we adapted common OOD detectors to the text generation case (Sec. A.6).

A.1 ADDITIONNAL DETAILS ON METRICS

OOD Detection is usually an unbalanced binary classification problem where the class of interest is OUT. Let us denote Z the random variable corresponding to actually being out of distribution. We can assess the performance of our OOD detectors focusing on the **False alarm rate** and on the **True detection rate**. The **False alarm rate** or False positive rate (FPR) is the proportion of samples missclassified as OUT. For a score threshold γ, we have FPR $= \Pr\big(a(\mathbf{x}) > \gamma \,|\, Z = 0\big)$. The **True detection rate** or True positive rate (TPR) is the proportion of OOD samples that are detected by the method. It is given by TPR $= \Pr\big(a(\mathbf{x}) > \gamma \,|\, Z = 1\big)$.

In order to evaluate the performance of our methods we will focus and report mainly the AUROC $\uparrow$ and the FPR $\downarrow$, we provide more detailed metrics and experiments in Sec. A.1.

Area Under the Receiver Operating Characteristic curve (AUROC $\uparrow$) Bradley (1997). The Receiver Operating Characteristic curve is curve obtained by plotting the True positive rate against the False positive rate. The area under this curve is the probability that an in-distribution example $\mathbf{X}_{in}$ has a anomaly score higher than an OOD sample $\mathbf{x}_{out}$: AUROC $\uparrow = \Pr(a(\mathbf{x}_{in}) > a(\mathbf{x}_{out}))$. It is given by $\gamma \mapsto \big(\Pr\big(a(\mathbf{x}) > \gamma \,|\, Z = 0\big), \Pr\big(a(\mathbf{x}) > \gamma \,|\, Z = 1\big)\big)$.

False Positive Rate at 95% True Positive Rate (FPR $\downarrow$). We accept to allow only a given false positive rate r corresponding to a defined level of safety and we want to know what share of positive samples we actually catch under this constraint. It leads to select a threshold γ_r such that the corresponding TPR equals r. At this threshold, one then computes: $\Pr(a(\mathbf{x}) > \gamma_r \,|\, Z = 0)$ with γ_r s.t. TPR$(\gamma_r) = r$. r is chosen depending of the difficulty of task at hand and the required level of safety.

For the sake of brevity we present only AUROC $\uparrow$and FPR $\downarrow$ metrics in our aggregated results but we also used Detection error and Area Under the Precision-Recall curve metrics and those are presented in our full results section (Ap. C).

Detection error. It is simply the probability of miss-classification for a given True positive rate.

Area Under the Precision-Recall curve (AUPR-IN/AUPR-OUT) Davis & Goadrich (2006). The Precision-Recall curve plots the recall (true detection rate) against the precision (actual proportion of OOD amongst the predicted OOD). The area under this curve $\gamma \mapsto \big(\Pr\big(Z = 1 \,|\, s(\mathbf{X}) \leqslant \gamma\big), \Pr\big(s(\mathbf{X}) \leqslant \gamma \,|\, Z = 1\big)\big)$ captures the trade-off between precision and recall made by the model. A high value represents a high precision and a high recall *i.e.* the detector captures most of the positive samples while having few False positive.

A.2 LANGUAGE PAIRS

Model	IN data	OUT data
Language shift		
DE-EN	Tatoeba DE	News FR
DE-EN	Tatoeba DE	Tatoeba NLD
ES-EN	Tatoeba ES	News FR
ES-EN	Tatoeba ES	Tatoeba CAT
ES-EN	Tatoeba ES	Tatoeba POR
NLD-EN	Tatoeba ES	AFR
Domain shift		
DE-EN	Tatoeba DE	EMEA DE
DE-EN	Tatoeba DE	Eurparl DE
ES-EN	Tatoeba ES	EMEA DE
ES-EN	Tatoeba ES	Eurparl DE

Table 6: Summary of models and studied shifts.

A.3 SAMPLES

Source sentence	Expected translation	Translation	BLEU
A en Tom li agrada la tecnologia.	Tom likes technology.	Tom li likes technology.	42.73
Ací està la teua bossa.	Here is your bag.	Ací está la teua bossa.	8.12
Això et posarà en perill.	That'll put you in danger.	Això et posarà en perill.	8.12
A Londres hi han molts parcs bonics.	There are many beautiful parks in London.	To London hi han molts parcs bonics.	6.57
Aquest pa és molt deliciós.	This bread is very delicious.	Aquest pa és molt deliciós.	8.12
A tots els meus amics els agraden els videojocs.	All my friends like playing videogames.	A tots els meus amics els agrade els videojocs.	4.20
Açò és un peix.	This is a fish.	Aaaaaaaaaaaaaaaaaaa ... aaaaaaaaaaaaaaaaaaaaaaaa	0.00
Moltes felicitats!	Congratulations!	Moltes congrats!	27.52
Bon any nou!	Happy New Year!	Bon any nou!	15.97
Aquell que menteix, robarà.	He that will lie, will steal.	The one who's mindless, he'll steal.	12.22
Jo sóc qui té la clau.	I'm the one who has the key.	Jo soc qui te la clau.	5.69
En Tom surt a treballar cada matí a dos quarts de set.	Tom leaves for work at 6:30 every morning.	In Tom surt to pull each matí to two quarts of set.	3.67
Ell m'ha dit que la seva casa era embruixada.	He told me that his house was haunted.	Ell m'ha dit that the seva house was haunted.	27.78
Aquest és el lloc on va nèixer el meu pare.	This is the place where my father was born.	Aquest is the lloc on va nèixer el meu pare.	8.30

Table 7: Example of behavior of a language model trained to handle Spanish inputs on Catalan inputs.

A.4 DIALOG DATASETS

Switchboard Dialog Act Corpus (`SwDA`) is a corpus of telephonic conversations. The corpus provides labels, topic and speaker information (Stolcke et al., 2000).

ICSI MRDA Corpus (`MRDA`) contains transcript 75h of naturally occuring meetings involving more than 50 people (Shriberg et al., 2004).

DaylyDialog Act Corpus (`DyDA`) contains daily common communications between people, covering topic such as small talk, meteo or daily activities (Li et al., 2017).

Interactive Emotional Dyadic Motion Capture IEMOCAP)(Tripathi et al., 2018) consists of transcripts of improvisations or scripted scenarii supposed to outline the expression of emotions.

A.5 CHOICES OF MODELS

A lot of pretrained model for conditional text generation are available. To perform our experiments we needed models that were already well installed and deployed and that would also support OOD settings. For translation tasks we needed specialized models for a notion of OOD to be easily defined. It would be indeed more hazardous to define a notion of OOD language when working with a multilingual model. The same is true for conversational models.

Neural Machine Translation model We benchmark our OOD method on translation models provided by Helsinky NLP Tiedemann & Thottingal (2020) on several pairs of languages with large and small shifts. We extended the experiment to detect domain shifts. These models are indeed specialized in each language pairs and are widely recognize in the neural machine translation field.

