Pulling Out All The Full Stops: Punctuation Sensitivity in Neural Machine Translation and Evaluation

Prathyusha Jwalapuram
Rakuten Institute of Technology
Rakuten Group, Inc.
p.jwalapuram@rakuten.com

Abstract

Much of the work testing machine translation systems for robustness and sensitivity has been adversarial or tended towards testing noisy input such as spelling errors, or non-standard input such as dialects. In this work, we take a step back to investigate a sensitivity problem that can seem trivial and is often overlooked: punctuation. We perform basic sentence-final insertion and deletion perturbation tests with full stops, exclamation and questions marks across source languages and demonstrate a concerning finding: commercial, production-level machine translation systems are vulnerable to mere single punctuation insertion or deletion, resulting in unreliable translations. Moreover, we demonstrate that both string-based and model-based evaluation metrics also suffer from this vulnerability, producing significantly different scores when translations only differ in a single punctuation, with model-based metrics penalizing each punctuation differently. Our work calls into question the reliability of machine translation systems and their evaluation metrics, particularly for real-world use cases, where inconsistent punctuation is often the most common and the least disruptive noise.[1]

1 Introduction and Related Work

Since the advent of the Transformer models (Vaswani et al., 2017), machine translation (MT) has seen tremendous improvement in performance, with several claims of parity with human translations (Wu et al., 2016; Hassan et al., 2018; Popel et al., 2020). However, one issue that is common to most deep learning models but does not hinder humans is sensitivity to small changes in the input, or a lack of robustness.

Robustness in machine translation refers to the ability of the models to produce consistent translations that preserve the meaning of the source sentence regardless of any noise in the input (Heigold et al., 2018). Changes in the input that preserve the semantics should not significantly change the output of the models. This can be a particularly critical quality for commercial machine translation systems, which are expected to translate real-world data including social media or internet text, which tend to be non-standard and noisy (Li et al., 2019).

From	Example
Ours	Iran gibt britischen Tanker frei Iran gibt britischen Tanker frei**!**
Niu et al. (2020)	Se kyllä **tuntuu** sangen luultavalta. Se kyllä **tumtuu** sangen luultavalta.
Michel et al. (2019)	Si seulement je pouvais me muscler **aussi** rapidement. Si seulement je pouvais me muscler **asusi** rapidement.
Ebrahimi et al. (2018)	... er ist Geigenbauer und **Psychotherapeut**. ... er ist Geigenbauer und **Psy6hothearpeiut**.
Tan et al. (2020)	When **is** the suspended team **scheduled** to **return**? When **are** the suspended team **schedule** to **returned**?
Wallace et al. (2020)	Did you know that adversarial examples can transfer to production models Did you know that adversarial examples can transfer to production models **Siehe Siehe Siehe Siehe Siehe Siehe Siehe**

Table 1: Common perturbations used in various robustness tests compared to our punctuation insertion. Original words in **bold**, perturbed text highlighted in **red**.

Models are typically tested for robustness by changing the input to introduce noise, called a perturbation, and checking whether the output is different. Several works have documented the sensitivity of machine translation models to various kinds of noise which commonly occurs in real-world data (Belinkov and Bisk, 2018; Niu et al., 2020; Tan et al., 2020). There has also been work on adversarial attacks, where algorithms with access to model gradients try to find optimal perturbations that result in a significant performance drop, or manipulate the model into producing malicious output (Ebrahimi et al., 2018; Wallace et al., 2020; Zhang et al., 2021). Most of these works have concentrated on robustness to variations in orthography and grammar. Table 1 shows some examples.

There has also been some work on MT evaluation metric robustness that has included similar pertur-

[1] https://github.com/rakutentech/Punctuation-NMT-ACL2023

Findings of the Association for Computational Linguistics: ACL 2023, pages 6116–6130
July 9-14, 2023 ©2023 Association for Computational Linguistics

bations at the character and word-level, and other linguistic phenomena such as synonyms, named entities, negation, numbers, and others (Sun et al., 2022; Freitag et al., 2022; Karpinska et al., 2022).

However, Michel et al. (2019) argue that many of these perturbations do not preserve the meaning on the source side. They propose that "meaning-preserving" perturbations should be limited to nearest neighbours in the embedding space and out-of-vocabulary word-internal character swaps.

In this work, we take a further step back from meaning-preserving spelling and grammatical perturbations, and ask: are machine translation models robust to **trivial changes in sentence-final punctuation**? Are the **metrics** used to evaluate machine translation robust to the same changes?

To investigate this, we test basic punctuation variation for which robustness may have been taken for granted. We perform simple sentence-final punctuation perturbations, restricting the experiments to two settings: insertion and deletion. Mimicking a very common form of natural noise, we insert or delete full stops, exclamation marks and question marks at the end of the input sentence (§2; see Table 1 for an example). Unlike common perturbation strategies, we make no changes to the content, words, or characters which may cause out-of-vocabulary or unseen tokens in the input. Our goal in this work is not to induce as drastic a drop in performance as possible, but to investigate the changes in translation that result from extremely minimal perturbations, and whether we are adequately able to detect these changes.

We test commercial MT systems from **Google**, **DeepL** and **Microsoft** on 3 language pairs from across resource levels and scripts: **German (De)**, **Japanese (Ja)** and **Ukrainian (Uk)** to **English (En)**. These systems are intended for real-world use, and can therefore be expected to already be robust to common noise in real-world data.

We first investigate whether commonly used evaluation metrics are robust to our perturbations, in order to ensure that our subsequent evaluation of the MT systems is fair (§3). We find that both string-based and model-based evaluation metrics are not robust to trivial sentence-final punctuation perturbations, significantly penalizing text with mismatched full stops, question marks or exclamations, sometimes more than text with more severe perturbations such as insertion or deletion of random characters.

Based on these results, we deviate from the standard robustness testing regime of perturbing the inputs and expecting the translations of both the original and the perturbed source text to match exactly. In the MT setting, adding a punctuation to the source text can naturally induce the model to also produce the corresponding punctuation in the translation. We therefore reset the punctuation changes in the translations in order to perform evaluation, and call for a review of standard MT robustness evaluation in such settings.

