Fine-Tuning MT systems for Robustness to Second-Language Speaker Variations

Md Mahfuz Ibn Alam and Antonios Anastasopoulos
Department of Computer Science, George Mason University
{malam21,antonis}@gmu.edu

Abstract

The performance of neural machine translation (NMT) systems only trained on a single language variant degrades when confronted with even slightly different language variations. With this work, we build upon previous work to explore how to mitigate this issue. We show that fine-tuning using *naturally occurring* noise along with pseudo-references (i.e. "corrected" non-native inputs translated using the baseline NMT system) is a promising solution towards systems robust to such type of input variations. We focus on four translation pairs, from English to Spanish, Italian, French, and Portuguese, with our system achieving improvements of up to 3.1 BLEU points compared to the baselines, establishing a new state-of-the-art on the JFLEG-ES dataset.[1]

1 Introduction

Neural machine translation (NMT) approaches have aided the machine translation field in achieving great advances in the recent years, starting with encoder-decoder models with attention (Bahdanau et al., 2014; Luong et al., 2015), to transformers using self-attention (Vaswani et al., 2018), to massively multilingual models that yield large improvements even in low-resource settings (Aharoni et al., 2019; Zhang et al., 2020).

Despite these very encouraging developments, the list of shortcomings of NMT is also quite vast (Koehn and Knowles, 2017), and one of the most crucial shortcomings is the lack of robustness to source-side noise.[2] When confronted with inputs that are even slightly different from the inputs that the models were trained on, the quality of the outputs significantly degrades. This observation has been confirmed for noise due to typos or character scrambling (Belinkov and Bisk, 2018), due to faulty speech recognition (Heigold et al., 2018), or due to naturally-occurring errors by second-language non-native speakers (Anastasopoulos et al., 2019).

However, this issue can particularly degrade the user experience for millions of potential users. For example, the number of non-native English speakers is three times larger than the number of native English speakers (c.f. around 1 billion for the former and about 300 million for the latter). Had one had access to large amounts of data for all different language varieties, it would be straightforward to train variety-specific MT models. Such data, though, are of course scarce.

In this paper we work on addressing this particular shortcoming, in an attempt to make NMT systems more robust to source-side variations that non-native speakers produce. Since English is the language with the largest amount of second-language learners and non-native speakers, we only focus on MT systems translating out of English, but we point out that such work is urgently needed for other colonial languages (i.e. French, Spanish) or majority languages (such as Russian, Mandarin, or Hindi) that are taking over minority ones.[3]

The main difference of our approach compared to previous work is that we do not attempt to synthesize different types of noise, but rather use *naturally-occurring* texts, as produced by non-native speakers. We utilize grammar error correction corpora and produce pseudo-references, which we then use to fine-tune a NMT system with a goal of increasing its robustness to such source-side noise. In our view, our approach has two main advantages and a single disadvantage over previous

[1] All datasets and code are publicly available here: https://github.com/mahfuzibnalam/finetuning_for_robustness.

[2] This is not to say that non-neural statistical approaches did not suffer from the same drawbacks.

[3] That is also not to say that robustness is not necessary for low-resource languages; to the contrary! We just focus on high-resource settings first as they are the ones that have the potential to affect a larger number of downstream users.

Proceedings of the 2020 EMNLP Workshop W-NUT: The Sixth Workshop on Noisy User-generated Text, pages 149–158

works. First, the types of *realistic* "non-native-like" noise that can be synthesized are limited, covering among others typos or simple morphological or syntactic mistakes (Belinkov and Bisk, 2018; Cheng et al., 2018; Anastasopoulos et al., 2019; Tan et al., 2020, et alia) but not covering the interplay between all these or any other higher level issues (e.g. word choice). Our approach has the potential to handle a larger spectrum of language variation, as it appears in naturally occurring data. Second, our choice of fine-tuning, rather than training from scratch as previous works have opted for, leads to lower training times and lower compute needed for similar improvements on robustness. The main drawback of our approach lies in the need for corrected (or "normalized") versions of "noisy" non-native sentences, but we take solace in the fact that at least for the majority of the high-resource languages (such as English, French, German, Russian, or Chinese) such datasets already exist. Very briefly, our contributions are summarized here:

- We show that fine-tuning a pre-trained system on noisy source-side data along with pseudo-references is a viable approach towards NMT robustness to grammar errors and input from non-native speakers.
- We show that fine-tuning of a multilingual NMT system on several languages is also advisable, yielding better performance for a subset of the languages.
- We also discuss the potential of achieving zero-shot robustness, as long as catastrophic forgetting issues can be overcome.

2 Related Work

Our work is inspired by and combines two lines of research: (1) robustness studies in NMT and (2) data augmentation.

Robust NMT Making robust models for NMT has recently gained popularity, with Shared Tasks organized in the Conference of Machine Translation (Li et al., 2019) and several solutions put forth (Berard et al., 2019; Helcl et al., 2019; Post and Duh, 2019; Zhou et al., 2019; Zheng et al., 2019, et alia). Liu et al. (2018); Karpukhin et al. (2019) focus on creating black-box methods for making synthetic or natural noises. Ebrahimi et al. (2018) use white-box methods and creates adversarial examples for character-level NMT. Anastasopoulos et al. (2019) show that including noisy synthetic data in the training data can increase the model's robustness without sacrificing performance on clean data, an approach that Tan et al. (2020) extend to more NLP tasks.

While these approaches are indeed meritorious and indeed improve a model's robustness, we argue that one needs to use *natural* noise instead, on account of two phenomena. The first is language change: the different variations that the models will have to contend with are not static, but rather constantly changing at an ever-increasing pace. Second, and perhaps a partial direct consequence of the first point, one cannot rely on synthetic examples to properly capture the wide variety of naturally-occurring variations. Besides, if one could properly model noise creation, they could also similarly model the inverse problem adequately, namely remove said noise, in which case a noise-removing preprocessing step would be most likely suffiecient to tackle the issue.

On working with real-world noise, the approach of Michel and Neubig (2018) is the most similar to ours. They collected "noisy" English, French, and Japanese sentences from Reddit, created translations, and split their dataset (MTNT) into train, development, and test, ranging from 5 to 36 thousand training examples. To build robust NMT systems, they first train a model on standard clean data and then fine-tune it on the training portion of MTNT using techniques from domain adaptation. The main difference between this worthy effort and our approach is two-fold. First, our approach does not require gold translations of the noisy inputs, which can be expensive and hard to collect, but we instead rely on the abundance of corrected second-language learner data (which we use to create pseudo-references, see §3). Second, we attest that Reddit language translation is much closer to a domain adaptation scenario, and includes additional noise types that are not pertinent to non-native language translation such as emoji, Reddit jargon such us "upvote" or "gild", and internet slang such as "tbh" and "smh".[4]

On working with pseudo-references, the approach of Cheng et al. (2019a) is the most similar to ours. They use ASR corpora to create synthetic ASR-induced noise and try to make NMT system more robust to this type of noise. As speech-to-transcription-to-translation datasets are

[4] "tbh" stands for "to be honest" and "smh" for "shake my head".

very costly to produce, they use standard speech-to-transcription datasets instead. They translate the gold transcription data set to get translation pseudo-references. Then they jointly train the model on noisy source transcription using the pseudo-reference translations as the target.

Data Augmentation Data augmentation techniques have become increasingly popular for MT and other NLP tasks, from back-translation of monolingual data (Sennrich et al., 2016) to counterfactual augmentation to address gender bias issues (Zmigrod et al., 2019, et alia). For our purposes, we will focus on data augmentation techniques aimed at increasing NMT robustness.