Conversational model We used a dialogGPT Zhang et al. (2019) model fine-tuned on the `Multi WOZ` dataset as chat bot model. The finetuning on daily dialogue type tasks ensure that the model is specialized, thus allowing us to get a good definition of samples not being in its range of expertise. Moreover, the choice of the architure, DialogGPT, guarantee that our results are valid on a very common architure.

A.6 GENERALIZATION OF EXISTING OOD DETECTORS TO SEQUENCE GENERATION

In this section, we extend classical OOD detection score to the conditional text generation settting. Common OOD detectors were built for classification tasks and we need to adapt them to conditional text generation. Our task can be viewed as a sequence of classification problems with a very large number of classes (the size of the vocabulary). We chose the most naive approach which consists of averaging the OOD scores over the sequence. We experimented with other aggregation such as the min/max or the standard deviation without getting interesting results.

Likelihood Score The most naive approach to build a OOD score is to rely solely on the log-likelyhood of the sequence. For a conditioning $\mathbf{x}$ we define the log-likelyhood score by $a_L(\mathbf{x}) = -\sum_{t=0}^{|\hat{\mathbf{y}}|-1} \log \mathbf{p}_\theta(\hat{\mathbf{y}}_{t+1}|\mathbf{x}, \hat{\mathbf{y}}_{\leqslant t})$

Average Maximum Softmax Probability score The maximum softmax probability Hendrycks & Gimpel (2017) takes the probability of the mode of the categorical distribution as score of OOD. We extend thise definition in the case of sequence of probability distribution by averaging this score along the sequence. For a given conditioning $\mathbf{x}$, we define the average MSP score $a_{\text{MSP}}(\mathbf{x}) = \frac{1}{|\hat{\mathbf{y}}|}\sum_{t=1}^{|\hat{\mathbf{y}}|} \max_{i\in[|0,K|]} \mathbf{p}_\theta^T(i|\mathbf{x}, \hat{\mathbf{y}}_{\leqslant t}))$. While it is closely linked to uncertainty measures it discards most of the information contained in the probability distribution. It discards the whole probability distribution. We claim that much more information can be retrieve by studying the whole distribution.

Average Energy score We extend the definition of the energy score described in Liu et al. (2020) to a sequence of probability distributions by averaging the score along the sequence. For a given conditioning $\mathbf{x}$ and a temperature T we define the average energy of the sequence:$a_E(\mathbf{x}) \triangleq -\frac{T}{|\hat{\mathbf{y}}|}\sum_{t=1}^{|\hat{\mathbf{y}}|} \log \sum_i^{|\Omega|} e^{f_\theta(\mathbf{x},\hat{\mathbf{y}}_{\leqslant t})_i/T}$. It corresponds to the normalization term of the $\mathrm{softmax}$ function applied on the logits. While it takes into account the whole distribution, it only takes into account the amount of unormalized mass before normalization without attention to how this mass is distributed along the features.

Mahalanobis distance Following Lee et al. (2018c) compute the Mahalanobis matrice based on the samples of a given reference set $\mathcal{R}$. In our case we are using encoder-decoder models we use the output of the last hidden layer of the encoder as embedding. Let's denote $\phi(\mathbf{x})$ this embedding for a conditionning $\mathbf{x}$. Let's μ and Σ be respectively the mean and the covariance of these embedding on the reference set. We define $a_{\text{M}}(\mathbf{x}) = \left(1 + (\phi(\mathbf{x}) - \mu)^\top \Sigma^{-1} (\phi(\mathbf{x}) - \mu)\right)^{-1}$.

B PARAMETERS TUNING

Detectors depend on their anomaly score to make decision and these scores can be parametric. First of all, soft probability based scores depend on the soft probability distribution and its scaling, therefore the temperature is a crucial parameter to tune to get the most performance. While a small temperature tend to make the distribution more picky, higher value spread the probability mass along the classes. Moreover, the renyi divergence and its related informaiton projection depend on a factor α. We provide here further results and analysis of those parameters on our results.

B.1 IMPACT OF α

Indeed, in Fig. 1 we present the impact of the size of the reference set and of the paramet α on Renyi information projection to distinguish dialog shifts, as expected, the larger the reference set, the better. However, we see that smaller values of α yield better results.

We recall that the Renyi divergence is defined as $D_\alpha(\mathbf{p}\|\mathbf{q}) = \frac{1}{\alpha-1}\log\left(\sum_{i=1}^{|\Omega|} \frac{p_i^\alpha}{q_i^{\alpha-1}}\right)$, where $\alpha \in \mathbb{R}_+ - \{1\}$. Smaller values of α distribute the weight of each feature more equally in the final divergence, more specifically they tend to give an equal weight to the very likely outcome as well as to the less likely ones, therefore giving more weight to the tail of the distribution. When α tends to 0 the Renyi divergence actually counts the number of nonzero common probabilities. That makes sens in terms of topic detection, it counts the common tokens considered during text generation.

C PERFORMANCE OF OUR DETECTORS IN OOD DETECTION

C.1 SUMMARY OF OUR RESULTS

In Fig. 2 we present the different level of performance of all the detectors we studied. We can see that in every task our detectors outperform the baselines but also that in dialog shift, while the Mahalanobis distance outperform clearly our detectors for $\mathbb{S}_0$, they still outperform baselines for their scenario by far.

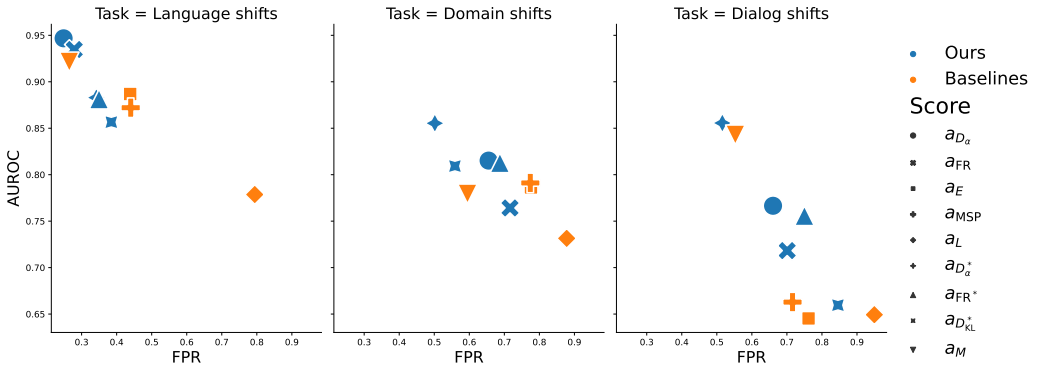

Figure 2: Trade-offs between AUROC ↑and FPR ↓ for each tasks and metrics

Table 8: Detailed results of the performances of our OOD detectors on different language shifts. The first language of the pair is the reference language of the model and the second one is the studied shift.