More importantly, we show that even commercial machine translation systems are extremely sensitive to trivial punctuation changes, particularly in languages such as Japanese and Ukrainian (§4). We show that both insertion and deletion of punctuation causes performance drops, which indicates that models may be biased to expect (or not expect) punctuation in certain types of sentences. We conduct a manual analysis and find that in more severe cases, a mere punctuation change can cause complete changes in the meaning of the translation or introduce hallucinations such as negation, with less severe changes including pronouns, named entities, tense, number, and others (§5).

Søgaard et al. (2018) provide some common examples of punctuation variation in real-world data and demonstrate how dependency parsers are sensitive to such punctuation differences. Ek et al. (2020) demonstrate the sensitivity of neural models to punctuation in Natural Language Inference tasks. Though there has also been some work on punctuation-based perturbation for machine translation (Bergmanis et al., 2020; Karpinska et al., 2022), the tendency has been to make more extreme perturbations than we adopt. Unlike previous work, we do not combine all punctuation changes into one bucket, and instead analyse each punctuation separately. We find that models are more sensitive to some punctuation than others. We also unify the usually independent work on machine translation robustness and evaluation metric robustness, and adjust our evaluation based on our observations.

Our work exposes serious real-world use-case implications and serves to show that while great strides have been made in both machine translation and its evaluation, we are a long way from building systems that are reliable for real-world use.

2 Test Set Creation

In this section, we describe the original test sets and perturbation operations we perform to build our test sets. Our perturbations reflect natural noise in punctuation occurrence: we only insert or delete punctuation such as full stops, exclamation marks and question marks from the ends of sentences.

2.1 Original Test Data

In order to build our perturbation test sets, we need a large test set with naturally occurring noise, *e.g.,* sentences which originally do not have full stops at the end (for insertion) or sentences ending with question marks (for deletion). Test sets typically have a majority of sentences ending with full stops, while other punctuation or punctuation-less sentences occur less often. In order to maximize these sentences, we combine test sets across FLORES101 (Goyal et al., 2021) and WMT 2020-2022 (Barrault et al., 2020, 2021; Kocmi et al., 2022) in both directions for German (De, high-resource), Japanese (Ja, medium-resource) and Ukrainian (Uk, medium-resource) to English (En).[2] We choose these 3 language pairs to optimize for diversity in resource levels and scripts, while ensuring we have adequate test data and commercial MT system support. FLORES101 and WMT2022 are general domain test sets, while WMT2020-2021 are news domain.

We then split the final combined test set based on whether the sentences originally end with a (*i*) full stop, (*ii*) exclamation mark, (*iii*) question mark, or (*iv*) no punctuation. In order to balance the test set sizes, we randomly choose 1000 sentences ending with a full stop. All test set sizes are given in Appendix A.1.

2.2 Perturbation Tests

Insertion. For the insertion perturbation, we start with the test set split that originally occurs with no ending punctuation, and then insert at the **end** of each sentence: a (*i*) full stop, (*ii*) exclamation mark, (*iii*) question mark, or a (*iv*) random character for comparison. The insertion of a single punctuation mark at the end of a sentence is an extremely minimal perturbation that does not change any content. We contrast this with the insertion of a random character at the end which changes the final word.

Deletion. For the deletion perturbation, we start with the test set splits that originally occur with a punctuation at the end of the sentences (full stop, exclamation or question mark) and delete them. We also contrast this with deleting the final character of the sentence, for which we use the split with no ending punctuation.

3 Evaluation Metrics

Before we evaluate the machine translation systems on our punctuation perturbation test sets, we first evaluate the evaluation metrics themselves to see if they are robust to these variations. This **meta-evaluation** is crucial; if the metrics are not reliable, we cannot be sure if changes in the scores are due to changes in translation content. We include the string-based metric **BLEU** (Papineni et al., 2002) for convention, and based on the recommendations from Kocmi et al. (2021), we use **chrF** (Popović, 2015), which is another string-based metric, and **COMET** (Rei et al., 2020), which is a model-based metric shown to have high correlations with human judgements, and also include **BLEURT-20** (Sellam et al., 2020) and **BERTScore** (Zhang* et al., 2020). Metric versions can be found in Appendix A.2.

3.1 Meta-Evaluation

Typical robustness tests for machine translation evaluate the translations of both the original and the perturbed source texts against the original reference text (Belinkov and Bisk, 2018; Michel et al., 2019; Bergmanis et al., 2020). The implicit assumption here is that given that the semantics are preserved, the ideal MT system should produce the same or a similar translation for both, and that the automatic metrics used to perform evaluation against the original reference translation will accurately measure the translation quality.

However, adding or deleting punctuation from the source input can lead to a predictable corresponding presence or absence of punctuation in the machine translation - which the reference translation lacks, since it may match the punctuation in the original source. In such circumstances, it is unclear if this significantly influences the evaluation quality perceived by the metrics.

Setup. In order to investigate whether automatic metrics are robust to the "translation of perturbed source but original reference" discrepancy, we conduct experiments comparing the scores produced by the metrics using the original and perturbed source texts as the "reference" and "translation" texts. More concretely, given the original source text $\mathbf{X}$, its perturbed version $\mathbf{X}'$, and a scoring metric $f(\mathbf{Y}, \mathbf{R})$ where $\mathbf{Y}$ is the translation and $\mathbf{R}$ is

[2] Approximation based on `https://www.statmt.org/wmt22/translation-task.html`.

Lang.	Insertion Test	BLEU	chrF	COMET	BLEURT	BERTScore
De	Original Source	100.0	100.0	124.7	96.6	100.0
	+ Full stop	-9.7	-0.3	-7.5	-4.2	-5.2
	+ Exclamation	-9.7	-0.3	-9.1	-6.4	-5.7
	+ Question	-9.7	-0.3	-24.9	-7.1	-6.0
	+ Random	-10.6	-0.3	-32.2	-9.1	-3.0
Ja	Original Source	100.0	100.0	129.6	97.6	100.0
	+ Full stop	-6.7	-0.7	-2.8	-6.3	-4.3
	+ Exclamation	-6.7	-0.7	-3.7	-7.8	-6.0
	+ Question	-6.7	-0.7	-9.9	-9.0	-5.5
	+ Random	-6.9	-0.7	-17.4	-11.7	-3.1
Uk	Original Source	100.0	100.0	132.6	99.0	100.0
	+ Full stop	-9.3	-0.4	-0.8	-2.6	-5.0
	+ Exclamation	-9.3	-0.4	-0.7	-4.5	-5.0
	+ Question	-9.3	-0.7	-2.5	-5.6	-5.5
	+ Random	-10.2	-0.4	-7.6	-8.5	-2.7
En	Original Source	100.0	100.0	125.3	97.3	100.0
	+ Full stop	-9.1	-0.4	-6.8	-4.1	-0.9
	+ Exclamation	-9.1	-0.4	-7.6	-6.3	-1.2
	+ Question	-9.1	-0.4	-18.2	-6.9	-1.7
	+ Random	-9.9	-0.4	-25.3	-7.8	-2.5

Table 2: Results comparing the punctuation insertion perturbed source texts against the original source texts using various metrics and showing the difference in scores. All comparisons use the original source text as the "reference" translation. COMET, BERTScore and BLEURT are reported x100 to match all score scales. Note that COMET and BLEURT do not always produce a score of 100.0 for perfect matches as they were not trained to produce scores within a specific range.

