Evaluating Robustness to Input Perturbations for Neural Machine Translation

Xing Niu, Prashant Mathur, Georgiana Dinu, Yaser Al-Onaizan
Amazon AI
{xingniu,pramathu,gddinu,onaizan}@amazon.com

Abstract

Neural Machine Translation (NMT) models are sensitive to small perturbations in the input. Robustness to such perturbations is typically measured using translation quality metrics such as BLEU on the noisy input. This paper proposes additional metrics which measure the relative degradation and changes in translation when small perturbations are added to the input. We focus on a class of models employing subword regularization to address robustness and perform extensive evaluations of these models using the robustness measures proposed. Results show that our proposed metrics reveal a clear trend of improved robustness to perturbations when subword regularization methods are used.

1 Introduction

Recent work has pointed out the challenges in building robust neural network models (Goodfellow et al., 2015; Papernot et al., 2016). For Neural Machine Translation (NMT) in particular, it has been shown that NMT models are brittle to small perturbations in the input, both when these perturbations are synthetically created or generated to mimic real data noise (Belinkov and Bisk, 2018). Consider the example in Table 1 where an NMT model generates a worse translation as a consequence of only one character changing in the input.

Improving robustness in NMT has received a lot of attention lately with data augmentation (Sperber et al., 2017; Belinkov and Bisk, 2018; Vaibhav et al., 2019; Liu et al., 2019; Karpukhin et al., 2019) and adversarial training methods (Cheng et al., 2018; Ebrahimi et al., 2018; Cheng et al., 2019; Michel et al., 2019) as some of the more popular approaches used to increase robustness in neural network models.

In this paper, we focus on one class of methods, subword regularization, which addresses NMT robustness without introducing any changes to the architectures or to the training regime, solely through dynamic segmentation of input into subwords (Kudo, 2018; Provilkov et al., 2019). We provide a comprehensive comparison of these methods on several language pairs and under different noise conditions on robustness-focused metrics.

Original input	Se kyllä tuntuu sangen luultavalta.
Translation	It certainly seems very likely.
Perturbed input	Se kyllä tumtuu sangen luultavalta.
Translation	It will probably darken quite probably.
Reference	It certainly seems probable.

Table 1: An example of NMT English translations for a Finish input and its one-letter misspelled version.

Previous work has used translation quality measures such as BLEU on noisy input as an indicator of robustness. Absolute model performance on noisy input is important, and we believe this is an appropriate measure for noisy domain evaluation (Michel and Neubig, 2018; Berard et al., 2019; Li et al., 2019). However, it does not disentangle model quality from the relative degradation under added noise.

For this reason, we propose two additional measures for robustness which quantify the changes in translation when perturbations are added to the input. The first one measures relative changes in translation quality while the second one focuses on consistency in translation output irrespective of reference translations. Unlike the use of BLEU scores alone, the metrics introduced show clearer trends across all languages tested: NMT models are more robust to perturbations when subword regularization is employed. We also show that for the models used, changes in output strongly correlate with decreased quality and the consistency measure alone can be used as a robustness proxy in the absence of reference data.

2 Evaluation Metrics

Robustness is usually measured with respect to translation quality. Suppose an NMT model M translates input x to y' and translates its perturbed version x_δ to y'_δ, the translation quality (TQ) on these datasets is measured against reference translations y: $\texttt{TQ}(y', y)$ and $\texttt{TQ}(y'_\delta, y)$. TQ can be implemented as any quality measurement metric, such as BLEU (Papineni et al., 2002) or 1 minus TER (Snover et al., 2006).

Previous work has used TQ on perturbed or noisy input as an indicator of robustness. However, we argue that assessing models' performance relative to that of the original dataset is important as well in order to capture models' sensitivity to perturbations. Consider the following hypothetical example:

$$M_1\text{: } \texttt{BLEU}(y'_1, y) = 40, \texttt{BLEU}(y'_{\delta 1}, y) = 38;$$
$$M_2\text{: } \texttt{BLEU}(y'_2, y) = 37, \texttt{BLEU}(y'_{\delta 2}, y) = 37.$$

Selecting M_1 to translate noisy data alone is preferable, since M_1 outperforms M_2 ($38 > 37$). However, M_1's quality degradation ($40 \rightarrow 38$) reflects that it is in fact more sensitive to perturbation δ comparing with M_2.

