A Study of Morphological Robustness of Neural Machine Translation

Sai Muralidhar Jayanthi,* Adithya Pratapa*
Language Technologies Institute
Carnegie Mellon University
`{sjayanth,vpratapa}@cs.cmu.edu`

Abstract

In this work, we analyze the robustness of neural machine translation systems towards grammatical perturbations in the source. In particular, we focus on morphological inflection related perturbations. While this has been recently studied for English→French translation (MORPHEUS) (Tan et al., 2020), it is unclear how this extends to Any→English translation systems. We propose MORPHEUS-MULTILINGUAL that utilizes UniMorph dictionaries to identify morphological perturbations to source that adversely affect the translation models. Along with an analysis of state-of-the-art pretrained MT systems, we train and analyze systems for 11 language pairs using the multilingual TED corpus (Qi et al., 2018). We also compare this to actual errors of non-native speakers using Grammatical Error Correction datasets. Finally, we present a qualitative and quantitative analysis of the robustness of Any→English translation systems. Code for our work is publicly available.[1]

1 Introduction

Multilingual machine translation is commonplace, with high-quality commercial systems available in over 100 languages (Johnson et al., 2017). However, translation from and into low-resource languages remains a challenge (Arivazhagan et al., 2019). Additionally, translation from morphologically-rich languages to English (and vice-versa) presents new challenges due to the wide differences in morphosyntactic phenomenon of the source and target languages. In this work, we study the effect of noisy inputs to neural machine translation (NMT) systems. A concrete practical application for this is the translation of text from non-native speakers. While the brittleness of NMT systems to input noise is well-studied (Belinkov and Bisk, 2018), most prior work has focused on translation from English (English→X) (Anastasopoulos et al., 2019; Alam and Anastasopoulos, 2020).

With over 800 million second-language (L2) speakers for English, it is imperative that the translation models should be robust to any potential errors in the source English text. A recent work (Tan et al., 2020) has shown that English→X translation systems are not robust to inflectional perturbations in the source. Inspired by this work, we aim to quantify the impact of inflectional perturbations for X→English translation systems. We hypothesize that inflectional perturbations to source tokens shouldn't adversely affect the translation quality. However, morphologically-rich languages tend to have free word order as compared to English, and small perturbations in the word inflections can lead to significant changes to the overall meaning of the sentence. This is a challenge to our analysis.

We build upon Tan et al. (2020) to induce inflectional perturbations to source tokens using the `unimorph_inflect` tool (Anastasopoulos and Neubig, 2019) along with UniMorph dictionaries (McCarthy et al., 2020) (§2). We then present a comprehensive evaluation of the robustness of MT systems for languages from different language families (§3). To understand the impact of size of parallel corpora available for training, we experiment on a spectrum of high, medium and low-resource languages. Furthermore, to understand the impact in real settings, we run our adversarial perturbation algorithm on learners' text from Grammatical Error Correction datasets for German and Russian (§3.3).

2 Methodology

To evaluate the robustness of X→English NMT systems, we generate inflectional perturbations to the tokens in source language text. In our methodology,

* Equal contribution.

[1] `https://github.com/murali1996/morpheus_multilingual`

Proceedings of the Seventeenth SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 49–59
August 5, 2021. ©

we aim to identify adversarial examples that lead to maximum degradation in the translation quality. We build upon the recently proposed MORPHEUS toolkit (Tan et al., 2020), that evaluated the robustness of NMT systems translating from English→X. For a given source English text, MORPHEUS works by greedily looking for inflectional perturbations by sequentially iterating through the tokens in input text. For each token, it identifies inflectional edits that lead to maximum drop in BLEU score.

We extend this approach to test X→English translation systems. Since their toolkit[2] is limited to perturbations in English only, in this work we develop our own inflectional methodology that relies on UniMorph (McCarthy et al., 2020).

2.1 Reinflection

UniMorph project[3] provides morphological data for numerous languages under a universal schema. The project supports over 100 languages and provides morphological inflection dictionaries for upto three part-of-speech tags, nouns (N), adjectives (ADJ) and verbs (V). While some UniMorph dictionaries include a large number of types (or paradigms) (German (≈15k), Russian (≈28k)), many dictionaries are relatively small (Turkish (≈3.5k), Estonian (<1k)). This puts a limit on the number of tokens we can perturb via UniMorph dictionary look-up. To overcome this limitation, we use the `unimorph_inflect` toolkit[4] that takes as input the lemma and the morphosyntactic description (MSD) and returns a reinflected word form. This tool was trained using UniMorph dictionaries and generalizes to unseen types. An illustration of our inflectional perturbation methodology is described in Table 1.