Lang.	Deletion Test	BLEU	chrF	COMET	BLEURT	BERTScore
De	Original Source	100.0	100.0	114.3	95.9	100.0
	- Fullstop	-3.7	-0.6	-7.0	-6.6	-3.2
	Original Source	100.0	100.0	123.5	97.1	100.0
	- Exclamation	-7.1	-1.2	-8.0	-7.0	-5.6
	Original Source	100.0	100.0	125.8	97.2	100.0
	- Question	-8.1	-1.5	-15.5	-7.4	-7.4
	Original Source	100.0	100.0	124.7	96.6	100.0
	- Random	-10.6	-1.3	-34.3	-12.0	-3.6
Ja	Original Source	100.0	100.0	127.6	98.0	100.0
	- Fullstop	-3.5	-1.6	-3.2	-5.7	-2.7
	Original Source	100.0	100.0	131.3	96.7	100.0
	- Exclamation	-7.5	-3.6	-2.3	-6.3	-6.3
	Original Source	100.0	100.0	131.9	97.6	100.0
	- Question	-6.7	-3.3	-2.7	-6.0	-6.1
	Original Source	100.0	100.0	129.6	97.6	100.0
	- Random	-7.3	-3.0	-23.9	-14.3	-3.3
Uk	Original Source	100.0	100.0	131.8	99.2	100.0
	- Fullstop	-5.4	-0.9	-1.0	-5.0	-3.2
	Original Source	100.0	100.0	132.6	99.9	100.0
	- Exclamation	-8.2	-1.5	-0.6	-6.0	-5.7
	Original Source	100.0	100.0	132.7	99.1	100.0
	- Question	-8.7	-1.7	-1.5	-6.5	-5.6
	Original Source	100.0	100.0	132.6	99.0	100.0
	- Random	-10.2	-1.6	-11.6	-11.4	-2.8
En	Original Source	100.0	100.0	116.7	97.9	100.0
	- Fullstop	-3.8	-0.7	-6.1	-6.2	-0.7
	Original Source	100.0	100.0	124.1	97.9	100.0
	- Exclamation	-6.7	-1.5	-7.1	-8.2	-1.4
	Original Source	100.0	100.0	126.2	97.7	100.0
	- Question	-7.9	-1.7	-12.2	-8.7	-1.7
	Original Source	100.0	100.0	125.3	97.3	100.0
	- Random	-9.9	-1.5	-34.2	-10.8	-2.9

Table 3: Results comparing the punctuation deletion perturbed source texts against the original source texts using various metrics and showing the difference in scores. All comparisons use the original source text as the "reference" translation. COMET, BERTScore and BLEURT are reported x100 to match all score scales. Note that COMET and BLEURT do not always produce a score of 100.0 for perfect matches as they were not trained to produce scores within a specific range.

the reference, we compute the score $f(\mathbf{X}, \mathbf{X})$ (perfect match) and $f(\mathbf{X}', \mathbf{X})$ (single punctuation mismatch).[3] We conduct this comparison for both the insertion and deletion tests, across all 4 languages (De, Ja, Uk and En).

The goal here is to measure, given all else is equal, whether punctuation insertion/deletion at the end of the sentence significantly affects the scores produced by the automatic metrics, and how this compares against a more typical perturbation of inserting or deleting a random final character. Ideally, the metrics should not produce different scores that are statistically significant given trivially perturbed inputs. We can then rely on scores produced by the metrics to perform robustness evaluations.

Insertion Results. The meta-evaluation results for the punctuation insertion tests are shown in Table 2. We see that the metrics produce significantly different scores even though the only difference is a single additional punctuation mark at the end of the sentence. The difference is particularly stark for BLEU, COMET and BLEURT while less pronounced for chrF and BERTScore, and is equally poor across languages. More interestingly, we see that while string-based matching metrics such as BLEU and chrF treat all punctuation equally, model-based metrics assign drastically lower scores for exclamation and question marks. In the case of BERTScore, punctuation insertion results in lower scores than random character insertion for all languages except English.

Deletion Results. The meta-evaluation results for the punctuation deletion tests are shown in Table 3. A similar trend is seen here, where the lack of a single punctuation at the end of the sentence causes a significant drop in scores across all metrics. We also see the same trend where missing exclamation or question marks result in more significant drops in scores. Furthermore, punctuation deletion more often results in lower scores than deleting a random final character compared to punctuation insertion. Surprisingly, results for Uk are relatively more sta-

[3] For COMET, which is of the form $f(\mathbf{Y}, \mathbf{S}, \mathbf{R})$ where S is the source text, we compute $f(\mathbf{X}, \mathbf{X}, \mathbf{X})$ and $f(\mathbf{X}', \mathbf{X}, \mathbf{X})$.

		chrF	COMET
Original	1) Deauthorize your e-book reader	79.3	94.9
Perturbed	1) Deauthorize your e-book reader .	78.6	93.0
	Score Δ	-0.7	-1.9
Original	Elon Musk lets Tesla shares rise	60.4	98.5
Perturbed	Elon Musk makes Tesla shares rise	59.4	97.7
	Score Δ	-1.0	-0.8

Table 4: Example of a case where minor punctuation difference earns similar or lower scores than more severe translation changes. Original translations are highlighted in yellow and changes in translation are highlighted in red.

ble than for En, particularly for COMET.

Note that all score differences here will register as statistically significant: the original source will always "win" against the perturbed source in all comparisons performed by tests such as paired bootstrap resampling or randomization.