Simple perturbations typically used include the infusion of character-level noise (e.g. character scrambling (Heigold et al., 2018) or typos (Belinkov and Bisk, 2018)) or word order scrambling (Sperber et al., 2017). Cheng et al. (2018, 2019b) propose a gradient based method to attack the translation model with adversarial source examples, but there's not guarantee that the adversarial attack results in realistic noise (Michel et al., 2019a). Anastasopoulos et al. (2019) add specific types of errors (such as subject-verb agreement or determiner errors) on the source-side of parallel data, while Tan et al. (2020) specifically perturb the inflectional morphology of words to create adversarial examples and show that adversarial fine-tuning them for a single epoch significantly improves robustness. Our work is highly motivated from these last two works, but instead of creating synthetic perturbed adversarial examples we use real noisy examples.

3 Fine-tuning for Robustness

Our goal is to achieve robustness to source-side variations that are similar to the mistakes that non-native English speakers make. To do so, we will utilize state-of-the-art pretrained systems and fine-tune them using pseudo-references over corpora that include *real-world* noise. The general outline of our approach is straightforward:

1. Start with a English-to-X NMT system pre-trained on any available data.
2. Obtain an English Grammar Error Correction dataset, which provides tuples $(\mathbf{x}, \tilde{\mathbf{x}})$ of original and corrected sentences.
3. Translate the corrected sentences obtaining pseudo-references $\tilde{\mathbf{y}} = \mathrm{NMT}(\tilde{\mathbf{x}})$.
4. Fine-tune the NMT system on $(\mathbf{x}, \tilde{\mathbf{y}})$ pairs.

Notation Throughout this work, we use the notation of Anastasopoulos (2019) to denote different types of data:

- $\mathbf{x}$: the original, noisy, potentially ungrammatical English sentence. Its tokens will be denoted as x_i.
- $\tilde{\mathbf{x}}$: the English sentence with the correction annotations applied to the original sentence $\mathbf{x}$, which is deemed fluent and grammatical. Again, its tokens will be denoted as $\tilde{x}_i$.
- $\tilde{\mathbf{y}}$: the output of the NMT system when $\tilde{\mathbf{x}}$ is provided as input (tokens: $\tilde{y}_j$). This will be our pseudo-reference for fine-tuning or evaluation.

For the sake of readability, we use the terms grammatical errors, noise, or edits interchangeably. In the context of this work, they will all denote the annotated grammatical errors in the source sentences ($\mathbf{x}$).

Data There are many publicly available corpora for non-native English that are annotated with corrections, which have been widely used for the Grammar Error Correction tasks (Bryant et al., 2019). We specifically use NUCLE (Dahlmeier et al., 2013), FCE (Yannakoudakis et al., 2011), and Lang-8 (Tajiri et al., 2012) for creating the pseudo-references. For evaluation we use the JFLEG dataset (Napoles et al., 2017) and its accompanying Spanish translations in the JFLEG-ES dataset (Anastasopoulos et al., 2019).

The NUS Corpus of Learner English (**NUCLE**) contains essays written by Singaporean students. It is generally considered the main benchmark for GEC. This dataset consists of 21.3K sentences. The First Certificate in English corpus (**FCE**) is also made of essays, written by learners who were sitting the English as Second or Other Language (ESOL) examinations. We use the publicly available version, which includes 17.6K sentences.

Lang-8 is a slightly different dataset than the previous two datasets. This dataset was built from user-provided corrections in an online learner forum. In comparison to the others, this dataset is much larger, consisting of 149.5K sentences. However, this datasets' error domain is very versatile. It does not consist any test and validation set.

The JHU FLuency-Extended GUG corpus (**JFLEG**) is a small corpus of only 1.3K sentences,

intended only for evaluation. It has an unique character that is different from other datasets, as it contains correction annotations that include extended fluency edits rather than just minimal grammatical ones.

The JFLEG corpus was translated into Spanish by Anastasopoulos et al. (2019) to create the **JFLEG-ES** corpus, which provides gold-standard Spanish translations for every JFLEG sentence.

Evaluation In cases where we have access to human references, we can simply evaluate with reference-based metrics (e.g. BLEU (Papineni et al., 2002)). Unfortunately, we only have references for the JFLEG-ES dataset in Spanish.