2.2 MORPHEUS-MULTILINGUAL

Given an input sentence, our proposed method, MORPHEUS-MULTILINGUAL, identifies adversarial inflectional perturbations to the input tokens that leads to maximum degradation in performance of the machine translation system. We first iterate through the sentence to extract all possible inflectional forms for each of the constituent tokens. Since, we are relying on UniMorph dictionaries, we are limited to perturbing only nouns, adjectives and verbs.[5] Now, to construct a perturbed sentence, we iterate through each token and uniformly sample one inflectional form from the candidate inflections. We repeat this process N (=50) times and compile our pool of perturbed sentences.[6]

To identify the adversarial sentence, we compute the chrF score (Popović, 2017) using the sacrebleu toolkit (Post, 2018) and select the sentence that results in the maximum drop in chrF score (if any). In our preliminary experiments, we found chrF to be more reliable than BLEU (Papineni et al., 2002) for identifying adversarial candidates. While BLEU uses word n-grams to compare the translation output with the reference, chrF uses character n-grams instead; which helps with matching morphological variants of words.

The original MORPHEUS toolkit follows a slightly different algorithm to identify adversaries. Similar to our approach, they first extract all possible inflectional forms for each of the constituent tokens. Then, they sequentially iterate through the tokens in the sentence, and for each token, they select an inflectional form that results in the worst BLEU score. Once an adversarial form is identified, they directly replace the form in the original sentence and continue to the next token. While a similar approach is possible in our setup, we found their algorithm to be computationally expensive as it prevents from performing efficient batching.

It is important to note that neither MORPHEUS-MULTILINGUAL nor the original MORPHEUS exhaustively searches over all possible sentences, due to memory and time constraints. However, our approach in MORPHEUS-MULTILINGUAL can be efficiently implemented and reduces the inference time by almost a factor of 20. We experiment on 11 different language pairs, therefore, the run time and computational costs are critical to our experiments.

3 Experiments

In this section, we present a comprehensive evaluation of the robustness of X→English machine translation systems. Since it is natural for NMT models to be more robust when trained on large amounts of parallel data, we experiment with two

[2] https://github.com/salesforce/morpheus

[3] https://unimorph.github.io/

[4] https://github.com/antonisa/unimorph_inflect

[5] Some dictionaries might contain fewer POS tags, for example, in German we are restricted to just nouns and verbs.

[6] N is a hyperparameter, and in our preliminary experiments, we find $N = 50$ to be sufficiently high to generate many uniquely perturbed sentences and also keep the process computationally tractable.

PRON	VERB	PART	PUNCT	ADV	NOUN	VERB	AUX
Sie	wissen	nicht	,	wann	Räuber	kommen	können
you-NOM.3PL	knowledge-PRS.3PL	not-NEG	,	when	robber-NOM.PL	come-NFIN	can-PRS.3PL
(*) Sie	wissten	nicht	,	wann	Räuber	kommen	können
(*) Sie	wissen	nicht	,	wann	Räuber	kommen	könne
(*) Sie	wisse	nicht	,	wann	Räuber	kommen	könnte

Table 1: Example inflectional perturbations on a German text.

sets of translation systems. First, we use state-of-the-art pre-trained models for Russian→English and German→English from `fairseq` (Ott et al., 2019).[7] Secondly, we use the multilingual TED corpus (Qi et al., 2018) to train transformer-based translation systems from scratch.[8] Using the TED corpus allows us to expand our evaluation to a larger pool of language pairs.

3.1 WMT19 Pretrained Models

We evaluate the robustness of best-performing systems from WMT19 news translation shared task (Barrault et al., 2019), specifically for Russian→English and German→English (Ott et al., 2019). We follow the original work and use *newstest2018* as our test set for adversarial evaluation. Using the procedure described in §2.2, we create adversarial versions of *newstest2018* for both the language pairs. In Table 2, we present the baseline and adversarial results using BLEU and chrF metrics. For both the language pairs, we notice significant drops on both metrics. Before diving further into the qualitative analysis of these MT systems, we first present a broader evaluation on MT systems trained on multilingual TED corpus.

lg	**Baseline**		**Adversarial**		
	BLEU	chrF	BLEU	chrF	NR
`rus`	38.33	0.63	18.50	0.47	0.81
`deu`	48.40	0.70	33.43	0.59	1.00

Table 2: Baseline & Adversarial results on *newstest2018* using `fairseq`'s pre-trained models. NR denotes Target-Source Noise Ratio (2).

3.2 TED corpus

The multilingual TED corpus (Qi et al., 2018) provides parallel data for over 50 language pairs, but in our experiments we only use a subset of these language pairs. We selected our test language pairs (X→English) to maximize the diversity in language families, as well as the resources available for training MT systems. Since we rely on UniMorph and `unimorph_inflect` for generating perturbations, we only select languages that have reasonably high accuracy in `unimorph_inflect` (>80%). Table 3 presents an overview of the chosen source languages, along with the information on language family and training resources.

