

Automatic Machine Translation Detection Using a Surrogate Multilingual Translation Model

Cristian García-Romero[†]

Miquel Esplà-Gomis^{†‡}

Felipe Sánchez-Martínez^{†‡}

[†]Dep. de Llenguatges i Sistemes Informàtics

[‡]Institut Universitari d'Investigació Informàtica

Universitat d'Alacant

E-03690 Sant Vicent del Raspeig (Spain)

{cristian.gr,miquel.espla,fsanchez}@ua.es

Abstract

Modern machine translation (MT) systems depend on large parallel corpora, often collected from the Internet. However, recent evidence indicates that (i) a substantial portion of these texts are machine-generated translations, and (ii) an overreliance on such synthetic content in training data can significantly degrade translation quality. As a result, filtering out non-human translations is becoming an essential pre-processing step in building high-quality MT systems. In this work, we propose a novel approach that directly exploits the internal representations of a surrogate multilingual MT model to distinguish between human and machine-translated sentences. Experimental results show that our method outperforms current state-of-the-art techniques, particularly for non-English language pairs, achieving gains of at least 5 percentage points of accuracy.

1 Introduction

Parallel corpora are an essential resource for the development of machine translation (MT) systems. They are used both to train models from scratch and to fine-tune pre-trained systems, particularly for domain adaptation and for enhancing performance in low-resource language pairs.

Automatic harvesting of parallel content from the Web is a common practice to build such corpora. However, some studies have highlighted a significant caveat: the Internet is increasingly populated with machine-translated content (Dodge et al., 2021; Ramírez-Sánchez et al., 2022; Thompson et al., 2024). Furthermore, empirical evidence shows that using large amounts of synthetic, automatically translated texts for training can degrade the translation performance of the resulting translation models (Wu et al., 2019; Jiao et al., 2021). This degradation—which has also been observed in generative-AI models

(Shumailov et al., 2024)—underscores the importance of distinguishing synthetic translations from human ones. The challenge is exacerbated by the fact that modern neural MT systems generate output that is not only grammatically fluent but also stylistically natural, making superficial quality checks insufficient.

In adjacent fields, such as the detection of AI-generated texts, a common approach is to leverage the very same large language models (LLMs) that produce the synthetic content to also detect it, following a white-box approach (Gehrmann et al., 2019). While promising, these self-referential methods rely on access to the generation model, which may not always be available. Consequently, recent research has increasingly focused on the use of surrogate models for AI-generated text detection (Mitchell et al., 2023; Venkatraman et al., 2024). These approaches extract interpretable features—such as log-probability distributions—from the internal representations of a surrogate model, which are then used to classify a given text as either human- or machine-generated.

Our approach builds on the latter research line but adopts a more flexible strategy: instead of relying on hand-crafted features derived from model internals, we directly use the latent representations from a pre-trained multilingual surrogate MT model—in our experiment, NLLB (NLLB Team et al., 2022)—to detect machine-translated sentences. This avoids potential information loss from manual feature extraction and allows the classifier to learn relevant patterns automatically. It also enables analysis of which components of the surrogate model (e.g., specific decoder blocks) are more informative. Crucially, by decoupling detection from the generation model, our method generalizes across all language pairs supported by the surrogate. Empirical evidence—shown in Fig. 1—further supports our approach: MT consistently shows lower per-word perplexity than

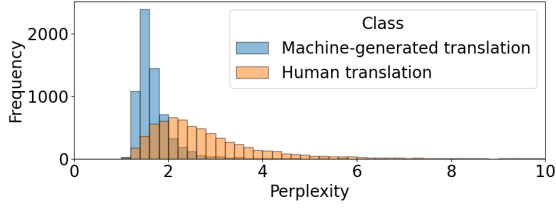


Figure 1: Per-word perplexity for human and MT-generated translations (produced with MADLAD) from our English–German training set (see Sec. 3). Perplexity is obtained with NLLB 3.3B. Similar trends are observed across language pairs and MT models.

human translations (HT) across a variety of MT generation models. Our findings show that pre-trained multilingual MT models encode useful signals for detecting machine-translated sentences, outperforming the current state of the art based on fine-tuned pre-trained multilingual encoder models (Chichirau et al., 2023).

The rest of the paper is organized as follows. The next section presents our approach in detail. Sec. 3 then outlines the experimental settings, whereas Sec. 4 reports and discusses the results. The final sections cover related work, conclusions, and a discussion of limitations.

2 Approach

We formulate the HT vs. MT problem as a binary classification task. Given a sentence pair (s, t) in languages S and T , where s is a (human-produced) source sentence and t is its translation (either HT or MT), we use a pre-trained multilingual MT model to obtain representations of t conditioned on s , which are then fed into a classifier.¹

Inspired by Sarvazyan et al. (2024), who address the detection of machine-generated text from LLMs, we propose a similar approach tailored to the HT vs. MT classification task, while avoiding the use of handcrafted features. In our method, illustrated in Fig. 2, the tokenized source sentence $s_{\text{tok}} = \{s_j\}_{j=1}^m$ is fed into a surrogate multilingual encoder-decoder MT model while the tokenized translation $t_{\text{tok}} = \{t_i\}_{i=1}^n$ is provided to the decoder via teacher forcing.² We then extract the target token-level representations $h_k^{(d)} = \{h_{k,i}^{(d)}\}_{i=1}^n$

¹A monolingual variant would consider only t .

²Adapting the method to a multilingual decoder-only model, such as EMMA-500 (Ji et al., 2024), would require minor adjustments.

from the hidden states of a chosen decoder block³ k , where $h_{k,i}^{(d)}$ is the hidden state of the k -th decoder block for token t_i . These representations are subsequently fed into a classifier after passing it through a projection layer to get $h^{(d')} = \{h_i^{(d')}\}_{i=1}^n$. This projection layer maps the surrogate model’s hidden states to the classifier’s embedding space and decouples the classifier from the surrogate model’s dimensionality (e.g., 1024 for NLLB 1.3B vs. 2048 for NLLB 3.3B). This ensures a consistent architecture across experiments, aiding interpretation and fair comparison across surrogate models of different sizes (see Sec. 4.1).

Our classifier is a transformer-based encoder with absolute positional embeddings to retain word order. It processes the output of the token-level representations $h^{(d')}$ generated by the projection layer and produces as output $h^{(c)} = \{h_i^{(c)}\}_{i=1}^n$. Following a BERT-style setup (Devlin et al., 2019), the representation of the first token, $h_1^{(c)}$, is then passed through a feed-forward layer to a single output neuron. This output is finally normalized with a sigmoid function to estimate the probability that the target sentence is an HT. We refer to this method as **Surrogate Machine Translation Detection**, or simply **SMTD**.

2.1 Combination with a Language Model

The classifier described above can optionally incorporate a representation from an encoder-based language model (LM) as additional input. This is done by processing a sequence of tokens consisting of a special $[\text{CLS}]$ token followed by the tokens $\{s'_j\}_{j=1}^{m'}$ and $\{t'_i\}_{i=1}^{n'}$, produced, respectively, by the LM’s tokenizer for s and t . Following a BERT-style approach (Devlin et al., 2019), we extract the hidden state of the first token, $h_{\text{CLS}}^{(l)}$, from the LM’s final block and pass it through a projection layer to yield $h_{\text{CLS}}^{(l')}$. $h_{\text{CLS}}^{(l')}$ is then prepended to the sequence of n token-level representations $h^{(d')}$ from the surrogate model to form the final input sequence for the transformer-based classifier. Finally, the representation of the first token produced as output by the transformer—in this case $h_{\text{CLS}}^{(c)}$, instead of $h_1^{(c)}$ —is used for the final classification (see Fig. 2).