To this end, we use the ratio between $\texttt{TQ}(y', y)$ and $\texttt{TQ}(y'_\delta, y)$ to quantify an NMT model M's invariance to specific data and perturbation, and define it as **robustness**:

$$\texttt{ROBUST}(M|x, y, \delta) = \frac{\texttt{TQ}(y'_\delta, y)}{\texttt{TQ}(y', y)}.$$

When evaluating on the dataset (x, y), $\texttt{ROBUST}(M|x, y, \delta) < 1$ means the translation quality of M is degraded under perturbation δ; $\texttt{ROBUST}(M|x, y, \delta) = 1$ indicates that M is robust to perturbation δ.

It is worth noting that: (1) ROBUST can be viewed as the normalized $\Delta\texttt{TQ} = \texttt{TQ}(y', y) - \texttt{TQ}(y'_\delta, y)$ because $\Delta\texttt{TQ}/\texttt{TQ}(y', y) = 1 - \texttt{ROBUST}$. We opt for the ratio definition because it is on a $[0, 1]$ scale, and it is easier to interpret than $\Delta\texttt{TQ}$ since the latter needs to be interpreted in the context of the TQ score. (2) High robustness can only be expected under low levels of noise, as it is not realistic for a model to recover from extreme perturbations.

Evaluation without References Reference translations are not readily available in some cases, such as when evaluating on a new domain. Inspired by unsupervised consistency training (Xie et al., 2019), we test if translation *consistency* can be used to estimate robustness against noise perturbations. Specifically, a model is consistent under a perturbation δ if the two translations, y'_δ and y' are similar to each other. Note that consistency is sufficient but not necessary for robustness: a good translation can be expressed in diverse ways, which leads to high robustness but low consistency.

We define **consistency** by

$$\texttt{CONSIS}(M|x, \delta) = \texttt{Sim}(y'_\delta, y').$$

Sim can be any *symmetric* measure of similarity, and in this paper we opt for $\texttt{Sim}(y'_\delta, y')$ to be the harmonic mean of $\texttt{TQ}(y'_\delta, y')$ and $\texttt{TQ}(y', y'_\delta)$, where TQ is BLEU between two outputs.

3 Experimental Set-Up

We run several experiments across different language families with varying difficulties, across different training data conditions (i.e. with different training data sizes) and evaluate how different subword segmentation strategies performs across noisy domains and noise types.

Implementation Details We build NMT models with the Transformer-base architecture (Vaswani et al., 2017) implemented in the Sockeye toolkit (Hieber et al., 2017). The target embeddings and the output layer's weight matrix are tied (Press and Wolf, 2017). Training is done on 2 GPUs, with a batch size of 3072 tokens and we checkpoint the model every 4000 updates. The learning rate is initialized to 0.0002 and reduced by 10% after 4 checkpoints without improvement of perplexity on the development set. Training stops after 10 checkpoints without improvement.

Tasks and Data We train NMT models on eight translation directions and measure robustness and consistency for them. EN↔DE and EN↔FI models are trained with pre-processed WMT18 news data and tested with the latest news test sets (newstest2019).

Recently, two datasets were built from user-generated content, MTNT (Michel and Neubig, 2018) and 4SQ (Berard et al., 2019). They provide naturally occurring noisy inputs and translations for EN↔FR and EN↔JA, thus enabling automatic evaluations. EN↔JA baseline models are trained and also tested with aggregated data provided by MTNT, i.e., KFTT+TED+JESC (KTJ). EN↔FR

	Languages	# sentences	# EN tokens
BASE	EN↔DE	29.3 M	591 M
	EN↔FR	22.2 M	437 M
	EN↔FI	2.9 M	71 M
	EN↔JA	3.9 M	43 M
MTNT	EN→FR	36.1 K	1,011 K
	FR→EN	19.2 K	779 K
	EN→JA	5.8 K	338 K
	JA→EN	6.5 K	156 K
4SQ	FR→EN	12.1 K	141 K

Table 2: Statistics of various training data sets.

baseline models are trained with aggregated data of Europarl-v7 (Koehn, 2005), NewsCommentary-v14 (Bojar et al., 2018), OpenSubtitles-v2018 (Lison and Tiedemann, 2016), and ParaCrawl-v5[1], which simulates the UGC training corpus used in 4SQ benchmarks, and they are tested with the latest WMT new test sets supporting EN↔FR (newstest2014).