Scenario		Score	AUPR-IN	AUPR-OUT	AUROC	ERR	FPR
s_0	Ours	a_{D_α}	**1.00**	**0.99**	**1.00**	**0.03**	**0.01**
		a_{FR}	0.99	**0.99**	0.99	0.05	0.04
	Baselines	a_E	0.98	0.95	0.98	0.05	0.05
		a_L	0.81	0.75	0.80	0.41	0.78
		a_{MSP}	0.99	**0.99**	0.99	0.04	0.03
s_1	Ours	$a_{D_\alpha^*}$	**1.00**	**1.00**	**1.00**	**0.03**	**0.01**
		$a_{D_{\text{KL}}^*}$	0.98	0.99	0.99	**0.03**	0.02
		a_{FR^*}	**1.00**	**1.00**	**1.00**	**0.03**	**0.01**
	Baselines	a_M	**1.00**	**1.00**	**1.00**	**0.03**	**0.01**

(a) deu-fr

Scenario		Score	AUPR-IN	AUPR-OUT	AUROC	ERR	FPR
s_0	Ours	a_{D_α}	**0.76**	**0.91**	**0.85**	**0.51**	**0.72**
		a_{FR}	0.71	0.88	0.80	0.53	0.76
	Baselines	a_E	0.74	0.88	0.82	0.56	0.79
		a_L	0.73	0.88	0.82	0.58	0.82
		a_{MSP}	0.70	0.88	0.81	0.56	0.80
s_1	Ours	$a_{D_\alpha^*}$	**0.71**	**0.94**	**0.87**	**0.23**	**0.31**
		$a_{D_{\text{KL}}^*}$	0.61	**0.94**	0.84	0.25	0.35
		a_{FR^*}	**0.71**	**0.94**	**0.87**	**0.23**	**0.31**
	Baselines	a_M	0.62	0.93	0.85	0.29	0.39

(b) nld-afr

Scenario		Score	AUPR-IN	AUPR-OUT	AUROC	ERR	FPR
s_0	Ours	a_{D_α}	**0.86**	**0.93**	**0.91**	**0.30**	**0.44**
		a_{FR}	0.85	0.92	0.89	0.35	0.51
	Baselines	a_E	0.71	0.82	0.77	0.58	0.87
		a_L	0.67	0.79	0.75	0.59	0.88
		a_{MSP}	0.68	0.83	0.73	0.58	0.87
s_1	Ours	$a_{D_\alpha^*}$	0.56	0.77	0.68	0.59	0.88
		$a_{D_{\text{KL}}^*}$	0.49	0.74	0.64	0.60	0.91
		a_{FR^*}	0.54	0.76	0.67	0.60	0.90
	Baselines	a_M	**0.71**	**0.88**	**0.81**	**0.41**	**0.60**

(c) spa-cat

Scenario		Score	AUPR-IN	AUPR-OUT	AUROC	ERR	FPR
s_0	Ours	a_{D_α}	**1.00**	**1.00**	**1.00**	**0.02**	**0.00**
		a_{FR}	**1.00**	**1.00**	**1.00**	0.03	0.01
	Baselines	a_E	0.99	0.97	0.98	0.04	0.04
		a_L	0.70	0.75	0.75	0.38	0.72
		a_{MSP}	0.98	0.99	0.99	0.05	0.04
s_1	Ours	$a_{D_\alpha^*}$	**0.99**	**0.99**	**0.99**	**0.03**	**0.02**
		$a_{D_{\text{KL}}^*}$	0.90	0.96	0.94	0.09	0.13
		a_{FR^*}	**0.99**	**0.99**	**0.99**	**0.03**	**0.02**
	Baselines	a_M	0.93	0.94	0.94	0.18	0.31

(d) spa-fr

Scenario		Score	AUPR-IN	AUPR-OUT	AUROC	ERR	FPR
s_0	Ours	a_{D_α}	**0.94**	**0.93**	**0.94**	**0.18**	**0.32**
		a_{FR}	**0.94**	**0.93**	**0.94**	0.20	0.34
	Baselines	a_E	0.82	0.75	0.79	0.44	0.83
		a_L	0.79	0.70	0.77	0.46	0.88
		a_{MSP}	0.77	0.77	0.74	0.45	0.84
s_1	Ours	$a_{D_\alpha^*}$	0.79	0.70	0.76	0.44	0.84
		$a_{D_{\text{KL}}^*}$	0.75	0.70	0.74	0.46	0.87
		a_{FR^*}	0.78	0.71	0.76	0.45	0.86
	Baselines	a_M	**0.92**	**0.95**	**0.94**	**0.15**	**0.24**

(e) spa-por

Scenario		Score	AUPR-IN	AUPR-OUT	AUROC	ERR	FPR
s_0	Ours	a_{D_α}	**1.00**	**0.99**	**1.00**	**0.03**	**0.00**
		a_{FR}	0.99	**0.99**	0.99	0.04	0.02
	Baselines	a_E	0.98	0.95	0.98	0.05	0.05
		a_L	0.77	0.77	0.80	0.37	0.69
		a_{MSP}	0.98	0.98	0.98	0.05	0.05
s_1	Ours	$a_{D_\alpha^*}$	**1.00**	**1.00**	**1.00**	**0.03**	**0.00**
		$a_{D_{\text{KL}}^*}$	0.97	0.99	0.99	0.04	0.03
		a_{FR^*}	**1.00**	**1.00**	**1.00**	**0.03**	**0.00**
	Baselines	a_M	0.99	**1.00**	0.99	0.04	0.03

(f) deu-nl

C.2 Detailed results of OOD detection performances

In this section we present the performances of our OOD detectors on each detailed tasks, *i.e.* for each pair of IN and OOD data with all the considered metrics. We show that our metrics outperform other OOD detectors baselines in almost all scenarios.

Table 9: Detailed results of the performances of our OOD detectors on different domain shifts. For Spanish (spa) and German (de), we present two domains shifts: Technical medical (EMEA) data and legal parlementary texts (parl) against common language emboddied by the Tatoeba dataset (tat).

Scenario		Score	AUPR-IN	AUPR-OUT	AUROC	ERR	FPR
$\mathbb{S}_0$	Ours	a_{D_α}	**0.90**	0.75	**0.86**	0.44	0.83
		a_{FR}	0.86	0.71	0.81	0.46	0.87
	Baselines	a_E	0.87	0.75	0.83	0.47	0.89
		a_L	0.86	0.73	0.82	0.48	0.90
		a_{MSP}	0.89	**0.76**	0.85	**0.43**	**0.81**
$\mathbb{S}_1$	Ours	$a_{D^*_\alpha}$	**0.89**	**0.87**	**0.89**	**0.27**	**0.49**
		$a_{D^*_{\text{KL}}}$	**0.89**	0.85	**0.89**	0.29	0.52
		a_{FR^*}	0.83	0.66	0.77	0.51	0.97
	Baselines	a_M	0.80	0.79	0.79	0.35	0.66