Preliminary experiments show that the most effective and analytically insightful configuration

³Blocks refer to the components in a Transformer layer (Vaswani et al., 2017): attention mechanism, residual connections, layer normalization, and feed-forward network.

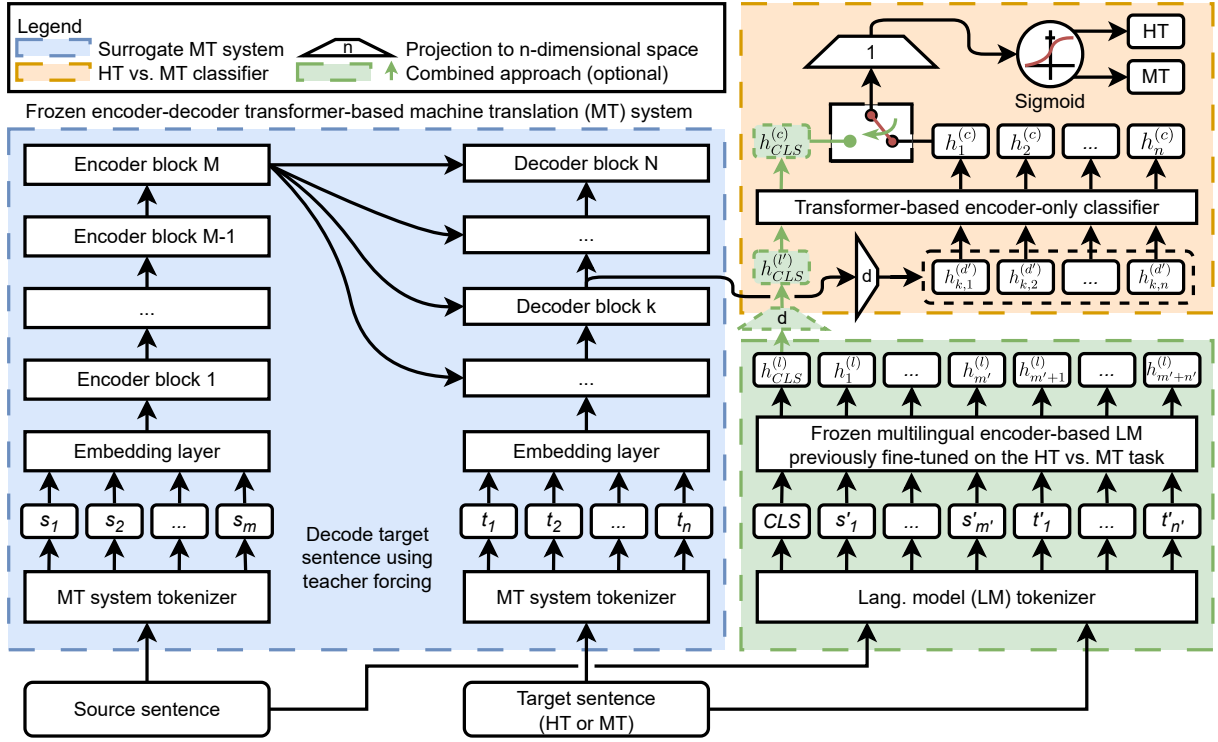


Figure 2: Architecture of SMatD/SMatD+LM. Some elements (e.g., positional embeddings) have been omitted for clarity. See Sec. 2 for a detailed explanation.

involves fine-tuning the LM for HT vs. MT classification, and then freezing it when integrating it with the surrogate MT representations. This two-stage approach isolates the contribution of the LM and facilitates quantifying the performance gains from combining both sources of information. We refer to this configuration as SMatD+LM.

3 Experimental Setting

This section outlines the datasets, evaluation metric and baselines used for comparison, followed by training details for both SMatD and SMatD+LM.

Datasets. We build on the dataset released by Chichirau et al. (2023), which leverages data from the WMT news translation shared tasks (2008–2019).⁴ This dataset consists of sentence pairs where the source segments are original, and the target segments are HT. It is English-centric, covering translations into English from German (de–en), Russian (ru–en), and Chinese (zh–en). The authors supplemented the existing human translations with additional MT outputs generated using

DeepL⁵ and Google Translate (hereafter Google).⁶

We extend this dataset by adding new MT systems and language pairs. The evaluated systems include state-of-the-art models: MADLAD-400 (Kudugunta et al., 2023) (multilingual; hereafter MADLAD or MADL), Opus-MT (Tiedemann and Thottingal, 2020) (bilingual; hereafter Opus)⁷ and Tower Instruct (Alves et al., 2024) (hereafter Tower).⁸ Tower is an instruction-tuned, decoder-only model; its inclusion enables us to examine whether such models behave differently from traditional encoder-decoder systems in distinguishing MT from HT.

The additional language pairs are: English–German (en–de), English–Russian (en–ru), German–Spanish (de–es), and Spanish–German (es–de). Data was collected from WMT using the same method as Chichirau et al. (2023), selecting editions based on language availability (see Table 1). We use the WMT 2019 edition for testing, the preceding one for development, and

⁵<https://www.deepl.com/translator>

⁶<https://translate.google.com/>

⁷Although Opus is bilingual and uses separate models per language pair, we refer to it as a single system for clarity.

⁸We use the prompt recommended by the authors: <https://huggingface.co/Unbabel/TowerInstruct-7B-v0.2>

⁴<https://www.statmt.org/wmt19/translation-task.html>

Lang. pair	Training	Development	Test
de-en	08–17	18	19
ru-en	15–17	18	19
en-de	08–11,13,15–17	18	19
en-ru	08–13,15–17	18	19
de-es	08–09	10	11,13
es-de	08–09	10	11,13
fi-en	-	-	19

Table 1: WMT editions used to create each data split.

earlier editions for training, except for de-es and es-de, where the WMT 2011 and 2013 editions are used for testing and the WMT 2010 edition is used for development.

Unlike Chichirau et al. (2023), we exclude Chinese from our experiments due to concerns about data reliability. First, zh-en exhibits unusually unstable per-token perplexity scores in our surrogate system, deviating from the consistent trends observed in Fig. 1. Second, Chichirau et al. (2023) also reported that zh-en behaves differently from the rest of language pairs. Third, translations of Chinese segments in the WMT datasets consistently yield significantly lower BLEU, chrF, and xCOMET scores—regardless of the MT system used in our experiments—compared to the rest of language pairs.

To evaluate generalization to unseen MT systems and language pairs, we introduce M2M-100 (Fan et al., 2021) (hereafter M2M) as a zero-shot MT system, and adopt Finnish–English (fi-en) data provided by Chichirau et al. (2023), generating additional translations with our MT systems, for zero-shot language evaluation.

Each source sentence is paired with its HT and with MT-generated translations from the corresponding MT systems, resulting in a balanced dataset per language and MT system. Table 2 shows the number of source sentences per language pair across the training, development, and test splits. Google, DeepL, and Tower were used for de-en and ru-en, while Tower, MADLAD, and Opus were used for the remaining pairs. For zero-shot evaluation, all MT systems generated fi-en translations, and M2M produced translations for all language pairs.