For all other datasets and languages, we treat the translations of the corrected clean English sources as pseudo-references, and use the metrics from (Anastasopoulos, 2019): Robustness Score (RB), f-BLEU, f-METEOR, and Noise Ratio (NR).

Robustness Score (RB) is defined as the percentage of translations of noisy sentences that are *exactly the same* as the translation of the respective corrected sentence.

f-BLEU and **f-METEOR** are slight modification of the popular BLEU and METEOR metrics. The only difference is that they use pseudo-references instead of true human-created ones, and hence are referred to as faux BLEU and faux METEOR. In our case, the pseudo-references are the translations of the corrected sentence.

Target-Source Noise Ratio (NR) is the ratio between the target- and source-side BLEU score between noisy and corrected sentences. All other measures do not take into consideration how large are the source-side differences. The intuition behind this metric is that if there is minimal perturbation $d(\mathbf{x}, \tilde{\mathbf{x}})$ on the input side then there should be minimal reflection on the target side perturbation $d(\mathbf{y}, \tilde{\mathbf{y}})$ as well. NR is computed as:

$$\text{NR}(\mathbf{x}, \tilde{\mathbf{x}}, \mathbf{y}, \tilde{\mathbf{y}}) = \frac{d(\mathbf{y}, \tilde{\mathbf{y}})}{d(\mathbf{x}, \tilde{\mathbf{x}})} = \frac{100 - \text{BLEU}(\mathbf{y}, \tilde{\mathbf{y}})}{100 - \text{BLEU}(\mathbf{x}, \tilde{\mathbf{x}})}$$

4 Experiments and Results

We name our models in a way that is convenient to understand. Our models are named as such: dataset_language; e.g. the NUCLE_ES model will refer to the model fine-tuned on the NUCLE dataset for Spanish language. We will overload the naming convention to also refer to datasets in the same way, e.g. the NUCLE_ES dataset.

Experimental Details All data are tokenized and true-cased using the Moses tools (Koehn et al., 2007).We use the SentencePiece (Kudo and Richardson, 2018) toolkit to split the sentences into sub-words. We use the unigram language model algorithm of the toolkit with 65,000 operations. We filter the fine-tuning dataset so that sentence length is capped at 80 words.

Target Side Creation Given the recent success and promise of massively multilingual systems (Johnson et al., 2016; Firat et al., 2016), we use as our original model the OPUS-MT multilingual Romance model[5] (Tiedemann and Thottingal, 2020), trained using Marian NMT (Junczys-Dowmunt et al., 2018) within the HuggingFace's Transformers library (Wolf et al., 2019). For every dataset we pass source sentences (both original and corrected versions) and obtain target side sentences. Then we use the corrected target side sentences as our ground truth for fine-tuning the same model.

Transformer Model Details We use a transformer architecture as they have shown to be much superior to recurrent architectures. We use HuggingFace's Transformers' BartForConditionalGeneration as our model and tokenizer. This model uses 12 layers, 16 attention heads, the embedding dimension is 1024, and positional feed-forward dimension is 4096. Dropout is set to 0.1. We use the same learning rate schedule as in (Vaswani et al., 2017) with 500 warm-up steps but only decay the learning rate until it reaches $3 * 10^{-5}$. We fine-tune our models on a V100 GPU for a maximum of 100 epochs (although best validation set performance is reached around 20 to 25 epochs). For testing we use the model with the best performance on the validation dataset. Our validation check interval is set to 0.2.

Evaluation We use METEOR (Denkowski and Lavie, 2014) to calculate the f-METEOR scores. We calculate BLEU and f-BLEU scores using Sacrebleu (Post, 2018). We compute statistical significance with paired bootstrap resampling (Koehn, 2004).

Results on English-Spanish We first discuss the results on the JFLEG-ES test set, which is the only dataset with human gold references.