Some issues with BLEU have been highlighted previously (Reiter, 2018; Kocmi et al., 2021); COMET, BLEURT and BERTScore presumably suffer from robustness issues as neural models. ChrF scores display smaller variations that are consistent across punctuation and languages, and therefore seem more reliable for robustness evaluations, corroborating the findings from Michel et al. (2019). Overall, we expand the metric sensitivity issues highlighted in Karpinska et al. (2022) for English in finer-detail for punctuation, and further confirm them for German, Japanese and Ukrainian.

Comparison with severe translation errors. We performed a manual segment-wise analysis of a subset of machine translation outputs. We find that in several cases, particularly for shorter sentences, translations with punctuation differences are penalized similar to translations with severe errors. See Table 4 for an example.

Broader Implications. More broadly, these results indicate that (*i*) statistically significant differences can be obtained merely by changing a single punctuation, (*ii*) models that fail to match the reference punctuation may be penalized more than they should be, and (*iii*) models that mistranslate a single word but correctly match the punctuation may be getting more credit than they should. For example, we found that up to 5% of the sentences in the WMT2022 Uk-En test set and up to 10% of the sentences in the WMT2022 Ja-En test set had mismatches between the ending punctuation in the source and reference. This could mean that model performance on these instances may be undervalued if the model reproduces the source punctuation. Conversely, we also found many instances of models producing acceptable punctuation that was not present in the original source (*e.g.*, $\approx 13\%$ of Microsoft's Uk-En output for full stop deletion perturbation test set had full stops), which may also get unfairly penalized.

More importantly, it may be worthwhile to reexamine how machine translation models are evaluated in robustness tests and after adversarial training, since resultant differences in scores may not be a reflection of actual translation quality.

4 Machine Translation Experiments

We now test the publicly available commercial machine translation systems of **Google**, **DeepL** and **Microsoft** through their paid APIs on our test sets.[4] Some of these commercial systems have previously been claimed to have reached human parity (Wu et al., 2016; Hassan et al., 2018). Commercial systems are generally expected to deal with nonstandard inputs as they are targeted for real-world use cases. We therefore expect that these systems have already been trained to be somewhat robust to various kinds of input noise.

For the insertion tests, we compare the translation of the original source text *without* punctuation against the translation of the perturbed source *with* sentence-final punctuation. For deletion tests, we compare the translation of the original source text *with* punctuation against the translation of the perturbed source *without* sentence-final punctuation.

4.1 Evaluation

Results from our meta-evaluation in §3.1 mean that we cannot get reliable results from evaluation metrics if we directly use the perturbed source translations and original references for evaluation; it will be hard to identify if changes in scores originate from translation differences or merely punctuation changes. One solution is to add the same punctuation perturbation to the reference that we add to the source. We find that this increases the overall scores since there is now an additional character that matches the reference in each sentence, rendering the score incomparable to the original translation.

Another solution is to reset the punctuation changes in the translations. We therefore remove corresponding sentence-final punctuation produced

[4] All translations are from December 2022.

in the translations for the source inputs perturbed through insertion that are *not also produced for the original source inputs*, and vice versa for deletion, thereby making the two translations comparable. Henceforth we use chrF scores due to its relative robustness and include COMET scores as it has been shown to have high correlations with human judgements (Kocmi et al., 2021; Freitag et al., 2022).

Inconsistency. Apart from measuring whether perturbations cause degradation in translations compared to a reference, another important criterion is the consistency. That is, given the original and the perturbed sources as input, we measure how different the translations produced for each are. Since here we want to also account for surface-level changes, we choose the string-based matching metric chrF based on results in §3.1 and findings from Michel et al. (2019). Given a source $\mathbf{X}$ and its translation $\mathbf{Y}$, and the perturbed source $\mathbf{X}'$ and its translation $\mathbf{Y}'$, we measure **consistency** at the sentence-level as the score chrF$(\mathbf{Y}', \mathbf{Y})$, where $\mathbf{Y}$ acts as the "reference". We designate a score < 75 to be a significant deviation in translation, and measure **percentage of inconsistency** by counting the number of $\mathbf{Y}'$ which have chrF< 75.

4.2 Results

The results for the punctuation insertion perturbation tests are given in Table 5. We see that in general, the insertion of sentence-final punctuation results in a statistically significant drop in scores, but also some significant improvements. The results for the punctuation deletion perturbation tests are given in Table 7. Overall, deletion causes more drops in performance than insertion, and far fewer improvements in scores.

Effect of Language. Unsurprisingly, based on inconsistency, we see that the models are far more robust to insertion perturbations for the high resource language pair De-En, with generally $< 10\%$ inconsistency. More interestingly, we see that while Ja-En and Uk-En are both medium resource, the models are far more robust for Uk-En at $0 - 23\%$ inconsistency, as compared to Ja-En which has between $18 - 35\%$ inconsistency across models.

We see the same inconsistency trends for deletion as for insertion: models are more robust to perturbations in De ($0 - 23\%$) and Uk ($0 - 25\%$) source texts than Ja ($10 - 37\%$). Overall, deletion leads to a higher range of inconsistency than insertion.

Effect of Punctuation. We see that the models are more likely to be robust to full stop insertion than exclamation and question marks: statistically significant differences in performance occur more often for the latter. In fact, DeepL and Microsoft models seem to benefit from having full stops and exclamation marks added, with results improving for Ja-En and Uk-En. In the case of question marks, we see that it causes a universal drop in scores across models and languages. For Uk-En, question mark insertion almost always causes more significant drops in scores than inserting a random character.

Unlike insertion, full stop deletion causes significant drops in scores, particularly for the DeepL and Microsoft models for Ja-En and Uk-En. Interestingly, question mark deletion does not cause a significant score drop in Ja-En for all models. This is possibly because the question mark is mostly optional in Ja, which uses the particle 'か' as a question marker.

Pre-processing. We see that both insertion and deletion can cause degradation in performance. This means that while pre-processing of the inputs to ensure consistent punctuation may lead to more consistent translations, it is unlikely to result in better quality translations.

5 Analysis and Discussion

Some examples of translation changes caused by the perturbations are given in Table 6. Both insertion and deletion cause a wide range of translation changes, with a few severe errors where the meaning is completely changed, such as by hallucinating or omitting negation. Others include changes in number, tense, pronouns, named entities, etc.