Following the convention, we also evaluate models directly on noisy MTNT (mtnt2019) and 4SQ test sets. We fine-tune baseline models with corresponding MTNT/4SQ training data, inheriting all hyper-parameters except the checkpoint interval which is re-set to 100 updates. Table 2 shows itemized training data statistics after pre-processing.

Perturbations We investigate two frequently used types of perturbations and apply them to WMT and KTJ test data. The first is synthetic misspelling: each word is misspelled with probability of 0.1, and the strategy is randomly chosen from single-character deletion, insertion, and substitution (Karpukhin et al., 2019). The second perturbation is letter case changing: each sentence is modified with probability of 0.5, and the strategy is randomly chosen from upper-casing all letters, lower-casing all letters, and title-casing all words (Berard et al., 2019).[2]

Since we change the letter case in the test data, we always report case-insensitive BLEU with '13a' tokenization using sacreBLEU (Post, 2018). Japanese output is pre-segmented with Kytea before running sacreBLEU.[3]

[1] https://paracrawl.eu/

[2] Character substitution uses neighbor letters on the QWERTY keyboard, so accented characters are not substituted. Japanese is "misspelled" for each character with probability of 0.1, and it only supports deletion and repetition. Letter case changing does not apply to Japanese.

[3] http://www.phontron.com/kytea/

Model Variations We focus on comparing different (stochastic) subword segmentation strategies: BPE (Sennrich et al., 2016), BPE-Dropout (Provilkov et al., 2019), and SentencePiece (Kudo, 2018). Subword regularization methods (i.e., BPE-Dropout and SentencePiece) generate various segmentations for the same word, so the resulting NMT model better learns the meaning of less frequent subwords and should be more robust to noise that yields unusual subword combinations, such as misspelling. We use them only in offline training data pre-processing steps, which requires no modification to the NMT model.[4]

4 Experimental Results

As shown in Table 3, there is no clear winner among the three subword segmentation models based on BLEU scores on original WMT or KTJ test sets. This observation is different from results reported by Kudo (2018) and Provilkov et al. (2019). One major difference from previous work is the size of the training data, which is much larger in our experiments – subword regularization is presumably preferable on low-resource settings.

However, both our proposed metrics (i.e., robustness and consistency) show clear trends of models' robustness to input perturbations across all languages we tested: BPE-Dropout > SentencePiece > BPE. This suggests that although we did not observe a significant impact of subword regularization on generic translation quality, the robustness of the models is indeed improved drastically.

Unfortunately, it is unclear if subword regularization can help translating real-world noisy input, as shown in Table 4. MTNT and 4SQ contain several natural noise types such as grammar errors, emojis, with misspelling as the dominating noise type for English and French. The training data we use may already cover common natural misspellings, perhaps contributing to the failure of regularization methods to improve over BPE in this case.

Robustness Versus Consistency Variation in output is not necessarily in itself a marker of reduced translation quality, but empirically, consistency and robustness nearly always provide same model rankings in Table 3. We conduct more comprehensive analysis on the correlation between them, and we collect additional data points by varying the noise level of both perturbations. Specif-

[4] We sample one subword segmentation for each source sequence with SentencePiece.