We also quantify the morphological richness for the languages listed in Table 3. As we are not aware of any standard metric to gauge morphological richness of a language, we use the reinflection dictionaries to define this metric. We compute the morphological richness using the Type-Token Ratio (TTR) as follows,

$$\mathrm{TTR}_{lg} = \frac{N_{\mathrm{types}}(lg)}{N_{\mathrm{tokens}}(lg)} = \frac{N_{\mathrm{paradigms}}(lg)}{N_{\mathrm{forms}}(lg)} \qquad (1)$$

In Table 3, we report the TTR_{lg} scores measured on UniMorph dictionaries as well as on the UniMorph-style dictionaries constructed from TED dev splits using `unimorph_inflect` tool. Note that, TTR_{lg}, as defined here, slightly differs from the widely known Type-Token ration used for measuring lexical diversity (or richness) of a corpus.

We run MORPHEUS-MULTILINGUAL to generate adversarial sentences for the validation splits of the TED corpus. We term a sentence adversarial if it leads to the maximum drop in chrF score. Note that, it is possible to have perturbed sentences that may not lead to any drop in chrF scores. In Figure 1, we plot the fraction of perturbed sentences along with adversarial fraction for each of the source languages. We see considerable perturbations for most languages, with the exception of Swedish, Lithuanian, Ukrainian, and Estonian.

[7] Due to resource constraints, we only experiment with the single models and leave the evaluation of ensemble models for future work.

[8] For the selected languages, we train an MT model with *'transformer_iwslt_de_en'* architecture from `fairseq`. We use a sentence-piece vocab size of 8000, and train up to 80 epochs with Adam optimizer (see A.2 in Appendix for more details)

lg	Family	Resource	TTR
heb	Semetic	High	(0.044, 0.191)
rus	Slavic	High	(0.080, 0.107)
tur	Turkic	High	(0.016, 0.048)
deu	Germanic	High	(0.210, 0.321)
ukr	Slavic	High	(0.103, 0.143)
ces	Slavic	High	(0.071, 0.082)
swe	Germanic	Medium	(0.156, 0.281)
lit	Baltic	Medium	(0.051, 0.084)
slv	Slavic	Low	(0.109, 0.087)
kat	Kartvelian	Low	(0.057, ——)
est	Uralic	Low	(0.026, 0.056)

Table 3: List of language chosen from multilingual TED corpus. For each language, the table presents the language family, resource level as the Type-Token ratio (TTR_{lg}). We measure the ratio using the types and tokens present in the reinflection dictionaries (UniMorph, lexicon from TED *dev*)

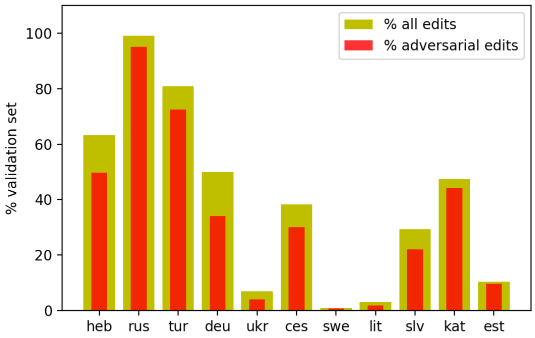

Figure 1: Perturbation statistics for selected TED languages

In preparing our adversarial set, we retain the original source sentence if we fail to create any perturbation or if none of the identified perturbations lead to a drop in chrF score. This is to make sure the adversarial set has the same number of sentences as the original validation set. In Table 4, we present the baseline and adversarial MT results. We notice a considerable drop in performance for Hebrew, Russian, Turkish and Georgian. As expected, the % drops are correlated to the perturbations statistics from Figure 1.

3.3 Translating Learner's Text

In the previous sections (§3.1, §3.2), we have seen the impact of noisy inputs to MT systems. While, these results indicate a need for improving the robustness of MT systems, the above-constructed adversarial sets are however synthetic. In this section, we evaluate the impact of morphological inflection related errors directly on learners' text.

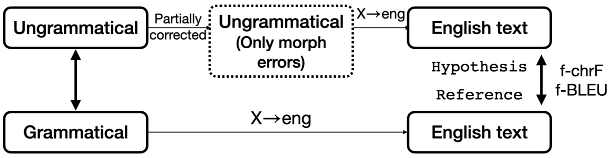

Figure 2: Schematic for preliminary evaluation on learners' language text. This is similar to the methodology used in Anastasopoulos (2019).

To this end, we utilize two grammatical error correction (GEC) datasets, German Falko-MERLIN-GEC (Boyd, 2018), Russian RULEC-GEC (Rozovskaya and Roth, 2019). Both of these datasets contain labeled error types relating to word morphology. Evaluating the robustness on these datasets will give us a better understanding of the performance on actual text produced by second language (L2) speakers.