Reordering. Often, inserting or deleting punctuation leads to a reordering of the words in the sentence. In many cases the reordering leads to mostly similar but slightly off translations (Example 4), with some cases causing significant differences in meaning (Example 8).

While we might expect punctuation perturbation to ideally cause no other changes in translation apart from the difference in punctuation itself, there could be cases of valid translation changes caused by the perturbation. For example, while *"1) Heben Sie die Autorisierung des Lesegeräts auf"* is originally translated as *"1) Deauthorize the reader"*, adding a question mark does not produce *"1) Deauthorize the reader?"* but instead *"1) Are you deauthorizing*

		Google			DeepL			Microsoft		
Lg.	Insertion	chrF	COMET	%Inc.	chrF	COMET	%Inc.	chrF	COMET	%Inc.
De-En	Original Source	62.6	66.1		63.5	67.8		63.2	67.1	
	+ Full stop	0.0	+0.1	0.0	-0.2	-0.4	8.3	**-0.5**	-0.7	7.2
	+ Exclamation	-0.2	0.0	11.6	-0.3	**-1.3**	7.9	**-0.7**	**-0.9**	7.2
	+ Question	**-0.5**	**-2.1**	13.5	**-0.6**	**-2.0**	8.3	**-0.7**	**-1.4**	8.7
	+ Random	-0.3	**-10.2**	10.4	**-0.9**	**-21.3**	9.4	**-0.7**	**-9.5**	6.5
Ja-En	Original	53.2	40.5		51.5	37.7		52.8	36.3	
	+Full stop	0.0	0.0	24.3	+0.2	+0.2	30.7	+0.2	**+1.9**	20.6
	+Exclamation	-0.2	-0.6	24.2	**-1.0**	**-1.9**	33.4	+0.1	+1.2	19.2
	+Question	**-0.4**	**-6.0**	28.0	-0.5	**-5.2**	35.8	**-0.4**	**-0.3**	25.4
	+Random	**-0.4**	**-7.0**	22.3	**-0.6**	**-12.1**	27.1	**-0.3**	**-7.1**	18.8
Uk-En	Original	64.7	58.3		63.1	56.1		61.4	42.4	
	+Full stop	0.0	0.0	0.8	**+0.9**	+1.1	12.6	0.0	**+1.9**	10.3
	+Exclamation	-0.1	**-0.9**	10.5	**+0.9**	+0.7	15.1	-0.3	**+1.4**	10.5
	+Question	**-2.1**	**-9.9**	23.9	**-0.8**	**-5.6**	20.8	**-1.8**	**-7.6**	22.1
	+Random	**-0.9**	**-5.0**	10.7	-0.4	**-6.9**	13.4	-0.2	**-7.6**	9.1

Table 5: Results for the punctuation insertion task for De/Ja/Uk-En for Google, DeepL and Microsoft MT systems, showing the differences in scores of the translations for perturbed source texts. **Lg.** indicates language pair, while **%Inc**onsistent is the percentage sentences which have chrF<75 with respect to the original translation. Results in **bold** are statistically significant (paired bootstrap resampling, $p < 0.05$).

					Δ		Con.
	#	Text	Original ($\mathbf{X}, \mathbf{Y}$)	Perturbed ($\mathbf{X}', \mathbf{Y}'$)	chrF	COMET	chrF
Insertion	1	Source	なにかアドバイス下さい	なにかアドバイス下さい 。			
		Google	give me some advice	Please give me some advice	+17.8	+8.0	91.7
	2	Source	Якщо ще колись не захочете мені писати, то я чекатиму	Якщо ще колись не захочете мені писати, то я чекатиму .			
		DeepL	If you ever want to write to me again, I will be waiting	If ever you do not want to write to me, I will be waiting	-2.1	-23.0	67.7
	3	Source	Elon Musk lässt Tesla-Aktien steigen	Elon Musk lässt Tesla-Aktien steigen !			
		Microsoft	Elon Musk lets Tesla shares rise	Elon Musk makes Tesla shares rise	-1.0	-0.8	76.9
	4	Source	アンドロイドのハドウェアをきれいにするコツ	アンドロイドのハドウェアをきれいにするコツ ！			
		DeepL	Tips for cleaning android hardware	Androids, tips on how to clean up your hardware	-19.2	-98.4	43.1
	5	Source	かけが良すぎるガデニングギミックにご用心	かけが良すぎるガデニングギミックにご用心 ？			
		Google	Beware of gardening gimmicks that look too good	Worried about gardening gimmicks that look too good	-11.8	-19.7	78.6
	6	Source	Der saubere Lake Tahoe vom Keimwandel verunreinigt	Der saubere Lake Tahoe vom Keimwandel verunreinigt ?			
		DeepL	The clean Lake Tahoe polluted by the germ change	Clean Lake Tahoe contaminated by gerrymandering	-14.5	-72.7	41.0
Deletion	7	Source	Проблема з температурою водонагрівача та ванною .	Проблема з температурою водонагрівача та ванною			
		Microsoft	The problem is with the temperature of the water heater and bath.	Problem with water heater temperature and bath.	+11.8	+38.6	56.0
	8	Source	Ja... wir haben hier alle Schusswaffen .	Ja... wir haben hier alle Schusswaffen			
		Microsoft	Yes... we all have firearms here.	Yes... we have all the firearms here.	-9.7	-7.6	75.5
	9	Source	«Справа дійсно зрушилася з місця !	«Справа дійсно зрушилася з місця			
		Google	“The matter really got out of hand !	“The matter has really moved from place to place !	-4.0	-19.4	43.4
	10	Source	Boron zum Grusse !	Boron zum Grusse			
		Google	Greetings Boron!	Greetings from Boron!	-2.6	-30.1	73.1
	11	Source	Тільки я буду трошки пізніше - десь о 8. Можна ?	Тільки я буду трошки пізніше - десь о 8. Можна			
		DeepL	Only I will be a little later - around 8 . Can I ?	Only I will be a little later - around 8 o’clock . You can?	-4.3	+0.5	80.6
	12	Source	LINEは何日に何回送るのが良いですか ？	LINEは何日に何回送るのが良いですか			
		Microsoft	What is the best time to send LINE messages per day?	How many times a day should I send LINE?	-8.5	+14.1	25.9