Evaluation. Following Chichirau et al. (2023) and Bhardwaj et al. (2020), we use accuracy as our evaluation metric. F1-score could also be used, but

Lang. pair	Sentences training	Sentences dev.	Sentences test
de-en	8,242	1,498	2,000
ru-en	4,382	1,500	2,000
en-de	6,501	1,500	1,997
en-ru	4,935	1,500	1,997
de-es	793	500	1,101
es-de	685	499	1,104
fi-en	-	-	1,996

Table 2: Number of source sentences per language pair in the train, dev., and test sets. Each source sentence is paired with its HT and the corresponding MT-generated translations. Counts for ru-en differ from Chichirau et al. (2023, Table 1) but align with the corresponding WMT test sets.

it has been shown to strongly correlate with accuracy for HT vs. MT classification using pre-trained LMs (van der Werff et al., 2022).

Baselines. We compare our approach to three different baselines: LLMixtic (Sarvazyan et al., 2024), a state-of-the-art method for detecting machine-generated text, and two other methods based on the current state of the art for detecting MT-generated translations (Chichirau et al., 2023).

LLMixtic (Sarvazyan et al., 2024), utilizes large language models (LLMs) as surrogate models and achieved state-of-the-art results in the automatic detection of monolingual machine-generated text, ranking first (out of 125 submissions) in the SemEval-2024 Task 8 (Wang et al., 2024). For our experiments, we used the authors’ implementation⁹ with the best-reported configuration of LLaMA-2 models (Touvron et al., 2023)—Llama-2-7b-hf, Llama-2-7b-chat-hf, Llama-2-13b-hf, and Llama-2-13b-chat-hf—and minor modifications—unlimited training with a patience of 6 epochs, and an inverse square root learning rate scheduler.¹⁰

The two other baselines correspond to two versions of the state-of-the-art method proposed by Chichirau et al. (2023) for detecting MT-generated translations, which uses a multilingual encoder-based LM. Specifically, we adopt mDeBERTaV3_{Base} (He et al., 2023) (hereafter

⁹<https://github.com/jogonba2/llmixtic>

¹⁰We used LLaMA-2 for all language pairs in our experiments because the newer LLaMA-3.1 models (Grattafiori et al., 2024) (Llama-3.1-8B and Llama-3.1-8B-Instruct) did not yield superior results for non-English language pairs, despite LLaMA-2 officially only supporting English.

mDeBERTaV3) in two configurations: a *monolingual* setup, where only the target sentence is used as input, and a *bilingual* setup, where both the source and target sentences are provided. The latter matches the configuration used for the LM in SMaTD+LM. As with SMaTD+LM, we use the first token ([CLS]) embedding to estimate the probability that the target sentence is HT or MT. We follow the best configuration from Chichi-[rau et al. \(2023\)](#), with minor adjustments: unlimited training with a patience of 6 epochs, 10% dropout (classifier, attention, and feed-forward layers), learning rate of 10^{-5} , and inverse square root learning rate scheduler with 400 warm-up steps. We use the AdamW optimizer (Loshchilov and Hutter, 2019) with default settings and no weight decay. All reported results were obtained using our reimplementation, which successfully reproduced the original results.

SMaTD training. For SMaTD, we use a standard encoder-only Transformer (Vaswani et al., 2017) as the binary classifier, trained from scratch, and NLLB-200 (NLLB Team et al., 2022) as the surrogate multilingual translation model.

We adopt the same training setup as the Chichi-[rau et al. \(2023\)](#)’s baselines, but with a learning rate of 10^{-4} and 10% dropout on positional embeddings. Based on preliminary experiments we use 3 encoder layers, 4 attention heads, a 2048-dimensional feed-forward layer, and a model dimensionality of 512. The parameters of the surrogate model remain frozen during training.

SMaTD+LM training. We follow the two-stages training procedure described in Sec. 2.1, using the same mDeBERTaV3 language model from the bilingual baseline, which is pre-trained on the HT vs. MT classification task. To prevent the classifier from relying solely on mDeBERTaV3 and ignore the surrogate model, we apply *stochastic depth* (Huang et al., 2016), a form of dropout at the representation level. Based on preliminary experiments, we set its probability to 0.7. All other hyperparameters match those used for training mDeBERTaV3 (baselines) and SMaTD.

4 Results and Discussion

We begin by analyzing how the size of the surrogate MT model and the choice of decoder block impact classification performance (Sec. 4.1). We then evaluate SMaTD and SMaTD+LM (Sec. 4.2),

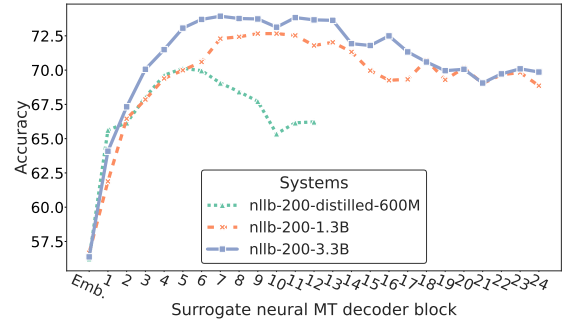


Figure 3: Accuracy on the development set for DeepL (de-en) evaluated using three different sizes of the NLLB surrogate MT model. For each surrogate model, we evaluate the use of the hidden state of different decoder blocks as input to the classifier; note that the 600M model has fewer decoder blocks by design.

examining their accuracy, their generalization to unseen languages and MT systems (Sec. 4.3), and their transferability when trained on multilingual or multi-MT datasets (Sec. 4.4).

All reported accuracy scores, computed over the test set, correspond to models selected based on their performance on the development set. Specifically, for each approach, we selected, out of three independent training executions, the model checkpoint that achieved the highest development set accuracy within its run (using early stopping with exhaustion of patience). See Appendix A for details on the variation over training executions. Statistically significant differences are reported in all tables of this section, as computed via approximate randomization (10,000 iterations; $p < 0.05$).

4.1 Surrogate MT Model Size and Layer Evaluation

Fig. 3 shows the accuracy for de-en with DeepL, comparing three NLLB surrogate model sizes (600M, 1.3B, 3.3B) across all decoder blocks. The results show that the 3.3B model consistently outperforms the others models, with 1.3B outperforming 600M at every decoder block. Similar patterns were observed for other language pairs and MT systems. Based on these observations, we report results exclusively for the 3.3B model in subsequent experiments.

Interestingly, Fig. 3 shows a consistent pattern: both the first and last decoder blocks underperform compared to middle blocks. To identify the best-performing block, we aggregate the accuracy across all language pairs and MT systems on the

development set: the lowest decoder block (embedding layer) yields the weakest performance, with an accuracy of 58.70%, while the 10th block achieves the highest accuracy (72.92%).¹¹ To assess the sensitivity of decoder block choice, we compute the difference in accuracy between the 10th block and the best-performing block in each experiment. The results indicate that performance is highly robust to the block choice (median: -0.68 ; mean: -0.98 ; std. dev.: 1.23). Accordingly, all subsequent experiments were performed using the 10th decoder block.¹²

4.2 Main Results

Table 3 shows the results for the baseline approaches—monolingual LLMixtic and language model (mLM), and the bilingual language model (LM)—together with SMaTD, and SMaTD+LM, the combination of the last two (see Sec. 2.1).

The results show a clear trend: either SMaTD+LM or SMaTD consistently outperform all baselines. In particular, SMaTD yields results comparable to the bilingual baseline (within ± 1 accuracy point) in certain cases (e.g., de-en for all MT systems). However, in most cases, SMaTD achieves statistically-significant improvements of at least one accuracy point. Remarkably, SMaTD+LM outperforms the LM baseline by a statistically-significant margin in all cases.