(a) de:news-EMEA

Scenario		Score	AUPR-IN	AUPR-OUT	AUROC	ERR	FPR
$\mathbb{S}_0$	Ours	a_{D_α}	**0.95**	**0.91**	**0.94**	**0.19**	**0.33**
		a_{FR}	0.91	0.89	0.91	0.24	0.42
	Baselines	a_E	0.91	0.83	0.90	0.32	0.59
		a_L	0.78	0.69	0.75	0.45	0.86
		a_{MSP}	0.90	0.88	0.89	0.36	0.68
$\mathbb{S}_1$	Ours	$a_{D^*_\alpha}$	**0.97**	0.94	**0.96**	0.13	0.20
		$a_{D^*_{\text{KL}}}$	0.95	**0.96**	**0.96**	**0.12**	**0.18**
		a_{FR^*}	0.95	0.89	0.94	0.25	0.44
	Baselines	a_M	0.92	0.92	0.92	0.20	0.35

(b) spa:tat-EMEA

Scenario		Score	AUPR-IN	AUPR-OUT	AUROC	ERR	FPR
$\mathbb{S}_0$	Ours	a_{D_α}	**0.75**	**0.75**	**0.76**	**0.35**	**0.66**
		a_{FR}	0.68	0.72	0.71	0.39	0.72
	Baselines	a_E	**0.75**	**0.75**	0.75	0.41	0.77
		a_L	0.65	0.69	0.71	0.42	0.79
		a_{MSP}	**0.75**	**0.75**	**0.76**	0.39	0.73
$\mathbb{S}_1$	Ours	$a_{D^*_\alpha}$	**0.89**	**0.88**	**0.90**	0.27	**0.48**
		$a_{D^*_{\text{KL}}}$	0.65	0.77	0.72	0.34	0.63
		a_{FR^*}	0.88	**0.88**	**0.90**	**0.26**	**0.48**
	Baselines	a_M	0.77	0.82	0.78	0.28	0.51

(c) spa:tat-parl

Scenario		Score	AUPR-IN	AUPR-OUT	AUROC	ERR	FPR
$\mathbb{S}_0$	Ours	a_{D_α}	**0.75**	**0.75**	**0.70**	**0.43**	**0.81**
		a_{FR}	0.60	0.64	0.63	0.45	0.85
	Baselines	a_E	**0.75**	**0.75**	0.67	0.45	0.85
		a_L	0.63	0.58	0.64	0.51	0.96
		a_{MSP}	**0.75**	**0.75**	0.67	0.46	0.87
$\mathbb{S}_1$	Ours	$a_{D^*_\alpha}$	**0.68**	**0.64**	**0.68**	**0.44**	**0.83**
		$a_{D^*_{\text{KL}}}$	0.65	0.62	0.67	0.48	0.90
		a_{FR^*}	0.62	0.63	0.65	0.45	0.86
	Baselines	a_M	0.64	0.62	0.63	0.46	0.87

(d) de:news-parl

Table 10: Detailed performance results of our OOD detectors on dialog shift against the Multi WOZ dataset as reference set.

Scenario		Score	AUPR-IN	AUPR-OUT	AUROC	ERR	FPR
$\mathbb{S}_0$	Ours	a_{D_α}	0.52	**0.87**	**0.72**	**0.52**	**0.69**
		a_{FR}	0.41	0.86	0.69	0.55	0.73
	Baselines	a_E	**0.63**	**0.87**	0.63	0.56	0.74
		a_L	0.40	0.66	0.47	0.73	1.00
		a_{MSP}	0.36	0.86	0.67	**0.52**	**0.69**
$\mathbb{S}_1$	Ours	$a_{D^*_\alpha}$	0.69	0.93	0.85	0.40	0.53
		$a_{D^*_{\text{KL}}}$	0.42	0.84	0.67	0.63	0.85
		a_{FR^*}	0.53	0.88	0.75	0.58	0.77
	Baselines	a_M	0.53	0.87	0.74	0.73	0.99

(a) melde

Scenario		Score	AUPR-IN	AUPR-OUT	AUROC	ERR	FPR
$\mathbb{S}_0$	Ours	a_{D_α}	0.42	**0.91**	**0.72**	0.57	**0.70**
		a_{FR}	0.41	0.90	**0.72**	0.57	0.71
	Baselines	a_E	**0.61**	0.89	0.63	0.61	0.76
		a_L	0.53	0.76	0.58	0.80	1.00
		a_{MSP}	0.35	0.90	0.68	**0.56**	**0.70**
$\mathbb{S}_1$	Ours	$a_{D^*_\alpha}$	0.67	0.95	0.86	0.44	0.55
		$a_{D^*_{\text{KL}}}$	0.29	0.85	0.63	0.72	0.90
		a_{FR^*}	0.41	0.90	0.73	0.61	0.77
	Baselines	a_M	0.42	0.89	0.70	0.78	0.99

(b) iemocap

Scenario		Score	AUPR-IN	AUPR-OUT	AUROC	ERR	FPR
$\mathbb{S}_0$	Ours	a_{D_α}	**0.87**	**0.87**	**0.87**	**0.31**	**0.56**
		a_{FR}	0.73	0.81	0.79	0.34	0.63
	Baselines	a_E	0.75	0.75	0.62	0.40	0.75
		a_L	0.81	0.79	0.82	0.39	0.72
		a_{MSP}	0.53	0.64	0.57	0.42	0.78
$\mathbb{S}_1$	Ours	$a_{D^*_\alpha}$	0.69	0.69	0.72	0.43	0.81
		$a_{D^*_{\text{KL}}}$	0.66	0.68	0.70	0.44	0.82
		a_{FR^*}	0.69	0.69	0.72	0.43	0.82
	Baselines	a_M	0.87	0.84	0.87	0.33	0.61

(c) dailydialog-default

Scenario		Score	AUPR-IN	AUPR-OUT	AUROC	ERR	FPR
$\mathbb{S}_0$	Ours	a_{D_α}	0.71	0.74	**0.72**	**0.36**	0.67
		a_{FR}	0.68	0.73	**0.72**	**0.36**	0.67
	Baselines	a_E	**0.75**	**0.75**	0.66	0.39	0.74
		a_L	0.69	0.48	0.57	0.50	1.00
		a_{MSP}	0.61	0.74	0.71	**0.36**	**0.66**
$\mathbb{S}_1$	Ours	$a_{D^*_\alpha}$	0.88	0.89	0.88	0.24	0.44
		$a_{D^*_{\text{KL}}}$	0.68	0.70	0.70	0.42	0.80
		a_{FR^*}	0.79	0.80	0.80	0.36	0.68
	Baselines	a_M	0.87	0.95	0.93	0.10	0.21

(d) swda

Scenario		Score	AUPR-IN	AUPR-OUT	AUROC	ERR	FPR
$\mathbb{S}_0$	Ours	a_{D_α}	0.71	**0.77**	**0.75**	0.33	0.60
		a_{FR}	0.69	0.74	0.74	0.36	0.67
	Baselines	a_E	**0.75**	0.75	0.70	0.37	0.68
		a_L	0.72	0.48	0.57	0.50	1.00
		a_{MSP}	0.62	0.76	0.73	**0.32**	**0.59**
$\mathbb{S}_1$	Ours	$a_{D^*_\alpha}$	0.86	0.88	0.87	0.23	0.42
		$a_{D^*_{\text{KL}}}$	0.66	0.71	0.70	0.41	0.77
		a_{FR^*}	0.77	0.80	0.79	0.36	0.66
	Baselines	a_M	0.88	0.96	0.94	0.10	0.21