The performance of our systems on the JFLEG-ES test set, as measured by detokenized BLEU is

[5]name: Helsinki-NLP/opus-mt-en-ROMANCE

System	en→es BLEU		
	Clean†	Noisy	Δ
(Anastasopoulos et al., 2019)	27.80	26.80	-1.00
Helsinki-NLP/opus-mt-en-ROMANCE	30.14	28.04	-2.10
-fine-tuned on:			
NUCLE	27.89	26.62	-1.27
FCE	26.93	25.54	-1.39
Lang-8	28.85	28.20	-0.65
all noisy in Spanish	29.37	**31.14***	1.77
all noisy in all four languages	28.16	29.04	0.88
all clean in Spanish	**30.39**	27.83	-2.56

Table 1: Translation quality (BLEU scores) on the JFLEG_ES data-set. †: average over 4 corrected sentences as input to the translation model. *statistically significantly better than the baseline, with $p < 0.05$.

System	Sentence	BLEU
Source (original)	it has some problems that *it* can effect *to humens*.	
OPUS-MT output	tiene algunos problemas que *puede* afectar a los *humens*.	47
Finetuned output	tiene algunos problemas que pueden afectar a los humanos.	89
Reference	esto tiene algunos problemas que pueden afectar a los humanos.	
Source (original)	last month, I needed to buy *digtal-camera*.	
OPUS-MT output	el mes pasado, necesitaba comprar *digtal-camera*.	26
Finetuned output	el mes pasado, necesitaba comprar una cámara digital.	66
Reference	el mes pasado necesitaba comprar una cámara digital.	

Table 2: Examples (cherry-picked) of sentences with high BLEU improvement after fine-tuning (English-Spanish on the the JFLEG-ES dataset).

summarized in Table 1. The "Clean" column refers to an average BLEU score over the four versions of source-side corrections provided by the JFLEG dataset, the "Noisy" column reports results with the original sentences as input, and the last column presents the difference (Δ) between the two.

The first thing to note is that the multilingual OPUS-MT model outperforms the previously published results of Anastasopoulos et al. (2019) by more than 2 BLEU points on both clean and noisy settings. This is unsurprising, if one considers that the OPUS-MT model has been trained on an order of magnitude more English-Spanish data (about 25x), and it has in addition been trained on other related Romance languages. However, we should also note that the difference of the two models is imbalanced: the improvement from all these additional data is +2.3 BLEU points when evaluated on clean data, but only +1.2 BLEU points when evaluated on the noisy pairs. This outlines the importance of evaluating MT systems not only on clean data but also on other language variants. Although imbalanced these improvements are significant and hence we treat our multilingual OPUS-MT model as the baseline in all following discussions.

Fine-tuning on individual datasets yields inconsistent results, with the BLEU score changing from -2.5 to +0.16. The highest drop is observed when fine-tuning on FCE; this is reasonable as JFLEG and FCE include errors on quite different domains (Napoles et al., 2017). This ablation allows us to identify Lang-8 as perhaps the most appropriate single dataset for this kind of tasks, most likely due to its size and diversity of errors and domains.

Using all available datasets, however, is significantly better. We find that our model performs particularly well when fine-tuned on pseudo-references from all corpora (the "all noisy in Spanish" model (sixth row) that is a concatenation of all the datasets in Spanish). We observe a 3.1 BLEU improvement for noisy data, while suffering a small decrease (0.8 BLEU points) on clean data. The dif-

en→x	system	RB	f-BLEU noisy (clean)	f-METEOR	NR
Spanish	original	9.56	62.61 (99.1)	0.79	0.99
	adapted on es	24.45	69.88* (81.76)	**0.84**	0.69
	adapted on all	**24.76**	**69.92*** (75.81)	0.83	**0.68**
Italian	original	9.40	59.06 (99.32)	0.70	1.08
	adapted on it	**34.48**	**84.64*** (81.54)	**0.87**	**0.46**
	adapted on all	23.35	67.71* (75.14)	0.76	0.72
French	original	10.32	61.96 (99.16)	0.77	1.00
	adapted on fr	**43.02**	**87.30*** (83.12)	**0.90**	**0.31**
	adapted on all	22.22	67.80* (74.38)	0.78	0.73
Portuguese	original	9.20	61.12 (99.01)	0.72	1.02
	adapted on pt	**24.29**	68.61* (81.60)	**0.77**	0.70
	adapted on all	24.00	**69.36*** (76.47)	**0.77**	**0.69**