Notably, the results for non-English language pairs are particularly strong, with gains of at least 5 accuracy points and even 10 in some cases (de-es for MADLAD, es-de for Opus and Tower). Since the bilingual baseline model is English-centric—it has been trained on CC-100 (Conneau et al., 2020, Fig. 1), where English is the most represented language—we argue that it might face greater challenges with non-English language pairs. In contrast, our surrogate model has been trained on multiple language directions beyond English-centric pairs (NLLB Team et al., 2022, Sec. 8.1). This difference may explain, at least partially, why SMaTD achieves such significant accuracy improvements. In line with this observation, SMaTD+LM generally outperforms SMaTD when applied to language pairs including English, while SMaTD performs best in the rest of cases.

¹¹Across decoder blocks, aggregated performance has a median of 69.87%, a mean of 68.53% and a standard deviation of 4.05.

¹²Note that the 10th layer may not be optimal in the cross-evaluation experiments reported in Sec. 4.3.

Train. & eval. lang.	Approach	Train. & eval. MT system		
		Google	DeepL	Tower
de-en	LLMixtic	49.53	49.98	56.70
	mLM	<u>63.95</u>	<u>63.10</u>	<u>61.45</u>
	LM	<u>75.52</u>	<u>72.53</u>	<u>70.38</u>
	SMaTD	75.88	72.90	71.32
	SMaTD+LM	†77.28	†75.25	†71.80
ru-en	LLMixtic	49.23	50.38	49.25
	mLM	<u>59.12</u>	<u>59.58</u>	<u>56.60</u>
	LM	<u>70.32</u>	<u>64.05</u>	<u>64.07</u>
	SMaTD	74.40	65.40	67.40
	SMaTD+LM	†74.08	†65.10	†67.42
		MADL.	Opus	Tower
en-de	LLMixtic	52.35	50.48	53.93
	mLM	<u>59.49</u>	<u>65.20</u>	<u>56.08</u>
	LM	<u>70.76</u>	<u>74.39</u>	<u>60.72</u>
	SMaTD	<u>72.26</u>	<u>76.09</u>	60.14
	SMaTD+LM	†74.21	†78.54	†62.27
en-ru	LLMixtic	51.18	53.98	51.83
	mLM	<u>63.52</u>	<u>68.00</u>	<u>58.76</u>
	LM	<u>68.70</u>	<u>70.38</u>	<u>60.79</u>
	SMaTD	<u>73.99</u>	71.48	64.17
	SMaTD+LM	†74.06	†73.41	†64.62
de-es	LLMixtic	55.86	51.50	55.86
	mLM	58.31	66.26	60.63
	LM	66.35	66.39	63.49
	SMaTD	80.20	75.11	71.71
	SMaTD+LM	†77.34	†73.66	†70.48
es-de	LLMixtic	55.53	50.41	50.82
	mLM	<u>59.10</u>	<u>62.64</u>	<u>55.30</u>
	LM	<u>67.30</u>	63.68	<u>61.68</u>
	SMaTD	<u>76.95</u>	77.31	72.46
	SMaTD+LM	†78.03	†73.55	†64.31

Table 3: Accuracy obtained by the five approaches under evaluation. Best scores are in bold. Scores whose difference from the score immediately above is statistically significant are underlined. † indicates statistically significant differences between SMaTD+LM and LM.

Contrary to expectations, the monolingual baseline LLMixtic performed the worst of all systems, even underperforming mLM. This poor result persisted despite exploring alternative configurations (e.g., LLaMA-3.1; see Sec. 3). Although LLMixtic excelled at SemEval’s English AI-text detection task—a result we were able to successfully replicate with the authors’ implementation—it failed to converge on our MT detection task. While a complete analysis is beyond our scope, this failure may be attributed to significant distri-

Train. & test lang.	Train. on MT system	Test LM on MT system				
		Google	DeepL	Tower	M2M	
de-en	Google	75.52	66.68	66.35	67.20	
	DeepL	72.85	72.53	68.83	64.60	
	Tower	70.42	67.30	70.38	65.43	
ru-en	Google	70.32	58.32	61.68	66.70	
	DeepL	63.40	64.05	57.73	59.35	
	Tower	65.72	58.10	64.07	64.17	
		MADL.	Opus	Tower	M2M	
en-de	MADL.	70.76	68.23	61.67	68.90	
	Opus	68.03	74.39	59.44	73.31	
	Tower	54.98	55.53	60.72	57.04	
en-ru	MADL.	68.70	61.64	58.19	70.56	
	Opus	65.70	70.38	56.61	71.58	
	Tower	61.97	57.74	60.79	64.62	
de-es	MADL.	66.35	61.76	63.67	64.35	
	Opus	67.57	66.39	64.31	67.17	
	Tower	64.08	58.08	63.49	61.85	
es-de	MADL.	67.30	66.08	62.91	68.39	
	Opus	63.45	63.68	60.24	64.58	
	Tower	63.09	62.36	61.68	63.81	

Train. & test lang.	Train. on MT system	Test SMTD on MT system				
		Google	DeepL	Tower	M2M	
de-en	Google	75.88	<u>70.05</u>	<u>69.60</u>	70.80	
	DeepL	74.30	72.90	70.10	<u>68.25</u>	
	Tower	<u>73.58</u>	<u>70.60</u>	71.32	<u>68.20</u>	
ru-en	Google	74.40	<u>61.43</u>	<u>64.47</u>	72.03	
	DeepL	<u>67.35</u>	65.40	<u>60.37</u>	<u>63.30</u>	
	Tower	<u>73.10</u>	<u>61.98</u>	67.40	<u>70.67</u>	
		MADL.	Opus	Tower	M2M	
en-de	MADL.	72.26	<u>71.58</u>	60.82	<u>73.36</u>	
	Opus	<u>68.48</u>	76.09	57.51	75.19	
	Tower	<u>59.81</u>	<u>62.57</u>	60.14	<u>65.47</u>	
en-ru	MADL.	73.99	68.43	<u>62.24</u>	76.16	
	Opus	<u>70.51</u>	71.48	<u>60.94</u>	<u>73.61</u>	
	Tower	<u>67.78</u>	<u>65.35</u>	64.17	<u>69.00</u>	
de-es	MADL.	80.20	<u>72.66</u>	64.35	<u>77.34</u>	
	Opus	<u>78.25</u>	75.11	65.49	77.70	
	Tower	<u>74.21</u>	<u>67.39</u>	71.71	<u>73.25</u>	
es-de	MADL.	76.95	76.40	63.13	<u>79.12</u>	
	Opus	<u>75.95</u>	77.31	60.73	79.62	
	Tower	<u>76.49</u>	<u>74.86</u>	72.46	<u>78.99</u>	

Table 4: Accuracy obtained by the bilingual baseline (LM; left table) and SMTD (right table) in the cross-MT setting. Rows correspond to the MT systems for which models were trained, and columns to the MT systems on which models were evaluated. The M2M model is only used for testing (zero-shot). Best scores are in bold. Underlined values indicate statistically significant superior performance across systems.

butional differences between MT and general AI-generated text, to LLMixtic requiring substantially more data than our dataset provided (though the dataset sufficed for the other models), or to the SemEval dataset being comparatively easier to classify.

In the following sections, only the bilingual baseline is used as a reference for two reasons: (i) it outperforms both monolingual baselines (LLMixtic and mLM) in all cases, and (ii) like SMTD, it leverages both source and target texts, enabling a fairer comparison.

4.3 Cross-Evaluation

We analyze the generalization capabilities of the bilingual baseline (LM) and SMTD when applied to languages and MT systems not seen during training in cross-MT and cross-lingual settings.