(e) mrda

Scenario		Score	AUPR-IN	AUPR-OUT	AUROC	ERR	FPR
$\mathbb{S}_0$	Ours	a_{D_α}	0.79	**0.80**	0.81	**0.36**	**0.68**
		a_{FR}	0.68	0.73	0.70	0.39	0.74
	Baselines	a_E	0.75	0.75	0.64	0.45	0.84
		a_L	**0.91**	0.73	**0.85**	0.49	0.94
		a_{MSP}	0.70	0.65	0.64	0.43	0.80
$\mathbb{S}_1$	Ours	$a_{D^*_\alpha}$	0.92	0.91	0.91	0.23	0.42
		$a_{D^*_{\text{KL}}}$	0.60	0.60	0.61	0.47	0.89
		a_{FR^*}	0.75	0.74	0.75	0.41	0.77
	Baselines	a_M	0.85	0.93	0.91	0.10	0.21

(f) dyda-da

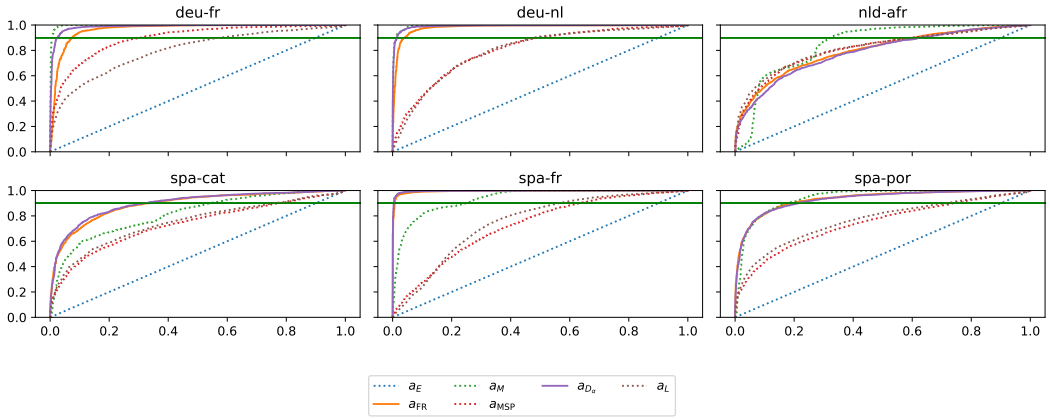

Figure 3: ROCAUC curves for our uncertainty based metrics compared to common baselines for language shifts detection. Baselines are represented in dashed lines.

C.3 ROC AUC CURVES

C.3.1 LANGUAGE SHIFTS

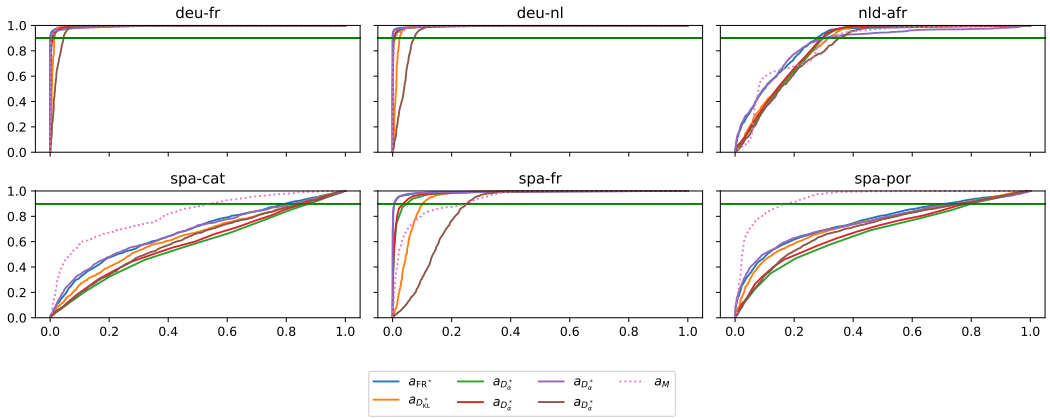

Figure 4: ROC-AUC curves for our reference based metrics compared to common baselines for language shifts detection. Baselines are represented in dashed lines.

C.3.2 DOMAIN SHIFTS

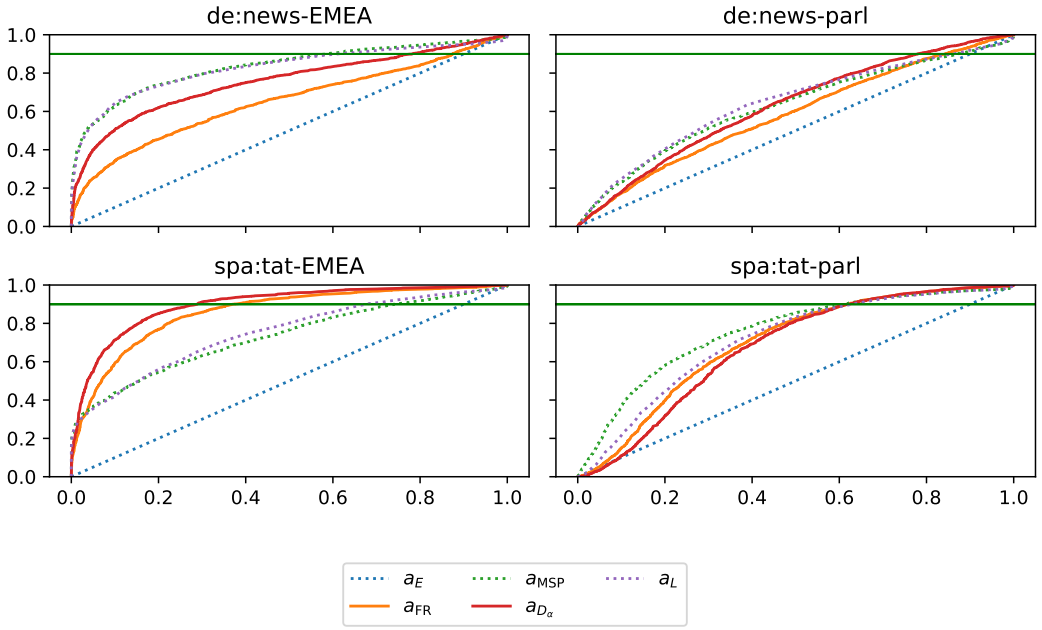

Figure 5: ROC-AUC curves for our uncertainty based metrics compared to common baselines for domain shifts detection. baselines are represented in dashed lines.