Table 3: Translation robustness evaluation (multiple metrics) for English to four Romance language translation. Adaptation significantly increases the Robustness percentage as well as f-BLEU. *statistically significantly better than the "original" baseline, with $p < 0.05$.

ferent datasets cover different types of errors and domains, and as a result the fine-tuning process does not get biased by a single type of domain.

The second-to-last row ("all noisy in all four languages") reports results when pseudo-references in all four experimental MT directions (EN to ES,FR,IT,PT) are used in the fine-tuning process of our multilingual model. In this case, we still observe 1 BLEU improvement over the noisy data, compared to the baseline, but the performance on clean data is further degraded.

Last, it was crucial to examine whether the improvements we obtained are due to our fine-tuned models becoming indeed more robust to errors, as opposed to adapting to the domain and other characteristics of the datasets we train and evaluate on. In the last row ("all clean in Spanish") we present the results following fine-tuning the models with the *corrected* sentences as inputs.[6] We confirm that indeed our model improves slightly on the clean data, but its performance does not improve on the noisy inputs. Hence, we can conclude that the effect of domain adaptation is minimal, and our fine-tuned model has indeed learn to deal with non-standard inputs.

Table 2 displays a couple of sentences where our fine-tuned system produces more fluent outputs that then pre-trained system, properly handling the source-side noise. The mistakes in the English source sentence and the MT outputs are highlighted with *red italics*. In the first example (top) our system can handle the typo "digtal" and correctly translate it as "digital." In the second example (bottom) our system, in addition to handling the typo "humens", it also correctly inflects the verb "pueden" (third person, plural) to agree with its subject, while the pretrained model produces an ungrammatical Spanish output (the verb "puede" is in third person singular and does not agree with its subject).

Results on other ROMANCE language In this section we report results obtained with the model fine-tuned on pseudo-references from all datasets for each of the four languages, as they were consistently better than any single-dataset fine-tuning approach. Table 3 presents the scores with all four evaluation metrics on all four En-to-X translation directions. For each language, we compare three models: the pre-trained one, one fine-tuned on pseudo-references for the respective language only, and one fine-tuned on all four languages simultaneously.

As we don't have human references for the other languages except Spanish, we use the Robustness Score, faux-BLEU, faux-METEOR, and Target-Source Noise Ratio metrics. As showcased by Table 3, in every language our approach yields a minimum of 7 f-BLEU points improvement over

[6]This would amount to a straightforward case of self-training, since it the target outputs were produced by the model itself prior to fine-tuning.

the original system when trained on that single language. In Italian and French, the improvement is particularly significant of at least 25 f-BLEU points. We have also listed the f-BLEU for clean test set which shows that the f-BLEU score decreased giving us assurance that the model is learning to be robust on noisy data. Also 75-80 BLEU score is still very significant.

The f-METEOR scores indicate similar trends. It is worth noting, though, that the differences between the original and the fine-tuned systems are less pronounced. We attribute this difference to the fact that most output differences are generally small local changes (e.g. on the inflection of a verb or a noun), which METEOR's paraphrase matching considers to be quite similar.

The Robustness Scores (RB) are also revealing: when the original system only returned the same output for the potentially noisy original sentence and the corrected one about 10% of the time, after fine-tuning all systems return the same outputs more than 24% of the time, reaching a RB score of more than 43% for French.

The Noise Ratio (NR) allows us to inspect if we actually manage to create a system that reduces the noise or not. An NR of less than 1 means that indeed our system reduces the source-side noise in its output, while a NR higher than one implies that the system amplifies the source-side differences (the lower NR the better). The pre-trained system consistently produces an NR of around 1, meaning that even though it does not amplify noise, it also does not reduce it. In comparison, our adapted models manage to reduce the source-side noise, with scores significantly lower than 1.