Cross-MT. Table 4 shows the results for the cross-MT evaluation. As expected, the best results are located along the diagonal (i.e., when the same

MT system is used for training and testing).¹³ Furthermore, SMTD consistently matches or surpasses the baseline when the MT systems used for training and testing are different. The de-en case is particularly interesting: while diagonal results for SMTD are similar to the baseline, most off-diagonal values are significantly higher, suggesting strong generalization. Notable improvements are also observed in the zero-shot setting (M2M): SMTD achieves statistically significant gains in accuracy in all cases, especially for de-es and es-de, with gains of 10–15 accuracy points, compared to diagonal differences of 5–10 points.

Cross-Lingual. Table 5 presents the results of the cross-lingual evaluation. As for the cross-MT evaluation, models trained on a particular language generally perform best when evaluated on the same language.¹⁴ Regarding off-diagonal results, SMTD yields stronger generalization. Notably, the results for DeepL improve cross-

¹³Diagonal results correspond to rows of Table 3.

¹⁴Diagonal results correspond to columns of Table 3.

Train. & test MT	Approach	Train. lang.	Test on lang.						
			de-en	ru-en	en-de	en-ru	de-es	es-de	fi-en
Google	LM	de-en	75.52	68.35	-	-	-	-	73.90
		ru-en	67.77	70.32	-	-	-	-	71.99
	SMaTD	de-en	75.88	74.50	-	-	-	-	77.18
		ru-en	69.85	74.40	-	-	-	-	72.04
DeepL	LM	de-en	72.53	59.00	-	-	-	-	66.01
		ru-en	62.37	64.05	-	-	-	-	66.31
	SMaTD	de-en	72.90	63.67	-	-	-	-	70.54
		ru-en	63.42	65.40	-	-	-	-	66.38
Tower	LM	de-en	70.38	67.37	58.69	55.13	58.36	56.20	66.28
		ru-en	63.85	64.07	55.68	52.35	55.31	55.48	64.18
		en-de	59.13	58.10	60.72	59.49	61.13	66.44	51.63
		en-ru	56.85	57.95	57.79	60.79	58.95	61.19	51.95
		de-es	59.97	61.83	52.58	52.68	63.49	57.20	56.39
		es-de	58.13	57.67	57.96	54.86	60.35	61.68	51.83
	SMaTD	de-en	71.32	69.82	57.39	61.27	62.85	57.07	61.87
		ru-en	65.47	67.40	57.84	56.94	60.35	58.92	58.64
		en-de	56.48	55.58	60.14	60.49	67.26	68.34	48.27
		en-ru	61.87	63.38	59.91	64.17	68.17	67.48	54.66
		de-es	51.80	53.03	53.73	53.58	71.71	59.10	46.39
		es-de	60.43	60.40	58.21	60.12	68.35	72.46	50.63
MADL.	LM	en-de	-	-	70.76	66.40	55.81	64.54	56.99
		en-ru	-	-	61.57	68.70	56.36	59.60	55.84
		de-es	-	-	53.43	55.51	66.35	59.78	58.44
		es-de	-	-	62.49	59.59	64.35	67.30	58.72
	SMaTD	en-de	-	-	72.26	72.53	77.61	74.91	67.76
		en-ru	-	-	66.35	73.99	77.70	64.04	67.08
		de-es	-	-	63.40	69.48	80.20	60.91	66.78
		es-de	-	-	67.88	68.45	71.80	76.95	70.74
Opus	LM	en-de	-	-	74.39	66.07	55.04	64.18	57.04
		en-ru	-	-	65.75	70.38	57.18	64.67	54.01
		de-es	-	-	54.53	54.01	66.39	59.47	58.37
		es-de	-	-	62.59	56.79	60.54	63.68	56.91
	SMaTD	en-de	-	-	76.09	63.75	69.35	63.22	64.28
		en-ru	-	-	67.23	71.48	73.39	74.64	60.25
		de-es	-	-	71.93	65.52	75.11	69.47	67.16
		es-de	-	-	71.91	60.64	75.07	77.31	65.01

Table 5: Accuracy for the bilingual baseline (LM) and SMaTD in the cross-lingual setting. Rows correspond to the language pairs for which models were trained, and columns to the language pairs on which models were evaluated. The language pair fi-en is only used for testing (zero-shot). Best scores are in bold. Underlined values indicate statistically significant superior performance across systems.

lingual accuracy by up to 6 accuracy points, despite similar performance on the diagonal. Tower shows some inconsistency: SMaTD achieves better cross-lingual results across all MT systems when evaluating en-ru, de-es, es-de and en-de, but results are mixed for the rest of language pairs; nevertheless, most improvements are statistically significant. For Opus and MADLAD, SMaTD achieves statistically-significant higher accuracies

in most cases. In the zero-shot setting (fi-en), SMaTD consistently outperforms the baseline, particularly when models are trained on de-en, or de-es and es-de when de-en is not present (i.e., MADLAD and Opus).

4.4 Transferability Assessment

This section explores the capabilities of SMaTD and the bilingual baseline (LM) to transfer knowl-

edge across languages and/or MT systems. To do so, we create two new datasets: (i) a multi-MT dataset that compiles, for each language pair, all MT samples in our initial dataset (see Sec. 3); and (ii) a multilingual dataset that combines, for each MT system, data from all available language pairs.

Class Balance in New Datasets. The multi-MT dataset is unbalanced in terms of the HT/MT class distribution: for each HT sample, there are three MT samples, each generated by a different MT system. To mitigate this during training, we dynamically sample a single MT sample per HT sample. As a result, a different class-balanced dataset is obtained for each training epoch.

The multilingual dataset is class-balanced but contains a different number of samples per language, leading to a language-unbalanced distribution. Following Arivazhagan et al. (2019), we apply temperature-based sampling, where each language l is assigned a sampling probability proportional to $p_l^{\frac{1}{T}}$, where p_l denotes the original proportion of samples for language l . Specifically, we use $\frac{1}{T} = 0.3$, as suggested by NLLB Team et al. (2022). Similarly to the multi-MT dataset, we obtain a different dataset per training epoch.

In contrast to the dynamic generation of training samples, described above, development datasets remain unchanged once created for the multilingual setting. In the multi-MT setting, we build new class-balanced development datasets, where the MT system is randomly sampled in advance, ensuring consistency across experiments. With this strategy we aim to reduce the risk of model bias due to the sensitivity of the evaluation metric to class imbalance. To ensure comparability across experiments, the test sets are the same used in the experiments reported in previous sections. Appendix A reports on the variability across transferability experiments, using new test sets created via the same methodology as the development datasets described above.

Multi-MT Results. Table 6 presents the results for the multi-MT evaluation, where SMTD consistently outperforms the baseline by a statistically significant margin in all languages and MT systems, except for en–de. This trend also holds in the zero-shot setting with M2M, where en–de is the only exception where the baseline outperforms SMTD. To assess transferability, we compare these results with those in Table 4. We observe

that (i) in most cases, models trained on all available MT systems do not outperform those trained on a single language and MT system (i.e., diagonal values) and (ii) both approaches consistently perform better across all languages for cross-MT results (i.e., off-diagonal values), except for de–es and es–de in the baseline (likely due to its English-centric pretraining). These results suggest moderate positive transferability, with minimal accuracy differences on average.

Multilingual Results. Table 7 presents the results on the multilingual evaluation. While the results for SMTD and the baseline are not consistent across MT systems, SMTD outperforms the baseline in most cases for different languages per MT system, except for some mixed results with Opus;¹⁵ this trend also holds for nearly all results in the zero-shot language pair (fi–en). In contrast to models trained on multiple MT systems above, models trained on multiple languages per MT system generally outperform those trained on a single language and MT system (see diagonal values in Table 5). These results suggest that incorporating training data from multiple languages improves transferability for both approaches, enabling a single model to operate across multiple languages, albeit for a single MT system.