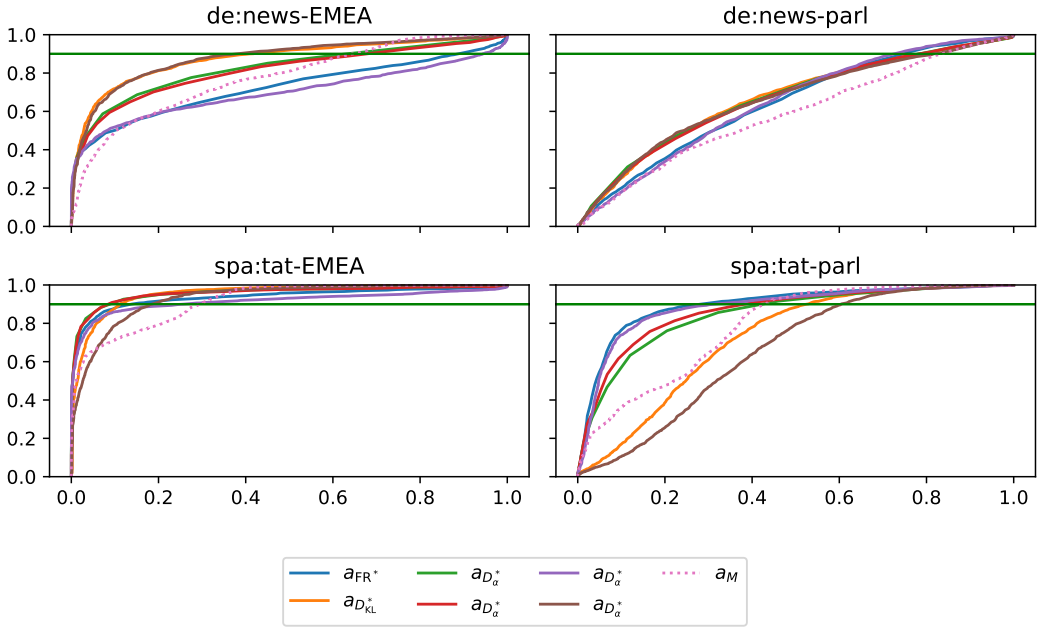

Figure 6: ROC-AUC curves for our reference based metrics compared to common baselines for domain shift detection. baselines are represented in dashed lines.

C.3.3 DIALOG SHIFTS

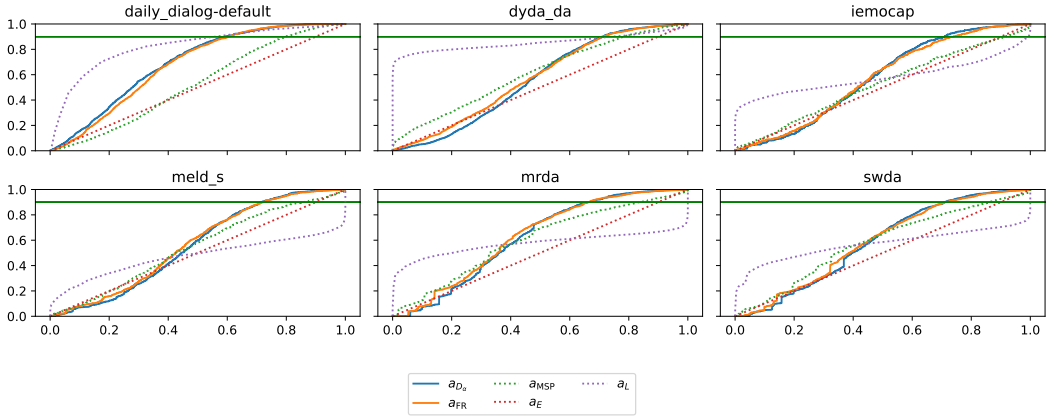

Figure 7: ROC-AUC curves for our uncertainty based metrics compared to common baselines for dialog shifts detection. baselines are represented in dashed lines.

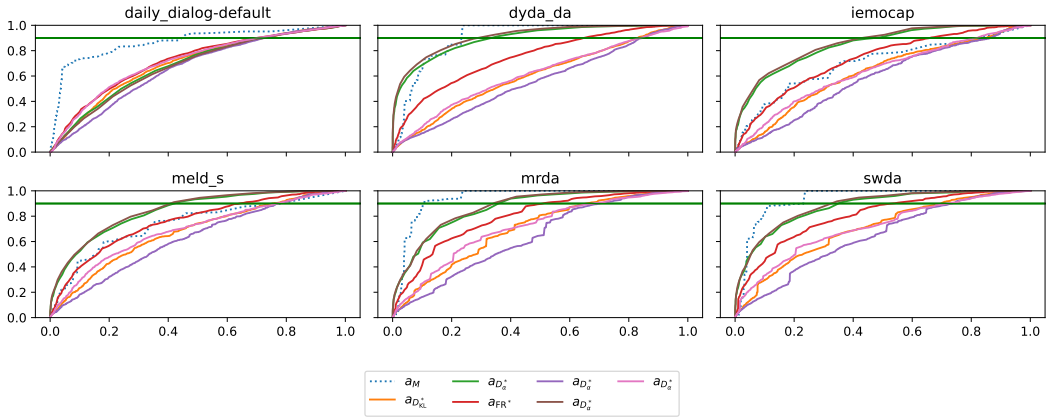

Figure 8: Rocauc curves for our reference based metrics compared to common baselines for dialog shift detection. baselines are represented in dashed lines.

D NTM PERFORMANCE

Surprisingly we show that common OOD detectors tend to exclude samples that are well handled by the model and keep some that are not leading to decreasing overall performance in terms of translation metrics. Moreoever it seems this phenomenon is more dominant in reference based detectors. We show that our uncertainty based detectors mostly avoir that downfall and provide good OOD detection and improved translation performances.

D.1 ABSOLUTE PERFORMANCES

It is clear (somewhat expected) that NMT models do not perform as well on OOD data as we can see in Tab. 11b. However, we find that our OOD detectors are able to remove most of the worst case samples and keep enough well translated samples so that with correct filtering our method actually allow the model to achieve somewhat acceptable BLEU scores.

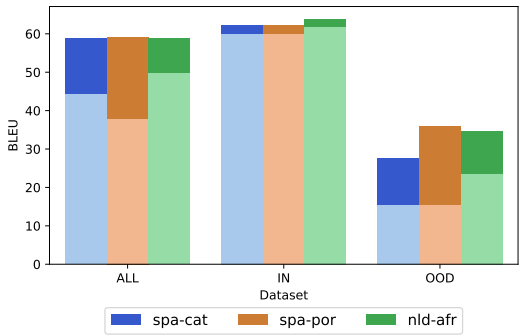

Figure 9: Gain in translation performances when filtering OOD samples with our method on different datasets and language pairs.