Unfortunately, we don't have gold English translations for the grammatically incorrect (or corrected) text from GEC datasets. While there is a related prior work (Anastasopoulos et al., 2019) that annotated Spanish translations for English GEC data, we are not aware of any prior work that provide gold English translations for grammatically incorrect data in non-English languages. Therefore, we propose a pseudo-evaluation methodology that allows for measuring robustness of MT systems. A schematic overview of our methodology is presented in Figure 2. We take the ungrammatical text and use the gold GEC annotations to correct all errors except for the morphology related errors. We now have ungrammatical text that only contains morphology related errors and it is similar to the perturbed outputs from MORPHEUS-MULTILINGUAL. Since, we don't have gold translations for the input Russian/German sentences, we use the machine translation output of the fully grammatical text as reference and the translation output of partially-corrected text as hypothesis. In Table 5, we present the results on both Russian and German learners' text.

Overall, we find that the pre-trained MT models from `fairseq` are quite robust to noise in learners' text. We manually inspected some examples, and found the MT systems to sufficiently robust to morphological perturbations and changes in the output translation (if any) are mostly warranted.

X→English	Code	# train	Baseline		Adversarial		
			BLEU	chrF	BLEU	chrF	NR
Hebrew	heb	211K	40.06	0.5898	33.94 (-15%)	0.5354 (-9%)	1.56
Russian	rus	208K	25.64	0.4784	11.70 (-54%)	0.3475 (-27%)	1.03
Turkish	tur	182K	27.77	0.5006	18.90 (-32%)	0.4087 (-18%)	1.43
German	deu	168K	34.15	0.5606	31.29 (-8%)	0.5373 (-4%)	1.82
Ukrainian	ukr	108K	25.83	0.4726	25.66 (-1%)	0.4702 (-1%)	2.96
Czech	ces	103K	29.35	0.5147	26.58 (-9%)	0.4889 (-5%)	2.11
Swedish	swe	56K	36.93	0.5654	36.84 (-0%)	0.5646 (-0%)	3.48
Lithuanian	lit	41K	18.88	0.3959	18.82 (-0%)	0.3948 (-0%)	3.42
Slovenian	slv	19K	11.53	0.3259	10.48 (-9%)	0.3100 (-5%)	3.23
Georgian	kat	13K	5.83	0.2462	4.92 (-16%)	0.2146 (-13%)	2.49
Estonian	est	10K	6.68	0.2606	6.53 (-2%)	0.2546 (-2%)	4.72

Table 4: Results on multilingual TED corpus (Qi et al., 2018)

Dataset	f-BLEU	f-chrF
Russian GEC	85.77	91.56
German GEC	89.60	93.95

Table 5: Translation results on Russian and German GEC corpora. An oracle (aka. fully robust) MT system would give a perfect score. We adopt the *faux*-BLEU terminology from Anastasopoulos (2019). f-BLEU is identical to BLEU, except that it is computed against a pseudo-reference instead of true reference.

Viewing these results in combination with results on TED corpus, we believe that X→English are robust to morphological perturbations at source as long as they are trained on sufficiently large parallel corpus.

4 Analysis

To better understand what makes a given MT system to be robust to morphology related grammatical perturbations in source, we present a thorough analysis of our results and also highlight a few limitations of our adversarial methodology.

Adversarial Dimensions: To quantify the impact of each inflectional perturbation, we perform a fine-grained analysis on the adversarial sentences obtained from multilingual TED corpus. For each perturbed token in the adversarial sentence, we identify the part-of-speech (POS) and the feature dimension(s) (dim) perturbed in the token. We uniformly distribute the % drop in sentence-level chrF score to each (POS, dim) perturbation in the adversarial sentence. This allows us to quantitatively compare the impact of each perturbation type (POS, dim) on the overall performance of MT model. Additionally, as seen in Figure 1, all inflectional perturbations need not cause a drop in chrF (or BLEU) scores. The adversarial sentences only capture the worst case drop in chrF. Therefore, to analyze the overall impact of the each perturbation (POS, dim), we also compute the impact score on the entire set of perturbed sentences explored by MORPHEUS-MULTILINGUAL.

Table 8 (in Appendix) presents the results for all the TED languages. First, the trends for adversarial perturbations is quite similar to all explored perturbations. This indicates that the adversarial impact of a perturbation is not determined by just the perturbation type (POS, dim) but is lexically dependent.

Evaluation Metrics: In the results presented in §3, we reported the performance using BLEU and chrF metrics (following prior work (Tan et al., 2020)). We noticed significant drops on these metrics, even for high-resource languages like Russian, Turkish and Hebrew, including the state-of-the-art `fairseq` models. To better understand these drops, we inspected the output translations of adversarial source sentences. We found a number of cases where the new translation is semantically valid but both the metrics incorrectly score them low (see S2 in Table 6). This is a limitation of using surface level metrics like BLEU/chrF.