The results above raise the question whether training a model on both multiple languages and multiple MT systems could further improve performance. Experiments conducted combining data from the multilingual and multi-MT datasets show that, consistent with our earlier findings, training on multiple languages improves performance, but incorporating translations from multiple MT systems degrades performance. In any case, a model trained on all available data achieves competitive results compared to models trained on a single language and system, offering the advantage of broader coverage with a single model.¹⁶

5 Related Work

Several studies have explored differences between human and machine-generated translations (Vanmassenhove et al., 2019; Roberts et al., 2020; Luo et al., 2024). However, modern MT systems now

¹⁵Opus is a bilingual MT system, where each language pair corresponds to a different MT model, unlike the other MT systems we evaluate. It is unclear whether Google and DeepL operate as bilingual systems.

¹⁶Results not provided due to space constraints.

Train. & test lang.	Test of LM on MT system				Train. & test lang.	Test of SMaTD on MT system			
	Google	DeepL	Tower	M2M		Google	DeepL	Tower	M2M
de-en	71.68	69.42	68.65	65.82	de-en	75.35	72.43	70.45	69.65
ru-en	65.97	60.50	61.52	63.05	ru-en	70.28	64.15	64.90	66.43
	MADL.	Opus	Tower	M2M		MADL.	Opus	Tower	M2M
en-de	68.58	70.16	61.79	70.73	en-de	66.07	67.73	60.59	69.50
en-ru	69.45	65.52	58.36	73.39	en-ru	71.98	68.80	61.19	75.59
de-es	62.62	60.31	60.76	62.08	de-es	78.25	74.02	67.35	76.79
es-de	60.73	60.64	59.33	60.82	es-de	75.91	76.18	71.38	78.53

Table 6: Accuracy obtained by the bilingual baseline (LM; left table) and SMaTD (right table) in the multi-MT setting. Results are provided per MT system (columns) for models trained on all the available MT systems for each language pair (rows). M2M is only used for testing (zero-shot). Best scores across approaches are in bold. Underlined values indicate statistically significant superior performance across systems.

Train. & test MT	Approach	Test on lang. data							
		de-en	ru-en	en-de	en-ru	de-es	es-de	fi-en	
Google	LM	74.80	72.60	-	-	-	-	76.53	
	SMaTD	75.52	75.85	-	-	-	-	78.43	
DeepL	LM	72.80	65.70	-	-	-	-	70.69	
	SMaTD	72.83	67.35	-	-	-	-	73.35	
Tower	LM	70.92	68.27	62.19	64.27	70.57	70.47	66.66	
	SMaTD	71.47	72.22	62.79	67.70	75.30	76.36	62.83	
MADL.	LM	-	-	69.03	70.46	71.71	71.51	66.31	
	SMaTD	-	-	70.28	77.17	82.74	82.25	72.44	
Opus	LM	-	-	75.21	74.16	73.93	77.67	66.03	
	SMaTD	-	-	75.84	71.06	77.84	81.97	65.16	

Table 7: Accuracy obtained by the bilingual baseline (LM) and SMaTD in the multilingual setting. Results are provided per language pair (columns) for models trained on all the available language pairs for a given MT system (rows). The language pair fi-en is only used for testing (zero-shot). Best scores across approaches are in bold. Underlined values indicate statistically significant superior performance across systems.

produce highly convincing texts, making the HT vs. MT classification task increasingly challenging—evidence suggests that higher-quality MT outputs are harder to detect (Aharoni et al., 2014).

Early approaches focused on identifying statistical MT outputs by leveraging fluency and linguistic features (Arase and Zhou, 2013; Li et al., 2015). Nguyen-Son and Echizen (2018) was the first to address neural MT detection, focusing on distinguishing MT from original texts (rather than human translations) using n-gram-based fluency and noise features. Most studies have concentrated on sentence-level classification, primarily motivated by the fact that neural MT systems are typically trained at this level, despite emerging trends toward coarser granularity (Kocmi et al., 2024). Bhardwaj et al. (2020) used feature-based

models, recurrent neural networks, and pre-trained monolingual and multilingual transformers across several English–French domains. Fu and Nederhof (2021) analyzed lexical diversity via n-grams and BERT models for English translations from multiple source languages. Nguyen-Son et al. (2021) proposed an alternative approach measuring differences in machine-generated texts iteratively back-translating them.

Given that classification becomes more challenging for shorter texts (Bhardwaj et al., 2020; Nguyen-Son et al., 2021), several studies have explored coarser granularities beyond the sentence level. Nguyen-Son et al. (2017) leveraged Zipfian distributions at the document level, while Nguyen-Son et al. (2018, 2019) assessed paragraph coherence. van der Werff et al. (2022) found that

Train. & eval. lang.	Approach	Train. & eval. MT system											
		Google				DeepL				Tower			
		Min.	Mean	Max.	SD	Min.	Mean	Max.	SD	Min.	Mean	Max.	SD
de-en	LLMixtic	49.53	50.20	50.55	0.58	49.18	49.65	49.98	0.42	49.43	53.42	56.70	3.69
	mLM	63.40	63.81	64.07	0.36	61.95	62.72	63.13	0.67	61.10	61.34	61.48	0.21
	LM	75.15	75.44	75.65	0.26	71.95	72.17	72.53	0.31	69.60	69.94	70.38	0.40
	SMaTD	74.80	75.44	75.88	0.57	72.90	73.48	74.00	0.55	71.00	71.27	71.47	0.24
	SMaTD+LM	77.28	77.64	77.93	0.33	74.15	74.59	75.25	0.58	71.80	72.21	72.50	0.36
ru-en	LLMixtic	48.90	49.19	49.45	0.28	50.38	50.57	50.68	0.17	49.25	49.38	49.60	0.20
	mLM	58.25	58.97	59.53	0.65	57.17	58.00	59.58	1.36	51.82	54.16	56.60	2.39
	LM	68.30	69.79	70.75	1.31	62.75	63.73	64.38	0.86	62.60	63.17	64.07	0.79
	SMaTD	73.43	73.89	74.40	0.49	64.75	64.98	65.40	0.36	66.62	67.40	68.17	0.78
	SMaTD+LM	71.72	72.92	74.08	1.18	64.90	64.98	65.10	0.10	65.97	66.70	67.42	0.73
		MADLAD				Opus				Tower			
en-de	LLMixtic	49.97	50.78	52.35	1.37	50.00	50.39	50.70	0.36	53.71	53.94	54.18	0.24
	mLM	58.96	59.20	59.49	0.27	64.55	64.85	65.20	0.33	55.18	56.19	57.31	1.07
	LM	70.23	70.45	70.76	0.27	73.69	73.99	74.39	0.36	60.72	60.92	61.04	0.18
	SMaTD	72.26	72.98	73.71	0.73	75.34	75.76	76.09	0.39	51.48	57.41	60.62	5.14
	SMaTD+LM	73.69	74.04	74.24	0.31	77.92	78.22	78.54	0.31	60.27	60.94	62.27	1.15
en-ru	LLMixtic	49.62	50.27	51.18	0.81	50.50	52.49	53.98	1.79	51.75	51.83	51.90	0.08
	mLM	56.61	61.37	63.97	4.13	68.00	69.09	69.78	0.95	57.06	57.79	58.76	0.88
	LM	61.79	66.25	68.70	3.87	70.38	70.91	71.38	0.50	60.77	60.97	61.37	0.34
	SMaTD	71.83	73.11	73.99	1.13	68.60	69.80	71.48	1.50	62.59	63.40	64.17	0.79
	SMaTD+LM	74.06	75.54	76.46	1.29	69.38	71.74	73.41	2.10	64.45	64.61	64.77	0.16
de-es	LLMixtic	55.86	56.51	57.18	0.66	51.00	51.38	51.63	0.33	53.13	54.81	55.86	1.47
	mLM	56.72	57.55	58.31	0.80	55.95	61.11	66.26	5.15	60.08	60.79	61.67	0.81
	LM	63.03	64.37	66.35	1.75	65.40	65.85	66.39	0.51	63.35	63.58	63.90	0.28
	SMaTD	76.52	78.29	80.20	1.84	74.98	75.58	76.66	0.93	51.00	64.65	71.71	11.83
	SMaTD+LM	73.43	75.61	77.34	1.99	73.66	74.69	76.07	1.24	64.67	67.35	70.48	2.93
es-de	LLMixtic	53.62	54.47	55.53	0.97	50.09	50.39	50.68	0.29	50.45	50.79	51.09	0.32
	mLM	57.84	58.62	59.10	0.69	62.64	64.13	66.17	1.83	52.85	54.15	55.30	1.23
	LM	63.59	65.08	67.30	1.96	63.68	64.52	65.81	1.13	59.78	60.98	61.68	1.04
	SMaTD	74.23	75.75	76.95	1.39	75.00	76.39	77.31	1.22	69.16	70.73	72.46	1.66
	SMaTD+LM	69.11	73.32	78.03	4.48	73.55	75.23	78.12	2.52	61.37	62.58	64.31	1.54