Can we achieve zero-shot robustness? An intriguing question that arose during our experiments, was whether one could fine-tune a multilingual system for robustness on only one language (e.g. Spanish) and consequently make the system more robust not only in that language but also in the other languages supported by the system. This avenue would significantly increase the value of not only our approach but also of the original multilingual systems: perhaps the community might eventually have access to large collections of true reference translations of non-native English, which would allow us to train systems robust to such source-side variations. Such datasets are unlikely to be available in multiple languages, though, hence the need for a way to improve a multilingual system's robustness using single-language data.

Finetune on:	en→es BLEU
Italian	5.04
French	10.02
Portuguese	3.86

Table 4: Simple finetuning on only a single language leads to catastrophic forgetting of the other languages, as the low translation quality (BLEU scores) on the JFLEG_ES data-set show.

We attempt a first step towards this direction, by evaluating on English-Spanish the systems that we fine-tuned solely on English-Italian, English-French, and English-Portuguese. Unfortunately, as outlined in Table 4, this simple approach does not work out-of-the-box. Fine-tuning on a single language pair leads to catastrophic forgetting (French, 1999) of the multilingual abilities of the system. This is a phenomenon commonly observed in continued learning or fine-tuning scenarios (Goodfellow et al., 2013) as well as on MT domain adaptation scenarios in particular (Freitag and Al-Onaizan, 2016), for the mitigation of which several approaches have been proposed (Lopez-Paz and Ranzato, 2017; Thompson et al., 2019; Michel et al., 2019b, et alia). As this research direction is beyond the scope of this paper, we leave the application of such approaches for future work.

5 Conclusion

In this work, we studied the effect of fine-tuning a NMT model using *real* source-side noise paired with pseudo-references obtained by translating Grammar Error Correction corpora. We confirmed previous works on the utility of training with source-side noise, as it leads to models more robust to non-native English inputs, but also showed that instead of using synthetically-induced noise, we can (a) use real-user data with pseudo-references and (b) fine-tune a pre-trained system, rather than training from scratch. We will release all pseudo-references and our code upon acceptance. Our approach of fine-tuning a pre-trained system with pseudo-references approach has particular appealing advantages (less training time, no need for costly translation references) and it improves the robustness of MT systems significantly on all language pairs we tested.

For future work, we will explore ways to integrate strategies for avoiding catastrophic forgetting,

in order to achieve *multilingual* robustness without needing to fine-tune a multilingual model on all interested languages, as well as incorporating robustness rewards through reinforcement learning in the fine-tuning process. In addition, we will investigate how the quality of the pseudo-references affects the downstream results, and we also plan to explore the trade-off between language-specific and multilingual fine-tuning.

Acknowledgements

The authors want to thank the reviewers for their insightful comments. The first author was funded by a George Mason Computer Science Department's 2020 PhD Research Initiation Award.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. 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 3874–3884, Minneapolis, Minnesota. 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.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. Cite arxiv:1409.0473Comment: Accepted at ICLR 2015 as oral presentation.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *Proc. ICLR*.

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

Christopher Bryant, Mariano Felice, Øistein E. Andersen, and Ted Briscoe. 2019. The BEA-2019 shared task on grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–75, Florence, Italy. Association for Computational Linguistics.

Qiao Cheng, Meiyuan Fang, Yaqian Han, Jin Huang, and Yitao Duan. 2019a. Breaking the data barrier: Towards robust speech translation via adversarial stability training.

Yong Cheng, Lu Jiang, and Wolfgang Macherey. 2019b. Robust neural machine translation with doubly adversarial inputs. *CoRR*, abs/1906.02443.

Yong Cheng, Zhaopeng Tu, Fandong Meng, Junjie Zhai, and Yang Liu. 2018. Towards robust neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1756–1766, Melbourne, Australia. Association for Computational Linguistics.

Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. 2013. Building a large annotated corpus of learner English: The NUS corpus of learner English. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 22–31, Atlanta, Georgia. Association for Computational Linguistics.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA. Association for Computational Linguistics.