Additionally, we require the new translation to be as close as possible to the original translation, but this can be a strict requirement on many occasions. For instance, if we changed a noun in the

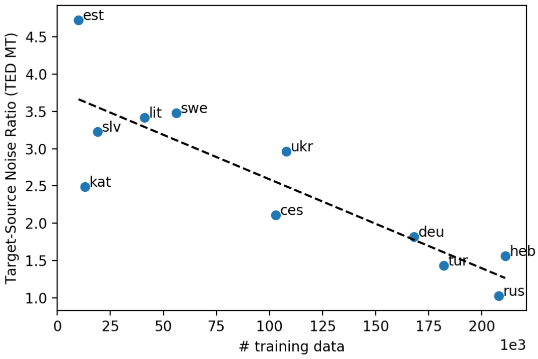

Figure 3: Correlation between Noise Ratio (NR) and # train. The results indicate that, larger the training data, the models are more robust towards source perturbations (NR≈1).

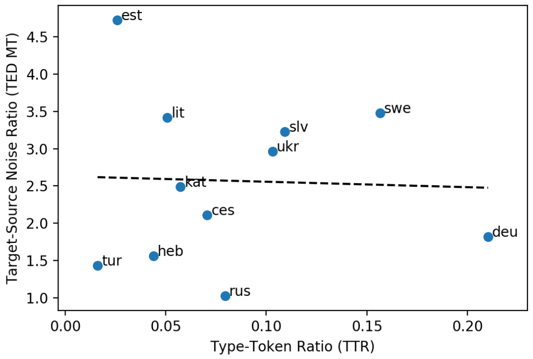

Figure 4: Correlation between Target-Source Noise Ratio (NR) on TED machine translation and Type-Token Ratio (TTR$_{lg}$) of the source language (from UniMorph). The results indicate that the morphological richness of the source language doesn't necessarily correlate to NMT robustness.

source from its singular to plural form, it is natural to expect a robust translation system to reflect that change in the output translation. To account for this behavior, we compute Target-Source Noise Ratio (NR) metric from Anastasopoulos (2019). NR is computed as follows,

$$\text{NR}(s, t, \tilde{s}, \tilde{t}) = \frac{100 - \text{BLEU}(t, \tilde{t})}{100 - \text{BLEU}(s, \tilde{s})} \qquad (2)$$

The ideal NR is $\sim$1, where a change in the source ($s \rightarrow \tilde{s}$) results in a proportional change in the target ($t \rightarrow \tilde{t}$). For the adversarial experiments on TED corpus, we compute the NR metric for each language pair and the results are presented in Table 4. Interestingly, while Russian sees a major drop in BLEU/chrF score, the noise ratio is close to 1. This indicates that the Russian MT is actually quite robust to morphological perturbations. Furthermore, in Figure 3, we present a correlation analysis between the size of parallel corpus available for training *vs* noise ratio metric. We see a very strong negative correlation, indicating that high-resource MT systems (e.g., heb, rus, tur) are quite robust to inflectional perturbations, inspite of the large drops in BLEU/chrF scores. Additionally, we noticed that morphological richness of the source language (measured via TTR in Table 3) doesn't play any significant role in the MT performance under adversarial settings (e.g., rus, tur *vs* deu). The scatter plot between TTR and NR for TED translation task is presented in Figure 4.

Morphological Richness: To analyze the impact of morphological richness of source, we look deeper into the Slavic language family. We experimented with four languages within the Slavic family, Czech, Ukranian, Russian and Slovenian. All except Slovenian are high-resource. These languages differ significantly in their morphological richness (TTR) with, $\text{TTR}_{\texttt{ces}} < \text{TTR}_{\texttt{slv}} << \text{TTR}_{\texttt{rus}} << \text{TTR}_{\texttt{ukr}}$.[9] As we have already seen in above analysis (see Figure 4), morphological richness isn't indicative of the noise ratio (NR), and this behavior is also true for Slavic languages. We now check if morphological richness determines the drop in BLEU/chrF scores? In fact, we find that this is also not the case. We see larger % drop for rus as compared to slv or ukr. We instead notice that the % drop in BLEU/chrF is dependent on the % edits we make to the validation set. The % edits we were able to make follows the order, $\delta_{\texttt{rus}} >> \delta_{\texttt{ces}} > \delta_{\texttt{slv}} >> \delta_{\texttt{ukr}}$ (see Figure 1).

While NR is driven by size of training set, and % drop in BLEU is driven by % edits to the validation set. The % edits in turn depends on the size of UniMorph dictionaries and not on morphological richness of the language. Therefore, we conclude that both the metrics, % drop in BLEU/chrF and NR are *dependent on the resource size* (parallel data and UniMorph dictionaries) and *not on the morphological richness of the language*.

Semantic Change: In our adversarial attacks, we aim to create a ungrammatical source via inflectional edits and evaluate the robustness of systems for these edits. While these adversarial attacks can help us discover any significant biases in the transla-

[9] TTR$_{lg}$ measured on lexicons from TED dev splits.