Table 8: Variability over the test set for the experiments reported in Sec. 4.2. Note that the values in the column reporting the maximum do not always match those in Table 3, as those correspond to the training execution achieving the highest accuracy on the development set (see Sec. 4). SD refers to standard deviation.

document-level evaluation outperforms sentence-level detection for German–English texts, using SVMs and various pre-trained monolingual transformer-based classifiers. Building on this, Chichirau et al. (2023) extended the analysis to seven source languages for English translations, incorporating multilingual models.

Concurrent to our work, Chen et al. (2025) propose combining a surrogate speech model with a monolingual encoder-based LM to differentiate original, human-produced (non-translated) text from MT-generated text. Our approach differs,

not only in the addressed task, but also in that we use a neural MT surrogate model and leverage its internal representations without fine-tuning, making it broadly applicable across the languages supported by the surrogate. Moreover, their motivation lies in leveraging speech features to better capture the linguistic nuances and deviations characteristic of machine-generated text, whereas our approach is motivated by modeling the underlying probability distributions of human and machine-generated translations directly, using a surrogate model trained for MT—which is more aligned

Train. & eval. lang.	Approach	Eval. on all MT systems			
		Min.	Mean	Max.	SD
de-en	LM	69.73	69.86	69.92	0.12
	SMaTD	72.87	73.19	73.65	0.41
ru-en	LM	61.83	62.60	63.17	0.70
	SMaTD	65.97	66.72	67.35	0.70
en-de	LM	64.87	66.36	67.25	1.30
	SMaTD	62.97	64.15	65.35	1.19
en-ru	LM	59.54	61.79	63.85	2.16
	SMaTD	64.70	67.52	70.28	2.79
de-es	LM	59.99	61.25	62.53	1.27
	SMaTD	50.00	65.18	73.12	13.15
es-de	LM	59.74	60.63	61.96	1.17
	SMaTD	72.74	73.49	74.32	0.80

Table 9: Variability over the multi-MT test set described in Sec. 4.4 and the corresponding multi-MT experiments presented in that section. SD refers to standard deviation.

with the HT vs. MT classification task.

6 Concluding Remarks

We have presented an approach, SMaTD, that directly leverages the internal representations of a surrogate neural MT model to address the HT vs. MT classification task. Additionally, we have also introduced SMaTD+LM, an extension that integrates an encoder-based language model.

Results across different surrogate model sizes indicate that larger models achieve better performance, with particularly strong results when using representations from middle decoder blocks. Importantly, we improve upon the current state of the art for this task, with notable gains—exceeding 5 accuracy points—for language pairs that do not involve English. By leveraging representations from a multilingual surrogate model that is not primarily focused on English, our approach effectively addresses non-English language pairs.

Cross-evaluation experiments show strong generalization capabilities, while transferability experiments indicate improvements when transferring across languages, but some degradation when transferring across MT systems.

Finally, the strength of our approach is further validated in a zero-shot setting, where our method achieves gains of 10–15 accuracy points on non-English pairs using an unseen MT system and language.

Train. & eval. MT	Approach	Eval. on all lang. data			
		Min.	Mean	Max.	SD
Google	LM	72.37	73.05	73.68	0.66
	SMaTD	74.06	74.15	74.26	0.10
DeepL	LM	70.63	71.46	72.27	0.82
	SMaTD	71.74	72.08	72.31	0.30
Tower	LM	67.27	67.60	68.20	0.53
	SMaTD	69.10	69.44	69.86	0.39
MADL	LM	72.97	73.73	74.25	0.68
	SMaTD	78.72	79.10	79.64	0.49
Opus	LM	77.14	77.79	78.51	0.69
	SMaTD	78.23	78.44	78.64	0.21

Table 10: Variability over the multilingual test set described in Sec. 4.4 and the corresponding multilingual experiments presented in that section. SD refers to standard deviation.

Code, datasets, and models are available at <https://github.com/transducens/SMaTD>.

7 Limitations

Our analysis is limited to the news domain, and we have not evaluated the robustness of our approach across domains. We also relied on a single surrogate system, without exploring alternative encoder-decoder or decoder-only architectures, or how specific components affect HT vs. MT classification. Additionally, the MT outputs used in our experiments were primarily generated by state-of-the-art encoder-decoder models with standard decoding strategies, leaving the effects of alternative decoding methods—as well as other paradigms, such as statistical or rule-based systems, and architectures like decoder-only transformers or recurrent neural networks—underexplored. Finally, our work does not address translation detection for systems trained at coarser granularities, such as document-level translation.

A Variability Across Training Executions

Table 8 reports the variability of the accuracy computed on the test set across three training executions with different random seeds for the main experiments (Sec. 4.2). Variability for the transferability experiments (Sec. 4.4) are presented in Table 9 and Table 10 for the multi-MT and multilingual experiments, respectively.

Acknowledgments

We thank the action editor and anonymous reviewers of the manuscript for their valuable comments which helped us improve the paper.

This paper is part of the work conducted in R+D+i projects PID2021-27999NB-I00 and PID2024-158157OB-C31 funded by the Spanish Ministry of Science and Innovation (MCIN), the Spanish Research Agency (AEI/10.13039/501100011033) and the European Regional Development Fund A way to make Europe. Cristian García-Romero is funded by Generalitat Valenciana and the European Social Fund through the research grant CIACIF/2021/365. Some of the computational resources used were funded by the Valencia Government and the European Regional Development Fund (ERDF) through project IDIFEDER/2020/003.