Table 6: Examples of changes in translation caused by perturbations. Punctuation perturbations at the end of the sentence are highlighted in blue , original translations are highlighted in yellow and the changes in the translations are highlighted in red . Given a translation $\mathbf{Y}$ of the original source $\mathbf{X}$, a translation $\mathbf{Y}'$ of the perturbed source $\mathbf{X}'$ and a reference $\mathbf{R}$, the Δ scores show the differences in chrF and COMET scores obtained as $f(Y', R) - f(Y, R)$, while **Con. chrF** measures the consistency through chrF scores between the two translations, obtained as $chrF(Y', Y)$. Best viewed in color.

the reader?”. This word reordering for an interrogative sentence, typical particularly for English, can be considered a valid change even though the chrF (65.5 → 52.7) and COMET (28.2 → -25.8) scores drop. There are also cases when the resultant reordering actually improves the scores of the translation despite being wrong *e.g.,* adding a question mark to *“Und ich muss nochmal Versandkosten*

		Google			DeepL			Microsoft		
Lg.	Deletion	chrF	COMET	%Inc.	chrF	COMET	%Inc.	chrF	COMET	%Inc.
De-En	Original Source	66.1	68.2		67.0	68.6		66.3	67.1	
	- Full stop	0.0	0.0	0.1	+0.1	0.0	4.7	0.0	-0.3	1.9
	Original Source	61.2	66.8		61.0	66.3		60.8	65.0	
	- Exclamation	-0.3	-0.2	7.8	-0.4	+1.6	14.7	-0.5	-1.9	4.9
	Original Source	61.3	61.8		61.3	56.0		60.8	59.9	
	- Question	**-0.9**	**-3.7**	18.4	**-0.9**	**-4.5**	23.0	**-1.2**	**-5.2**	17.3
	Original Source	62.6	66.1		63.5	67.8		63.2	67.1	
	- Random	**-0.6**	**-8.3**	10.4	**-1.2**	**-3.2**	9.9	**-1.5**	**-15.4**	11.2
Ja-En	Original Source	55.6	52.0		56.5	56.0		55.5	53.9	
	- Full stop	-0.1	-0.3	13.1	**-0.3**	**-1.5**	13.2	**-0.2**	**-1.8**	10.0
	Original Source	46.8	41.6		48.1	43.3		46.7	39.9	
	- Exclamation	-0.8	**-3.2**	26.1	-0.7	-2.9	31.4	**-1.3**	**-2.7**	26.6
	Original Source	48.9	41.3		51.8	48.3		47.6	39.5	
	- Question	-0.3	-2.9	34.9	**-1.8**	**-6.4**	29.6	-0.2	-2.2	19.4
	Original Source	53.2	40.5		51.5	37.7		52.8	36.3	
	- Random	**-1.7**	**-13.2**	33.4	**-2.7**	**-14.0**	37.5	**-2.2**	**-18.5**	33.5
Uk-En	Original Source	65.2	64.6		65.2	65.1		62.6	59.3	
	- Full stop	+0.1	-0.2	0.6	**-0.4**	**-1.0**	5.5	**-0.2**	**-0.8**	4.2
	Original Source	64.6	69.7		63.5	66.0		59.4	58.5	
	- Exclamation	**-0.9**	-0.9	5.8	-0.3	+0.5	8.7	**-0.9**	-3.9	10.1
	Original Source	61.7	57.7		61.6	60.9		59.0	50.0	
	- Question	**-1.7**	**-5.6**	21.6	**-1.3**	**-6.2**	24.7	**-2.4**	-6.7	25.4
	Original Source	64.7	58.3		63.1	56.1		61.4	42.4	
	Random	**-1.0**	**-7.5**	12.2	**-1.3**	**-7.9**	15.5	**-2.9**	**-16.2**	17.5

Table 7: Results for the punctuation deletion task for De/Ja/Uk-En for Google, DeepL and Microsoft MT systems, showing the differences in scores of the translations for perturbed source texts. **Lg.** indicates language pair, while **%Inc**onsistent is the percentage of sentences which have chrF<75 with respect to the original translation. Results in **bold** are statistically significant (paired bootstrap resampling, $p < 0.05$).

zahlen" changes the translation from *"And I have to pay shipping again"* to *"And do I have to pay shipping costs again?"* (instead of *"And I have to pay shipping again?"*) and improves both chrF (17.5 → 23.4) and COMET (26.3 → 45.6) scores, presumably due to the presence of the word *"costs"* that now matches the reference (*"And I still need to pay the delivery costs"*). Similarly for question mark deletion, removing the question mark from "Заняття в понеділок і середу відрізняються?" changes the translation from *"Are Monday and Wednesday classes different?"* to *"Monday and Wednesday classes are different"*, dropping the chrF (76.2 → 74.9) and COMET (91.0 → 83.7) scores. Expecting translations of both original and perturbed source texts to match is a standard evaluation setting for robustness tests, even for more severe perturbations resulting in drastic changes and out-of-vocabulary inputs (see Table 1). Given these results, we reiterate our call from §3 to re-examine this evaluation setup for settings similar to this work.

However, there are several cases where the interrogative nature of the source is not dependent on the question mark and the model correctly produces a translation that is also interrogative but different. For example, deleting the question mark from "ありますか? " changes the translation from *"Is there?"* to *"do you have"*. Example 12 shows another case where the model correctly recognizes the perturbed source as a question, but produces a significantly different translation. Example 5 and Example 6 are also cases of translation differences that are more severe than reordering.

Sentence Style Association. Although we see some critical translation changes due to perturbing full stops (Example 2), a majority of the translations underwent a change in sentence style. In particular, we found that inserting a full stop resulted in models producing longer, complete sentences, while deleting the full stop resulted in shorter, headline-style sentences. This was observed across systems (Example 1 and 7), which indicates that this stylistic change presumably comes from what is commonly seen in training data: the models have seemingly learnt to associate a lack of full stop with article headlines from news domain data. In the case of Example 7, the changed translation better matches the reference ("*Water heater temp and bath issue.*"), leading to improvement in scores in both chrF ($46.8 \rightarrow 58.6$) and COMET ($36.7 \rightarrow 75.3$).