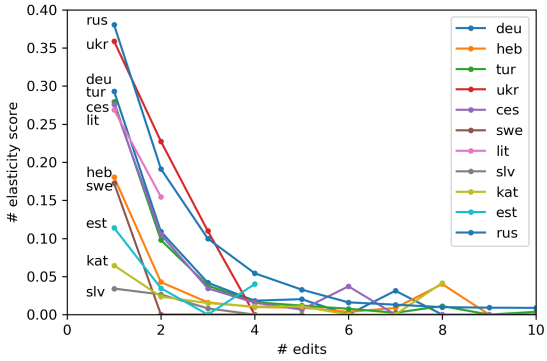

Figure 5: Elasticity score for TED languages

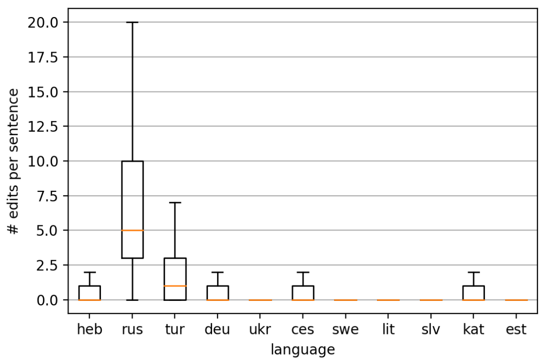

Figure 6: Boxplots for the distribution of # edits per sentence in the adversarial TED validation set.

tion systems, they can often lead to unintended consequences. Consider the example Russian sentence S1 (s) from Table 6. The sentence is grammatically correct, with the subject Тренер ('Coach') and object игрока ('player') in NOM and ACC cases respectively. If were perturb this sentence to A-S1 ($\tilde{s}$), the new words Тренера ('Coach'), and игрок ('player') are now in ACC and NOM cases respectively. Due to case assignment phenomenon in Russian, this perturbation ($s \rightarrow \tilde{s}$) has essentially swapped the subject and object roles in the Russian sentence. As we can see in the example, the English translation, $\tilde{t}$ (A-T1) does in fact correctly capture this change. This indicates that our attacks can sometimes lead to significant change in the semantics of the source sentence. Handling such cases would require deeper understanding of each language grammar and we leave this for future work.

Elasticity: As we have seen in discussion on noise ratio, it is natural for MT systems to transfer changes in source to the target. However, inspired by (Anastasopoulos, 2019), we wanted to understand how this behavior changes as we increase the number of edits in the source sentence. For this purpose, we first bucket all the explored perturbed sentences based on the number of edits (or perturbations) from the original source. Within each bucket, we compute the fraction of perturbed source sentences that result in same translation as the original source. We define this fraction as the *elasticity* score, i.e. whether the translation remains the same under changes in source. Figure 5 presents the results and we find the elasticity score dropping quickly to zero as the # edits increase. Notably, ukr drops quickly to zero, while rus retains reasonable elasticity score for higher number of edits.

Aggressive edits: Our algorithm doesn't put any restrictions on the number of tokens that can be perturbed in a given sentence. This can lead to aggressive edits, especially in languages like Russian that are morphologically-rich and the reinflection lexicons are sufficiently large. As we illustrate in Figure 6, median edits per sentence in rus is 5, significantly higher than the next language (tur at 1). Such aggressive edits in Russian can lead to unrealistic sentences, and far from our intended simulation of learners' text. We leave the idea of thresholding # edits to future work.

Adversarial Training: In an attempt to improve robustness of NMT systems against morphological perturbations, we propose training NMT models with augmenting adversarially perturbed sentences. Due to computational constraints, we evaluate this setting only for slv. We follow the strategy outlined in Section 2 to obtain adversarial perturbations for TED corpus training data. We observe that the adversarially trained model performs marginally poorer (BLEU 10.30 from 10.48 when trained without data augmentation). We hypothesize that this could possibly due to small training data, and believe that this training setting can better benefit models with already high BLEU scores. We leave extensive evaluation and further analysis on adversarial training to future work.

5 Conclusion

In this work, we propose MORPHEUS-MULTILINGUAL, a tool to analyze the robustness of X→English NMT systems under morphological perturbations. Using this tool, we experiment with 11 different languages selected from diverse language families with varied training resources.