References

- Roei Aharoni, Moshe Koppel, and Yoav Goldberg. 2014. [Automatic detection of machine translated text and translation quality estimation](#). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 289–295, Baltimore, Maryland. Association for Computational Linguistics.
- Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and Andre Martins. 2024. [Tower: An open multilingual large language model for translation-related tasks](#). In *First Conference on Language Modeling*.
- Yuki Arase and Ming Zhou. 2013. [Machine translation detection from monolingual web-text](#). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1597–1607, Sofia, Bulgaria. Association for Computational Linguistics.
- Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. 2019. [Massively multilingual neural machine translation in the wild: Findings and challenges](#). *arXiv preprint arXiv:1907.05019*.
- Shivendra Bhardwaj, David Alfonso Hermelo, Phillippe Langlais, Gabriel Bernier-Colborne, Cyril Goutte, and Michel Simard. 2020. [Human or neural translation?](#) In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6553–6564, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Yongjian Chen, Mireia Farrús, and Antonio Toral. 2025. [The potential of speech features to discriminate between original and machine-translated texts](#). In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.
- Malina Chichirau, Rik van Noord, and Antonio Toral. 2023. [Automatic discrimination of human and neural machine translation in multilingual scenarios](#). In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 217–226, Tampere, Finland. European Association for Machine Translation.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groen-

- evel, Margaret Mitchell, and Matt Gardner. 2021. [Documenting large webtext corpora: A case study on the colossal clean crawled corpus](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. [Beyond english-centric multilingual machine translation](#). *Journal of Machine Learning Research*, 22(107):1–48.
- Yingxue Fu and Mark-Jan Nederhof. 2021. [Automatic classification of human translation and machine translation: A study from the perspective of lexical diversity](#). In *Proceedings for the First Workshop on Modelling Translation: Translatology in the Digital Age*, pages 91–99, online. Association for Computational Linguistics.
- Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. 2019. [GLTR: Statistical detection and visualization of generated text](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 111–116, Florence, Italy. Association for Computational Linguistics.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits,
- Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Gefert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yearly, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Celebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Rapparthi, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Van-

denhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vitor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuwei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkan Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippas Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Geor-

gia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damla, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabza, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Grosse, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lind-

- say, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. [The llama 3 herd of models](#).
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. [DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing](#). In *The Eleventh International Conference on Learning Representations*.
- Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q. Weinberger. 2016. [Deep networks with stochastic depth](#). In *Computer Vision – ECCV 2016*, pages 646–661. Springer International Publishing.
- Shaoxiong Ji, Zihao Li, Indraneil Paul, Jaakko Paavola, Peiqin Lin, Pinzhen Chen, Dayyán O’Brien, Hengyu Luo, Hinrich Schütze, Jörg Tiedemann, and Barry Haddow. 2024. [EMMA-500: Enhancing massively multilingual adaptation of large language models](#). *arXiv preprint arXiv:2409.17892*.
- Rui Jiao, Zonghan Yang, Maosong Sun, and Yang Liu. 2021. [Alternated training with synthetic and authentic data for neural machine translation](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1828–1834, Online. Association for Computational Linguistics.
- Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinhórf Steingrímsson, and Vilém Zouhar. 2024. [Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.
- Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. [MADLAD-400: A multilingual and document-level large audited dataset](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 67284–67296. Curran Associates, Inc.
- Yitong Li, Rui Wang, and Hai Zhao. 2015. [A machine learning method to distinguish machine translation from human translation](#). In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation: Posters*, pages 354–360, Shanghai, China.
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). In *International Conference on Learning Representations*.
- Jiaming Luo, Colin Cherry, and George Foster. 2024. [To diverge or not to diverge: A morphosyntactic perspective on machine translation vs human translation](#). *Transactions of the Association for Computational Linguistics*, 12:355–371.
- Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. [DetectGPT: Zero-shot machine-generated text detection using probability curvature](#). In *Proceedings of the 40th International Conference on Machine Learning*, pages 24950–24962.

- Hoang-Quoc Nguyen-Son and Isao Echizen. 2018. [Detecting computer-generated text using fluency and noise features](#). In *International Conference of the Pacific Association for Computational Linguistics*, pages 288–300, Singapore. Springer Singapore.
- Hoang-Quoc Nguyen-Son, Huy H. Nguyen, Ngoc-Dung T. Tieu, Junichi Yamagishi, and Isao Echizen. 2018. [Identifying computer-translated paragraphs using coherence features](#). In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, Hong Kong. Association for Computational Linguistics.
- Hoang-Quoc Nguyen-Son, Tran Thao, Seira Hidano, Ishita Gupta, and Shinsaku Kiyomoto. 2021. [Machine translated text detection through text similarity with round-trip translation](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5792–5797, Online. Association for Computational Linguistics.
- Hoang-Quoc Nguyen-Son, Tran Phuong Thao, Seira Hidano, and Shinsaku Kiyomoto. 2019. [Detecting machine-translated paragraphs by matching similar words](#). In *International Conference on Computational Linguistics and Intelligent Text Processing*, pages 521–532. Springer.
- Hoang-Quoc Nguyen-Son, Ngoc-Dung T. Tieu, Huy H. Nguyen, Junichi Yamagishi, and Isao Echi Zen. 2017. [Identifying computer-generated text using statistical analysis](#). In *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1504–1511. IEEE.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Gonzalez Mejia, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. [No language left behind: Scaling human-centered machine translation](#). *arXiv preprint arXiv:2207.04672*.
- Gema Ramírez-Sánchez, Marta Bañón, Jaume Zaragoza-Bernabeu, and Sergio Ortiz Rojas. 2022. [Human evaluation of web-crawled parallel corpora for machine translation](#). In *Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 32–41, Dublin, Ireland. Association for Computational Linguistics.
- Nicholas Roberts, Davis Liang, Graham Neubig, and Zachary C. Lipton. 2020. [Decoding and diversity in machine translation](#). In *Proceedings of the Resistance AI Workshop at 34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, Vancouver, CA.
- Areg Mikael Sarvazyan, José Ángel González, and Marc Franco-salvador. 2024. [Genaios at SemEval-2024 task 8: Detecting machine-generated text by mixing language model probabilistic features](#). In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 101–107, Mexico City, Mexico. Association for Computational Linguistics.
- Iliia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. [AI models collapse when trained on recursively generated data](#). *Nature*, 631(8022):755–759.
- Brian Thompson, Mehak Dhaliwal, Peter Frisch, Tobias Domhan, and Marcello Federico. 2024. [A shocking amount of the web is machine translated: Insights from multi-way parallelism](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1763–1775, Bangkok, Thailand. Association for Computational Linguistics.
- Jörg Tiedemann and Santhosh Thottingal. 2020. [OPUS-MT – building open translation services for the world](#). In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lis-

boa, Portugal. European Association for Machine Translation.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. [Llama 2: Open foundation and fine-tuned chat models](#).

Eva Vanmassenhove, Dimitar Shterionov, and Andy Way. 2019. [Lost in translation: Loss and decay of linguistic richness in machine translation](#). In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 222–232, Dublin, Ireland. European Association for Machine Translation.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010, Long Beach, USA.

Saranya Venkatraman, Adaku Uchendu, and Dongwon Lee. 2024. [GPT-who: An information density-based machine-generated text detector](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 103–115, Mexico City, Mexico. Association for Computational Linguistics.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, and Thomas Arnold. 2024. [SemEval-2024 task 8: Multidomain, multi-model and multilingual machine-generated text detection](#). In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 2057–2079, Mexico City, Mexico. Association for Computational Linguistics.

Tobias van der Werff, Rik van Noord, and Antonio Toral. 2022. [Automatic discrimination of human and neural machine translation: A study with multiple pre-trained models and longer context](#). In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 161–170, Ghent, Belgium. European Association for Machine Translation.

Lijun Wu, Yiren Wang, Yingce Xia, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. 2019. [Exploiting monolingual data at scale for neural machine translation](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4207–4216, Hong Kong, China. Association for Computational Linguistics.