Javid Ebrahimi, Daniel Lowd, and Dejing Dou. 2018. On adversarial examples for character-level neural machine translation. *CoRR*, abs/1806.09030.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 866–875, San Diego, California. Association for Computational Linguistics.

Markus Freitag and Yaser Al-Onaizan. 2016. Fast domain adaptation for neural machine translation. arXiv:1612.06897.

Robert M French. 1999. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135.

Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. 2013. An empirical investigation of catastrophic forgetting in gradient-based neural networks.

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 Papers)*, pages 68–80, Boston, MA. Association for Machine Translation in the Americas.

Jindřich Helcl, Jindřich Libovický, and Martin Popel. 2019. CUNI system for the WMT19 robustness task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 539–543, Florence, Italy. Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda B. Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's multilingual neural machine translation system: Enabling zero-shot translation. *CoRR*, abs/1611.04558.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, et al. 2018. Marian: Fast neural machine translation in c++. arXiv:1804.00344.

Vladimir Karpukhin, Omer Levy, Jacob Eisenstein, and Marjan Ghazvininejad. 2019. Training on synthetic noise improves robustness to natural noise in machine translation. *CoRR*, abs/1902.01509.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. 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.

Hairong Liu, Mingbo Ma, Liang Huang, Hao Xiong, and Zhongjun He. 2018. Robust neural machine translation with joint textual and phonetic embedding. *CoRR*, abs/1810.06729.

David Lopez-Paz and Marc'Aurelio Ranzato. 2017. Gradient episodic memory for continual learning. In *Advances in neural information processing systems*, pages 6467–6476.

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. *CoRR*, abs/1508.04025.

Paul Michel, Xian Li, Graham Neubig, and Juan Pino. 2019a. On evaluation of adversarial perturbations for sequence-to-sequence models. 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 3103–3114, Minneapolis, Minnesota. Association for Computational Linguistics.

Paul Michel and Graham Neubig. 2018. MTNT: A testbed for machine translation of noisy text. *CoRR*, abs/1809.00388.

Paul Michel, Elisabeth Salesky, and Graham Neubig. 2019b. Regularizing trajectories to mitigate catastrophic forgetting. Preprint.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2017. JFLEG: A fluency corpus and benchmark for grammatical error correction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 229–234, Valencia, Spain. 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.

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.

Matt Post and Kevin Duh. 2019. JHU 2019 robustness task system description. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 552–558, Florence, Italy. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Matthias Sperber, Jan Niehues, and Alex Waibel. 2017. Toward robust neural machine translation for noisy input sequences. In *Proc. IWSLT*.

Toshikazu Tajiri, Mamoru Komachi, and Yuji Matsumoto. 2012. Tense and aspect error correction for ESL learners using global context. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 198–202, Jeju Island, Korea. Association for Computational Linguistics.

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.

Brian Thompson, Jeremy Gwinnup, Huda Khayrallah, Kevin Duh, and Philipp Koehn. 2019. Overcoming catastrophic forgetting during domain adaptation of neural machine translation. 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 2062–2068, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Ashish Vaswani, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. 2018. Tensor2Tensor for neural machine translation. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*, pages 193–199, Boston, MA. Association for Machine Translation in the Americas.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. arXiv:1910.03771.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189, Portland, Oregon, USA. Association for Computational Linguistics.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

Renjie Zheng, Hairong Liu, Mingbo Ma, Baigong Zheng, and Liang Huang. 2019. Robust machine translation with domain sensitive pseudo-sources: Baidu-OSU WMT19 MT robustness shared task system report. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 559–564, Florence, Italy. Association for Computational Linguistics.

Shuyan Zhou, Xiangkai Zeng, Yingqi Zhou, Antonios Anastasopoulos, and Graham Neubig. 2019. Improving robustness of neural machine translation with multi-task learning. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 565–571, Florence, Italy. Association for Computational Linguistics.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy. Association for Computational Linguistics.