	Model	BLEU	ROBUST	CONSIS	BLEU	ROBUST	CONSIS
		EN→DE (newstest2019)			DE→EN (newstest2019)		
original	BPE	39.70±0.71	–	–	40.01±0.65	–	–
	BPE-Dropout	39.65±0.73	–	–	40.16±0.66	–	–
	SentencePiece	39.85±0.75	–	–	40.25±0.67	–	–
+ misspelling	BPE	29.38±0.60	74.01±0.95	60.59±0.80	33.48±0.61	83.69±0.96	71.51±0.74
	BPE-Dropout	33.13±0.70	83.55±0.92	70.74±0.77	35.97±0.64	89.58±0.78	78.33±0.64
	SentencePiece	31.87±0.66	79.99±0.97	66.40±0.76	35.26±0.66	87.61±0.91	74.09±0.74
+ case-changing	BPE	31.61±0.74	79.63±1.31	73.26±1.19	33.72±0.69	84.27±1.15	73.19±1.13
	BPE-Dropout	35.04±0.73	88.37±0.97	80.04±0.99	36.34±0.69	90.48±0.95	78.96±0.96
	SentencePiece	33.49±0.73	84.05±1.09	76.24±1.09	34.48±0.71	85.65±1.10	74.55±1.10
		EN→FR (newstest2014)			FR→EN (newstest2014)		
original	BPE	41.47±0.48	–	–	39.24±0.50	–	–
	BPE-Dropout	40.72±0.48	–	–	39.22±0.50	–	–
	SentencePiece	41.05±0.48	–	–	39.14±0.50	–	–
+ misspelling	BPE	34.01±0.45	82.01±0.66	71.59±0.53	32.62±0.48	83.13±0.63	73.05±0.49
	BPE-Dropout	35.98±0.46	88.36±0.59	78.49±0.48	34.71±0.48	88.51±0.60	79.27±0.50
	SentencePiece	34.78±0.45	84.72±0.59	75.28±0.51	33.44±0.48	85.43±0.62	75.28±0.50
+ case-changing	BPE	34.75±0.54	83.81±0.97	79.34±0.93	32.31±0.54	82.34±0.96	76.56±0.95
	BPE-Dropout	38.28±0.47	94.00±0.55	86.28±0.58	35.78±0.50	91.24±0.65	84.47±0.65
	SentencePiece	36.49±0.50	88.87±0.74	82.73±0.76	33.51±0.54	85.61±0.84	78.18±0.88
		EN→FI (newstest2019)			FI→EN (newstest2019)		
original	BPE	20.43±0.55	–	–	24.31±0.59	–	–
	BPE-Dropout	20.01±0.54	–	–	24.51±0.57	–	–
	SentencePiece	20.63±0.57	–	–	24.67±0.60	–	–
+ misspelling	BPE	15.20±0.46	74.42±1.39	52.76±0.89	21.27±0.54	87.47±1.14	70.06±0.89
	BPE-Dropout	17.39±0.50	86.95±1.43	63.63±0.86	22.40±0.55	91.38±1.06	75.18±0.83
	SentencePiece	16.73±0.51	81.09±1.52	57.45±0.85	21.89±0.57	88.76±1.19	70.57±0.87
+ case-changing	BPE	15.65±0.53	76.63±1.71	68.27±1.44	20.71±0.58	85.20±1.32	74.85±1.16
	BPE-Dropout	17.19±0.53	85.92±1.39	72.76±1.30	23.10±0.58	94.26±1.09	79.67±1.00
	SentencePiece	15.72±0.54	76.19±1.72	67.73±1.40	21.50±0.58	87.16±1.26	76.29±1.12
		EN→JA (KTJ)			JA→EN (KTJ)		
original	BPE	24.28±0.53	–	–	22.80±0.51	–	–
	BPE-Dropout	24.11±0.51	–	–	22.21±0.52	–	–
	SentencePiece	22.63±0.45	–	–	22.99±0.50	–	–
+ misspelling	BPE	19.82±0.47	81.66±1.09	54.84±0.73	18.20±0.45	79.83±1.20	52.34±0.74
	BPE-Dropout	22.01±0.49	91.30±0.95	63.21±0.78	18.89±0.47	85.06±1.17	56.43±0.78
	SentencePiece	19.85±0.41	87.69±1.05	61.25±0.80	18.97±0.46	82.53±1.15	56.40±0.73
+ case-changing	BPE	20.35±0.51	83.83±1.13	68.10±1.25	–	–	–
	BPE-Dropout	21.44±0.49	88.91±1.00	72.96±1.13	–	–	–
	SentencePiece	19.99±0.44	88.32±1.06	73.52±1.10	–	–	–

Table 3: BLEU, robustness (in percentage), and consistency scores of different subword segmentation methods on original and perturbed test sets. We report mean and standard deviation using bootstrap resampling (Koehn, 2004). Subword regularization makes NMT models more robust to input perturbations.