rus	S1	Source (s)	Тренер полностью поддержал игрока.
	T1	Target (t)	The coach fully supported the player.
	A-S1	Source ($\tilde{s}$)	Тренера полностью поддержал игрок.
	A-T1	Target ($\tilde{t}$)	The coach was fully supported by the player.
deu	S2	Source (s)	Dinosaurier benutzte Tarnung, um seinen Feinden auszuweichen
	T2	Target (t)	Dinosaur used camouflage to evade its enemies (1.000)
	A-S2	Source ($\tilde{s}$)	Dinosaurier benutze Tarnung, um seinen Feindes auszuweichen
	A-T2	Target ($\tilde{t}$)	Dinosaur Use camouflage to dodge his enemy (0.512)
rus	S3	Source (s)	У нас вообще телесные наказания не редкость.
	T3	Target (t)	In general, corporal punishment is not uncommon. (0.885)
	A-S3	Source ($\tilde{s}$)	У нас вообще телесных наказании не редкостях.
	A-T3	Target ($\tilde{t}$)	We don't have corporal punishment at all. (0.405)
rus	S4	Source (s)	Вот телесные наказания - спасибо, не надо.
	T4	Target (t)	That's corporal punishment - thank you, you don't have to. (0.458)
	A-S4	Source ($\tilde{s}$)	Вот телесных наказаний - спасибах, не надо.
	A-T4	Target ($\tilde{t}$)	That's why I'm here. (0.047)
deu	S5	Source (s)	Die Schießereien haben nicht aufgehört.
	T5	Target (t)	The shootings have not stopped. (0.852)
	A-S5	Source ($\tilde{s}$)	Die Schießereien habe nicht aufgehört.
	A-T5	Target ($\tilde{t}$)	The shootings did not stop, he said. (0.513)
rus	S6	Source (s)	Всякое бывает.
	T6	Target (t)	Anything happens. (0.587)
	A-S6	Source ($\tilde{s}$)	Всякое будете бывать.
	A-T6	Target ($\tilde{t}$)	You'll be everywhere. (0.037)
kat	S7	Source (s)	ნამდვილი სკოლა.
	T7	Target (t)	It 's a real school. (0.821)
	A-S7	Source ($\tilde{s}$)	ნამდვილნი სკოლები.
	A-T7	Target ($\tilde{t}$)	There 's a man who 's friend. (0.107)
est	S8	Source (s)	Ning meie laste tuleviku varastamine saab ühel päeval kuriteoks.
	T8	Target (t)	And our children 's going to be the future of our own day. (0.446)
	A-S8	Source ($\tilde{s}$)	Ning meie laptegs tuleviku varastamine saab ühel päeval kuriteoks.
	A-T8	Target ($\tilde{t}$)	And our future is about the future of the future. (0.227)
est	S9	Source (s)	Nad pagevad üle piiride nagu see.
	T9	Target (t)	They like that overdights like this. (0.318)
	A-S9	Source ($\tilde{s}$)	Nad pagevad üle piirete nagu see.
	A-T9	Target ($\tilde{t}$)	They dress it as well as well. (0.141)
rus	S10	Source (s)	Мой дедушка был необычайным человеком того времени.
	T10	Target (t)	My grandfather was an extraordinary man at that time. (0.802)
	A-S10	Source ($\tilde{s}$)	Мой дедушка будё необычайна человеков того времи.
	A-T10	Target ($\tilde{t}$)	My grandfather is incredibly harmful. (0.335)

Table 6: Qualitative analysis. (1) semantic change, (2) issues with evaluation metrics, (3,4,5,6,7,10) good examples for attacks, (8) poor attacks, (9) poor translation quality ($s \rightarrow t$)

We evaluate NMT models trained on TED corpus as well as pretrained models readily available as part of `fairseq` library. We observe a wide range of 0-50% drop in performances under adversarial setting. We further supplement our experiments with an analysis on GEC-learners corpus for Russian and German. We qualitatively and quantitatively analyze the perturbations created by our methodology and presented its strengths as well as limitations, outlining some avenues for future research towards building more robust NMT systems.

References

Md Mahfuz Ibn Alam and Antonios Anastasopoulos. 2020. Fine-tuning MT systems for robustness to second-language speaker variations. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 149–158, Online. Association for Computational Linguistics.

Antonios Anastasopoulos. 2019. An analysis of source-side grammatical errors in NMT. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 213–223, Florence, Italy. Association for Computational Linguistics.

Antonios Anastasopoulos, Alison Lui, Toan Q. Nguyen, and David Chiang. 2019. Neural machine translation of text from non-native speakers. 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 3070–3080, Minneapolis, Minnesota. Association for Computational Linguistics.

Antonios Anastasopoulos and Graham Neubig. 2019. Pushing the limits of low-resource morphological inflection. 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 984–996, Hong Kong, China. 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, et al. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations*.

Adriane Boyd. 2018. Using Wikipedia edits in low resource grammatical error correction. In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pages 79–84, Brussels, Belgium. Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Arya D. McCarthy, Christo Kirov, Matteo Grella, Amrit Nidhi, Patrick Xia, Kyle Gorman, Ekaterina Vylomova, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, Timofey Arkhangelskiy, Nataly Krizhanovsky, Andrew Krizhanovsky, Elena Klyachko, Alexey Sorokin, John Mansfield, Valts Ernštreits, Yuval Pinter, Cassandra L. Jacobs, Ryan Cotterell, Mans Hulden, and David Yarowsky. 2020. UniMorph 3.0: Universal Morphology. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3922–3931, Marseille, France. European Language Resources Association.

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

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.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. 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.

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 529–535, New Orleans, Louisiana. Association for Computational Linguistics.

Alla Rozovskaya and Dan Roth. 2019. Grammar error correction in morphologically rich languages: The case of Russian. *Transactions of the Association for Computational Linguistics*, 7:1–17.