Robustness. Previous works have correlated consistency with robustness (Niu et al., 2020; Wallace et al., 2020), the implication being that less consistent outputs are lower in translation quality. We find that this is not necessarily the case for our perturbation setting. For instance, Example 1 shows a translation that has high consistency (91.7 chrF compared to original translation), while Example 7 has low consistency (56.0 chrF). However, in both cases the translations of the perturbed sources score significantly higher than the original translations. Similarly, Example 12 has a very low consistency score (25.9) but the chrF reduces (-8.5) while the COMET increases ($+14.1$). COMET is more reflective of the translation quality here: given the reference (' '*How many LINE messages are okay to send in a day?"*), the translation of the perturbed source is closer to the actual translation. Conversely, instances with relatively high consistency (Example 3, 5, 8, 10) all drop in scores and have significant translation issues.

Other Changes. Some other changes in the translations include changes in number, tense, pronouns, named entities, capitalization, and so on. Some of the less severe errors such as changes in capitalization or extra demonstratives also incur heavy drops in chrF and COMET scores. Some more examples of the translations produced for perturbed inputs can be found in Appendix A.3.

6 Conclusions

In this work, we unite the robustness evaluation of both machine translation systems and their evaluation metrics, and discuss ways in which both fail to be adequately robust to trivial punctuation change. This shows that models and metrics are in fact far more sensitive and a lot less reliable in real-world use cases than is commonly expected. We show that both metrics and machine translation systems treat each punctuation differently, with machine translation systems showing associations between punctuation and sentence styles. We also highlight the implications of these sensitivities for robustness research and evaluation for machine translation.

Although it may not necessarily be a hard task to train systems that are robust to punctuation, our goal is to highlight one of the issues that has possibly been overlooked due to its triviality. We hope that future research in robustness, evaluation metrics and machine translation accounts for these sensitivities while performing evaluation and model training.

Limitations

Test Set Size. One of the main limitations of our work is relatively smaller test set sizes. This stems from the way our perturbation experiments are set up - we can only use existing test sentences which already end with specific punctuation in order to measure the effect of deleting them, or start with sentences which do not have sentence final punctuation in order to measure the effect of inserting them. In general, a majority of the official test sets have sentences ending in full stops; this results in having a smaller test set to work with. This is also the same issue that presumably gives rise to sensitivity issues in the trained models.

However, given that our focus has been on each particular punctuation, instead of merging them all together, we find that our test sets are larger than the ones used in previous work for each punctuation. Combined with the fact that we ensure to perform significance testing and manual analysis, we believe our results are reliable. Appendix A.1 includes details and a discussion.

Target Language. Although we test models across several source languages, the target language is always English. This makes our analysis of induced errors limited to phenomena that occur in English, for example, changes in number, reordering of words for question marks, or changes in capitalization, etc. Languages without capitalization or number marking but with morphological richness and other phenomena are likely to have different errors. For example, inserting a full stop changes the translation to include 'Please' and makes the sentence more polite in Example 1 in Table 6. For languages like Japanese, which have complex systems of marking varying levels of honorifics, punctuation perturbations may result in more interesting changes to the translations.

A vast majority of previous work has performed perturbations on languages using the Latin alphabet, so we consider our work a step forward, considering that we also evaluate metrics on Japanese and Ukrainian texts. However, it is also important to evaluate sensitivity when both directions are non-English, for example, Ukraininan to Japanese translation. A lack of adequate parallel data in such directions usually precludes such experiments. We hope to undertake this in future work.

Acknowledgments

We would like to thank the reviewers for their reviews and suggestions. We would also like to thank our colleagues Alberto Poncelas and Maksim Tkachenko for their valuable inputs.

References

Duarte Alves, Ricardo Rei, Ana C Farinha, José G. C. de Souza, and André F. T. Martins. 2022. Robust MT evaluation with sentence-level multilingual augmentation. In *Proceedings of the Seventh Conference on Machine Translation*, pages 469–478, Abu Dhabi. Association for Computational Linguistics.

Loic Barrault, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussa, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Tom Kocmi, Andre Martins, Makoto Morishita, and Christof Monz, editors. 2021. *Proceedings of the Sixth Conference on Machine Translation*. Association for Computational Linguistics, Online.

Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Yvette Graham, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, and Matteo Negri, editors. 2020. *Proceedings of the Fifth Conference on Machine Translation*. Association for Computational Linguistics, Online.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. *ArXiv*, abs/1711.02173.

Toms Bergmanis, Arturs Stafanovivcs, and Marcis Pinnis. 2020. Robust neural machine translation: Modeling orthographic and interpunctual variation. In *Baltic HLT*.

Xiaoyu Chen, Daimeng Wei, Hengchao Shang, Zongyao Li, Zhanglin Wu, Zhengzhe Yu, Ting Zhu, Mengli Zhu, Ning Xie, Lizhi Lei, Shimin Tao, Hao Yang, and Ying Qin. 2022. Exploring robustness of machine translation metrics: A study of twenty-two automatic metrics in the WMT22 metric task. In *Proceedings of the Seventh Conference on Machine Translation*, pages 530–540, Abu Dhabi. Association for Computational Linguistics.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. HotFlip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 31–36, Melbourne, Australia. Association for Computational Linguistics.

Adam Ek, Jean-Philippe Bernardy, and Stergios Chatzikyriakidis. 2020. How does punctuation affect neural models in natural language inference. In *Proceedings of the Probability and Meaning Conference (PaM 2020)*, pages 109–116, Gothenburg. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and F. T. André Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU - neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation*, pages 46–68, Abu Dhabi. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjan Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2021. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William D. Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving human parity on automatic Chinese to English news translation. *ArXiv*, abs/1803.05567.

Georg Heigold, Stalin Varanasi, Günter Neumann, and Josef van Genabith. 2018. How robust are character-based word embeddings in tagging and MT against wrod scramlbing or randdm nouse? In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 68–80, Boston, MA. Association for Machine Translation in the Americas.

Marzena Karpinska, Nishant Raj, Katherine Thai, Yixiao Song, Ankita Gupta, and Mohit Iyyer. 2022.

Demetr: Diagnosing evaluation metrics for translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, Maja Popovic, and Mariya Shmatova. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation*, pages 1–45, Abu Dhabi. Association for Computational Linguistics.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Xian Li, Paul Michel, Antonios Anastasopoulos, Yonatan Belinkov, Nadir Durrani, Orhan Firat, Philipp Koehn, Graham Neubig, Juan Pino, and Hassan Sajjad. 2019. Findings of the first shared task on machine translation robustness. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 91–102, Florence, Italy. Association for Computational Linguistics.