Scenario		Score	spa-cat	spa-por	nld-afr	spa:tat-parl	de:news-parl	spa:tat-EMEA	de:news-EMEA
		$\times$	59.89	59.89	61.72	59.89	34.28	59.89	34.28
$\mathbb{s}_0$	Ours	a_{D_α}	63.97	62.57	66.19	64.89	36.88	63.97	36.82
		a_{FR}	62.58	62.44	65.40	64.00	36.42	64.00	36.42
	Baselines	a_E	62.78	62.78	64.79	62.78	36.10	62.78	36.10
		a_L	64.74	64.74	66.71	64.74	36.93	64.74	36.93
		a_{MSP}	58.50	58.50	60.12	58.50	33.20	58.50	33.20
$\mathbb{s}_1$	Ours	$a_{D^*_\alpha}$	60.68	60.68	62.69	60.68	33.99	60.68	33.99
		a_{FR^*}	60.70	60.70	62.70	60.70	34.00	60.70	34.00
	Baselines	a_M	59.95	59.95	62.99	59.95	34.48	59.95	34.48

(a) IN

Scenario		Score	spa-cat	spa-por	nld-afr	spa:tat-parl	de:news-parl	spa:tat-EMEA	de:news-EMEA
		$\times$	15.49	15.38	23.52	33.29	28.64	59.48	52.26
$\mathbb{s}_0$	Ours	a_{D_α}	23.00	36.28	35.39	42.33	33.48	64.58	54.90
		a_{FR}	27.53	36.09	33.20	38.61	31.94	62.71	55.15
	Baselines	a_E	14.23	13.08	27.26	35.51	32.40	60.67	54.63
		a_L	16.60	14.53	32.80	39.97	34.44	61.36	54.46
		a_{MSP}	17.22	19.06	23.74	30.87	26.75	42.47	34.40
$\mathbb{s}_1$	Ours	$a_{D^*_\alpha}$	19.74	23.20	33.06	33.56	27.97	49.73	45.00
		a_{FR^*}	19.71	23.10	33.15	33.52	27.93	49.64	44.98
	Baselines	a_M	15.50	23.52	27.60	31.29	26.82	49.25	44.48

(b) OOD

Scenario		Score	spa-cat	spa-por	nld-afr	spa:tat-parl	de:news-parl	spa:tat-EMEA	de:news-EMEA
		$\times$	44.26	37.64	49.73	46.59	31.46	59.69	43.27
$\mathbb{s}_0$	Ours	a_{D_α}	59.77	59.81	62.58	57.16	35.40	64.27	46.49
		a_{FR}	58.93	59.77	60.40	53.45	34.35	63.33	46.31
	Baselines	a_E	48.69	40.43	57.95	49.15	34.50	61.63	45.96
		a_L	54.18	48.10	61.57	54.43	35.87	62.93	46.37
		a_{MSP}	46.34	42.71	47.86	44.15	29.75	55.93	33.49
$\mathbb{s}_1$	Ours	$a_{D^*_\alpha}$	49.91	48.54	58.77	56.85	30.75	59.63	37.73
		a_{FR^*}	49.92	48.48	58.83	56.84	30.75	59.64	37.75
	Baselines	a_M	51.82	56.41	57.46	48.57	30.97	57.74	37.80

(c) ALL

Table 11: Absolue translation performances in terms of BLEU on the different subset (IN, OOD, ALL) of each dataset of our translation OOD performance benchmark.

Scenario		Score	spa-cat	spa-por	nld-afr	spa:tat-parl	de:news-parl	spa:tat-EMEA	de:news-EMEA
$\mathbb{s}_0$	Ours	a_{D_α}	+4.08	+2.68	+4.46	+5.00	+2.60	+4.08	+2.53
		a_{FR}	+2.69	+2.54	+3.68	+4.11	+2.14	+4.11	+2.14
	Baselines	a_E	+2.89	+2.89	+3.07	+2.89	+1.82	+2.89	+1.82
		a_L	+4.85	+4.85	+4.98	+4.85	+2.65	+4.85	+2.65
		a_{MSP}	-1.39	-1.39	-1.60	-1.39	-1.09	-1.39	-1.09
$\mathbb{s}_1$	Ours	$a_{D^*_\alpha}$	+0.79	+0.79	+0.97	+0.79	-0.29	+0.79	-0.29
		a_{FR^*}	+0.81	+0.81	+0.97	+0.81	-0.29	+0.81	-0.29
	Baselines	a_M	+0.06	+0.06	+1.27	+0.06	+0.19	+0.06	+0.19

(a) IN

Scenario		Score	spa-cat	spa-por	nld-afr	spa:tat-parl	de:news-parl	spa:tat-EMEA	de:news-EMEA
$\mathbb{s}_0$	Ours	a_{D_α}	+7.51	+20.90	+11.87	+9.04	+4.83	+5.10	+2.64
		a_{FR}	+12.05	+20.71	+9.68	+5.32	+3.30	+3.23	+2.90
	Baselines	a_E	-1.25	-2.31	+3.74	+2.22	+3.76	+1.18	+2.37
		a_L	+1.11	-0.85	+9.27	+6.68	+5.80	+1.88	+2.20
		a_{MSP}	+1.73	+3.68	+0.22	-2.42	-1.89	-17.01	-17.86
$\mathbb{s}_1$	Ours	$a_{D^*_\alpha}$	+4.25	+7.81	+9.54	+0.27	-0.67	-9.75	-7.25
		a_{FR^*}	+4.22	+7.72	+9.63	+0.23	-0.71	-9.84	-7.28
	Baselines	a_M	+0.01	+8.14	+4.07	-2.00	-1.83	-10.23	-7.78

(b) OOD

Scenario		Score	spa-cat	spa-por	nld-afr	spa:tat-parl	de:news-parl	spa:tat-EMEA	de:news-EMEA
$\mathbb{s}_0$	Ours	a_{D_α}	+15.51	+22.17	+12.85	+10.57	+3.94	+4.58	+3.22
		a_{FR}	+14.67	+22.13	+10.67	+6.85	+2.89	+3.65	+3.04
	Baselines	a_E	+4.43	+2.79	+8.22	+2.56	+3.03	+1.94	+2.69
		a_L	+9.92	+10.46	+11.84	+7.84	+4.41	+3.24	+3.10
		a_{MSP}	+2.09	+5.07	-1.87	-2.44	-1.71	-3.76	-9.78
$\mathbb{s}_1$	Ours	$a_{D^*_\alpha}$	+5.65	+10.91	+9.04	+10.26	-0.71	-0.05	-5.55
		a_{FR^*}	+5.66	+10.85	+9.10	+10.25	-0.71	-0.04	-5.52
	Baselines	a_M	+7.56	+18.78	+7.73	+1.98	-0.50	-1.95	-5.47