Samson Tan, Shafiq Joty, Min-Yen Kan, and Richard Socher. 2020. It's morphin' time! Combating linguistic discrimination with inflectional perturbations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2920–2935, Online. Association for Computational Linguistics.

A Appendix

A.1 UniMorph Example

An example from German UniMorph dictionary is presented in Table 7.

Paradigm	**Form**	**MSD**
abspielen ('play')	abgespielt ('played')	V.PTCP;PST
abspielen ('play')	abspielend ('playing')	V.PTCP;PRS
abspielen ('play')	abspielen ('play')	V;NFIN

Table 7: Example inflections for German verb *abspielen* ('play') from the UniMorph dictionary.

A.2 MT training

For all the languages in TED corpus, we train Any→English using the `fairseq` toolkit. Specifically, we use the 'transformer_iwslt_de_en' architecture, and train the model using Adam optimizer. We use an inverse square root learning rate scheduler with warm-up update steps of 4000. In the linear warm-up phase, we use an initial learning rate of 1e-7 until a configured rate of 2e-4. We use cross entropy criterion with label smoothing of 0.1.

A.3 Dimension Analysis

Dimension	ces	deu	est	heb	kat	lit	rus	slv	swe	tur	ukr
ADJ.Animacy	-	-	-	-	-	-	3.51$_{(0.89)}$	-	-	-	-
ADJ.Case	4.31$_{(0.81)}$	-	-	-	10.67$_{(2.59)}$	-	4.78$_{(0.91)}$	-	5.05$_{(5.05)}$	-	6.04$_{(1.10)}$
ADJ.Comparison	-	-	-	-	-	7.99$_{(0.46)}$	-	-	-	-	-
ADJ.Gender	3.83$_{(0.78)}$	-	-	-	-	6.81$_{(-1.35)}$	5.30$_{(1.00)}$	-	-	-	-
ADJ.Number	4.07$_{(0.78)}$	-	-	-	13.90$_{(1.52)}$	6.31$_{(-2.26)}$	4.67$_{(0.94)}$	-	5.05$_{(5.05)}$	7.92$_{(2.23)}$	6.25$_{(1.29)}$
ADJ.Person	-	-	-	-	-	-	-	-	-	8.89$_{(2.43)}$	-
N.Animacy	-	-	-	-	-	-	6.53$_{(1.19)}$	-	-	-	-
N.Case	6.94$_{(0.81)}$	6.39$_{(1.26)}$	12.35$_{(1.50)}$	-	15.38$_{(0.98)}$	-	6.65$_{(1.20)}$	-	4.29$_{(1.05)}$	14.39$_{(2.37)}$	10.28$_{(7.66)}$
N.Definiteness	-	-	-	-	-	-	-	-	8.36$_{(1.61)}$	-	-
N.Number	5.44$_{(0.77)}$	5.70$_{(1.27)}$	8.10$_{(1.33)}$	16.22$_{(5.92)}$	14.46$_{(0.66)}$	-	6.12$_{(1.22)}$	-	4.30$_{(1.52)}$	13.08$_{(2.31)}$	21.20$_{(15.96)}$
N.Possession	-	-	-	12.63$_{(4.31)}$	-	-	-	-	-	-	-
V.Aspect	-	-	-	-	14.17$_{(-0.38)}$	-	-	-	-	-	-
V.Gender	-	-	-	-	-	-	6.52$_{(1.51)}$	-	-	-	-
V.Mood	13.17$_{(2.78)}$	15.89$_{(2.77)}$	-	-	11.11$_{(0.58)}$	-	-	21.49$_{(3.73)}$	-	-	-
V.Number	8.23$_{(2.72)}$	32.86$_{(8.12)}$	-	13.78$_{(4.60)}$	9.02$_{(1.33)}$	-	6.23$_{(1.44)}$	21.47$_{(-9.47)}$	-	-	-
V.Person	6.58$_{(2.69)}$	6.22$_{(1.50)}$	-	10.86$_{(4.99)}$	12.37$_{(1.33)}$	-	6.10$_{(1.29)}$	-	-	-	-
V.Tense	-	-	-	17.52$_{(7.13)}$	13.09$_{(1.05)}$	-	6.59$_{(1.61)}$	-	-	-	-
V.CVB.Tense	-	-	-	-	-	-	6.70$_{(0.87)}$	-	-	-	9.09$_{(2.62)}$
V.MSDR.Aspect	-	-	-	-	14.39$_{(4.68)}$	-	-	-	-	-	-
V.PTCP.Gender	10.28$_{(2.75)}$	-	-	-	-	-	-	-	-	-	-
V.PTCP.Number	9.31$_{(2.51)}$	-	-	-	-	-	-	-	-	-	-

Table 8: Fine-grained analysis of X→English translation performance w.r.t the perturbation type (POS, Morphological feature dimension). The number reported in this table indicate the average *% drop* in sentence level chrF for an adversarial pertubation on a token with POS on the dimension (dim). The numbers in the parentheses indicate average *% drop* for all the tested perturbations including the adversarial perturbations.