Paul Michel, Xian Li, Graham Neubig, and Juan Miguel Pino. 2019. On evaluation of adversarial perturbations for sequence-to-sequence models. In *North American Chapter of the Association for Computational Linguistics*.

Xing Niu, Prashant Mathur, Georgiana Dinu, and Yaser Al-Onaizan. 2020. Evaluating robustness to input perturbations for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8538–8544, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Martin Popel, Markéta Tomková, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondrej Bojar, and Z. Žabokrtský. 2020. Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature Communications*, 11.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

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

Ehud Reiter. 2018. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Anders Søgaard, Miryam de Lhoneux, and Isabelle Augenstein. 2018. Nightmare at test time: How punctuation prevents parsers from generalizing. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 25–29, Brussels, Belgium. Association for Computational Linguistics.

Jiao Sun, Thibault Sellam, Elizabeth Clark, Tu Vu, Timothy Dozat, Dan Garrette, Aditya Siddhant, Jacob Eisenstein, and Sebastian Gehrmann. 2022. Dialect-robust evaluation of generated text. *ArXiv*, abs/2211.00922.

Samson Tan, Shafiq R. Joty, Min-Yen Kan, and Richard Socher. 2020. It's morphin' time! combating linguistic discrimination with inflectional perturbations. *ArXiv*, abs/2005.04364.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *ArXiv*, abs/1706.03762.

Eric Wallace, Mitchell Stern, and Dawn Song. 2020. Imitation attacks and defenses for black-box machine translation systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5531–5546, Online. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Z. Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason R. Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Gregory S. Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *ArXiv*, abs/1609.08144.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

Xinze Zhang, Junzhe Zhang, Zhenhua Chen, and Kun He. 2021. Crafting adversarial examples for neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1967–1977, Online. Association for Computational Linguistics.

A Appendix

A.1 Testset Sizes

Test Split	De-En	Ja-En	Uk-En
No Final Punctuation	748	956	515
Final Full Stop	1000	1000	1000
Final Exclamation Mark	102	207	69
Final Question Mark	283	284	287

Table 8: Testset sizes

Test set sizes for our perturbation tests are given in Table 8. Note that all punctuation insertion tests use the **No Final Punctuation** split, while the deletion tests use the respective ending punctuation splits. Random insertion and deletion both use the **No Final Punctuation** split.

The test set sizes reflect a general imbalance in sentence-final punctuation in parallel corpora that may be causing the sensitivity in the models. In order to be able to insert or delete punctuation, we are limited to sentences which originally have no punctuation or the specific punctuation we intend to delete. This is a requirement unique to our extremely minimal setup, since more indiscriminate punctuation perturbations can be possibly carried out on a larger scale.

For comparison, the FLORES101 dataset has 1012 sentences, and the WMT2020-2022 datasets range from 785 to 2037 sentences. Some challenge sets for metrics in the WMT Metrics Tasks (Freitag et al., 2022) included 50 sentences per phenomenon for 3 language pairs (Alves et al., 2022) and 721 sentences covering 5 error types for Zh-En (Chen et al., 2022).

A.2 Metric Versions

Metric signatures and versions used for evaluation are given in Table 9.

A.3 More Translation Examples

Table 10 shows some more examples of translation changes in response to perturbations. We see more instances of changes in sentence style, fluency, hallucination and others.

Metric	Version
BLEU	`nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.3.1`
BLEU [Ja]	`nrefs:1|case:mixed|eff:no|tok:ja-mecab-0.996-IPA|smooth:exp|version:2.3.1`
chrF	`nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.3.1`
COMET	1.1.3 `wmt20-comet-da`
BERTScore [En]	`roberta-large_L17_no-idf_version=0.3.12(hug_trans=4.25.1)`
BERTScore [Other]	`bert-base-multilingual-cased_L9_no-idf_version=0.3.12(hug_trans=4.25.1)`
BLEURT	0.0.2 BLEURT-20

Table 9: Metric versions and signatures. We use the sacreBLEU (Post, 2018) implementations for BLEU and chrF, and the huggingface implementations for BLEURT and BERTScore.

#	Text	Original	Perturbed
13	Source	2019年初りセル催中	2019年初りセル催中！
	DeepL	2019 New Year's Sale is underway!	2019 First Year Sale is on now!
14	Source	У нас військова служба обов'язковою для всіх чоловіків від 16 до 29 років.	У нас військова служба обов'язковою для всіх чоловіків від 16 до 29 років
	DeepL	In Ukraine, military service is compulsory for all men aged 16 to 29.	In our country, military service is compulsory for all men aged 16 to 29.
15	Source	Яблуко від яблуні недалеко, як відомо, пада...	Яблуко від яблуні недалеко, як відомо, пада..
	Google	As you know, the apple does not fall far from the apple tree...	As you know, the apple falls far from the apple tree.
16	Source	BGM	BGM？
	Google	Background music	BGM
17	Source	Vorwürfe gegen Trump verschärfen sich	Vorwürfe gegen Trump verschärfen sich!
	Microsoft	Allegations against Trump intensify	Accusations against Trump are intensifying
18	Source	Я розмовляю укранською, російською та чеською мовами інтенсивно вчуся.	Я розмовляю укранською, російською та чеською мовами інтенсивно вчуся
	DeepL	I speak Ukrainian, Russian and Czech and I am studying intensively.	I speak Ukrainian, Russian and Czech intensively studying.
19	Source	Гришко вже пішов у ясла але не все так просто. . . Дуже сильно плаче	Гришко вже пішов у ясла але не все так просто. . . Дуже сильно плаче.
	DeepL	Grishko has already gone to the nursery, but it's not so easy ... He cries a lot . . .	Hryshko has already gone to the nursery, but not everything is so simple He cries a lot
20	Source	印象に残る日曜日は？	印象に残る日曜日は
	Microsoft	What Sunday left a lasting impression on you?	Memorable Sundays?

Table 10: Examples of changes in translation caused by perturbations. Punctuation perturbations at the end of the sentence are highlighted in blue, original translations are highlighted in yellow and the changes in the translations are highlighted in red.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitations section

□ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Left blank.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 2

☑ B1. Did you cite the creators of artifacts you used?
Section 2

□ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

□ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

□ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 2

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Appendix A.1

C ☒ Did you run computational experiments?

Left blank.

□ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
No response.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

□ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
No response.

□ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
No response.

□ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
No response.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.