(c) ALL

Scenario		Dataset Score	spa-cat ALL	spa-cat IN	spa-cat OOD	spa-por ALL	spa-por IN	spa-por OOD	nld-afr ALL	nld-afr IN	nld-afr OOD	spa:tat-parl ALL	spa:tat-parl IN	spa:tat-parl OOD	de:news-parl ALL	de:news-parl IN	de:news-parl OOD	spa:tat-EMEA ALL	spa:tat-EMEA IN	spa:tat-EMEA OOD	de:news-EMEA ALL	de:news-EMEA IN	de:news-EMEA OOD
$\mathbb{s}_0$	Ours	a_{D_α}	42%	19%	83%	55%	20%	91%	38%	20%	77%	39%	20%	58%	29%	20%	39%	21%	19%	23%	14%	20%	8%
		a_{FR}	42%	20%	83%	56%	20%	91%	33%	18%	67%	32%	20%	43%	26%	20%	31%	17%	20%	14%	15%	20%	10%
	Baselines	a_E	27%	20%	40%	27%	20%	35%	33%	20%	61%	20%	20%	20%	29%	20%	39%	12%	20%	4%	15%	20%	9%
		a_L	32%	19%	58%	39%	19%	60%	35%	19%	68%	30%	19%	42%	30%	20%	40%	12%	19%	6%	13%	20%	6%
		a_{MSP}	26%	19%	38%	33%	19%	46%	17%	20%	11%	16%	19%	13%	14%	20%	8%	52%	19%	85%	47%	20%	75%
$\mathbb{s}_1$	Ours	$a_{D^*_\alpha}$	30%	20%	47%	41%	20%	62%	37%	20%	73%	53%	20%	87%	13%	20%	7%	56%	20%	92%	39%	20%	59%
		a_{FR^*}	29%	20%	47%	40%	20%	61%	37%	20%	74%	53%	20%	87%	12%	18%	6%	56%	20%	92%	38%	18%	58%
	Baselines	a_M	37%	20%	67%	56%	20%	91%	35%	20%	68%	34%	20%	47%	26%	20%	32%	50%	20%	79%	40%	20%	60%

Table 13: Share of the datasets removed when taking γ so that we keep 80% of the IN distribution.

D.2 GAINS

D.3 EFFECT OF A LARGER THRESHOLD ON NMT PERFORMANCE

Table 15: Detailed impacts on NMT performance results per tasks (Domain- or Language-shifts) of the different OOD detectors with a threshold defined to keep 99% of the IN data. We present results on the different part of the data: IN data, OOD data and the combination of both, ALL. For each we report the absolute average BLEU score (Abs.), the average gains in BLEU (G.s) compared to a setting without OOD filtering (f_θ only) and the share of the subset removed by the detector (R.Sh.).

			Domain shifts									Language shifts								
			IN			OOD			ALL			IN			OOD			ALL		
			Abs.	G.s	R.Sh	Abs.	G.s	R.Sh	Abs.	G.s	R.Sh	Abs.	G.s	R.Sh	Abs.	G.s	R.Sh	Abs.	G.s	R.Sh
		$\times$	47.1	+0.0	0.0%	43.4	+0.0	0.0%	45.3	+0.0	0.0%	60.5	+0.0	0.0%	18.1	+0.0	0.0%	43.9	+0.0	0.0%
$\mathcal{S}_0$	Ours	a_{D_α}	47.3	+0.2	1.0%	44.2	+0.8	5.5%	45.8	+0.5	3.2%	60.7	+0.2	1.0%	21.8	+3.7	34.5%	49.2	+5.4	14.6%
		a_{FR}	47.3	+0.2	1.0%	44.1	+0.7	7.2%	45.7	+0.4	4.1%	60.7	+0.2	1.0%	22.3	+4.2	37.0%	49.7	+5.9	15.8%
	Bas.	a_E	47.3	+0.2	1.0%	44.0	+0.5	1.9%	45.6	+0.4	1.4%	60.9	+0.4	1.0%	18.7	+0.6	17.6%	46.0	+2.1	7.3%
		a_L	47.3	+0.2	0.9%	44.0	+0.6	1.9%	45.6	+0.4	1.4%	60.8	+0.3	0.9%	19.1	+0.9	18.4%	46.2	+2.3	7.6%
		a_{MSP}	47.0	-0.1	1.0%	40.3	-3.1	14.7%	43.5	-1.8	7.8%	60.4	-0.1	1.0%	18.5	+0.3	4.3%	44.3	+0.5	2.5%
$\mathcal{S}_1$	Ours	$a_{D^*_\alpha}$	47.0	-0.1	0.9%	40.3	-3.1	26.5%	43.9	-1.4	13.7%	60.5	-0.0	0.9%	19.3	+1.1	10.5%	45.3	+1.4	4.8%
		a_{FR^*}	47.0	-0.1	0.9%	40.3	-3.1	26.6%	43.9	-1.3	13.8%	60.5	-0.0	0.9%	19.3	+1.1	10.5%	45.3	+1.4	4.8%
	Bas.	a_M	47.0	-0.1	1.0%	41.6	-1.8	18.1%	44.6	-0.7	9.6%	60.5	-0.0	1.0%	18.4	+0.3	12.1%	45.3	+1.4	5.9%

E NEGATIVE RESULTS

E.1 DIFFERENT AGGREGATION OF OOD METRICS

Most of our detectors are initially classification OOD detectors that we adapted for text generation by averaging them over the generated sequences and using this aggregated score as a score for the whole sequence. We experimented with other aggregations such as the standard deviation or the min/max along the sequence. If the standard deviation gave relatively good results they were still less interesting that the naive average.

Table 14: Correlation between OOD scores and translation metrics BLEU and BERT-S

			BERT-S			BLEU		
			ALL	IN	OUT	ALL	IN	OUT
$\mathcal{S}_0$	Ours	a_{D_α}	-0.49	-0.32	-0.55	-0.38	-0.27	-0.40
		a_{FR}	-0.48	-0.33	-0.49	-0.39	-0.29	-0.41
	Base.	a_E	-0.17	-0.28	0.00	-0.15	-0.20	0.00
		a_L	-0.22	-0.25	0.21	-0.29	-0.27	-0.07
		a_{MSP}	-0.26	-0.14	-0.41	-0.13	-0.09	-0.13
$\mathcal{S}_1$	Ours	$a_{D^*_\alpha}$	-0.35	-0.21	-0.63	-0.27	-0.20	-0.37
		a_{FR^*}	-0.33	-0.21	-0.62	-0.25	-0.20	-0.39
	Base.	a_M	-0.27	-0.14	-0.22	-0.18	-0.10	-0.03

E.2 NEGENTROPY OF BAG OF DISTRIBUTIONS

We introduced in Sec. 3.3 the bag of distributions as a way to aggregate a sequence of probability distribution and compare it to a set of reference using information projections Sec. 3.3. A natural idea would be to apply the Negentropy methods (Sec. 3.2) to these aggregated distributions.

More formally given a sequence of probability distribution $\mathcal{S}_\theta(\mathbf{x}) = \{p_\theta^T(\mathbf{x}, \hat{\mathbf{y}}_{\leqslant t})\}_{t=1}^n$ we would compute its bag of distributions:

$$\bar{p}_\theta(x) \triangleq \frac{1}{|y|} \sum_{t=1}^{|y|} p_\theta(x, y_{\leqslant t}) \tag{7}$$

And then compute as novelty score:

$$J_D(\mathbf{p}) = D(\mathbf{p} \| \mathcal{U}) \tag{8}$$

Further experiments have shown that this process was unable to discriminate OOD samples or improve performance translation. We suspect that the uncertainty at each step is key to capture the behavior of the language model and that this uncertainty information is lost when averaging probability distribution along the sequence.