		MTNT (mtnt2019)				4SQ
	Model	EN→JA	JA→EN	EN→FR	FR→EN	FR→EN
baseline	BPE	10.75±0.49	9.68±0.59	34.15±0.93	45.84±0.89	30.96±0.85
	BPE-Dropout	10.76±0.47	9.26±0.64	33.39±0.95	45.84±0.90	31.28±0.84
	SentencePiece	10.52±0.51	9.52±0.68	33.75±0.91	45.94±0.92	31.44±0.85
fine-tuning	BPE	14.88±0.52	10.47±0.69	35.11±0.95	46.49±0.90	34.83±0.86
	BPE-Dropout	15.26±0.53	11.13±0.68	34.80±0.93	46.88±0.88	34.72±0.84
	SentencePiece	14.68±0.53	11.19±0.72	34.71±0.93	46.89±0.90	34.59±0.86

Table 4: BLEU scores of using different subword segmentation methods on two datasets with natural noise. Subword regularization methods do not achieve consistent improvement over BPE, nor with or without fine-tuning.

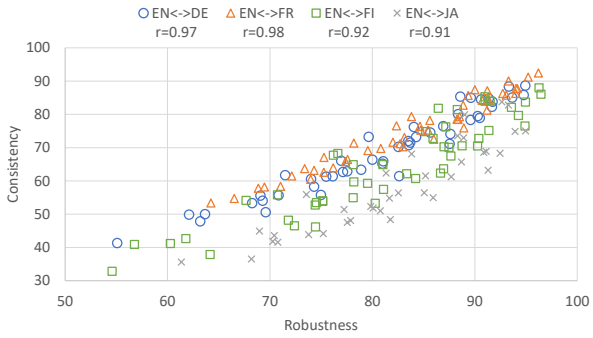

Figure 1: Robustness (in percentage) and consistency are highly correlated within each language pair. Correlation coefficients are marked in the legend.

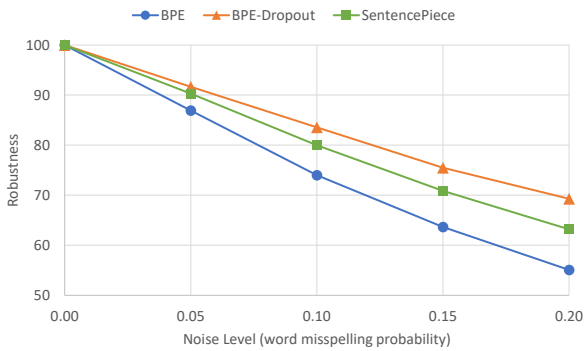

Figure 2: Varying the synthetic word misspelling probability for EN→DE models does not change the model ranking w.r.t. robustness (in percentage).

ically, we use the following word misspelling probabilities: $\{0.05, 0.1, 0.15, 0.2\}$ and the following sentence case-changing probability values: $\{0.3, 0.5, 0.7, 0.9\}$.

As illustrated in Figure 1, consistency strongly correlates with robustness (sample Pearson's $r = 0.91$ to 0.98) within each language pair. This suggests that for this class of models, low consistency signals a drop in translation quality and the consistency score can be used as a robustness proxy when the reference translation is unavailable.

Robustness Versus Noise Level In this paper, robustness is defined by giving a fixed perturbation function and its noise level. We observe consistent model rankings across language pairs, but is it still true if we vary the noise level?

To test this, we plot the robustness data points from the last section against the noise level. Focusing on the misspelling perturbation for EN→DE models, Figure 2 shows that varying the word misspelling probability does not change the ranking of the models, and the gap in the robustness measurement only increases with larger amount of noise. This observation applies to all perturbations and language pairs we investigated.

5 Conclusion

We proposed two additional measures for NMT robustness which can be applied when both original and noisy inputs are available. These measure robustness as relative degradation in quality as well as consistency which quantifies variation in translation output irrespective of reference translations. We also tested two popular subword regularization techniques and their effect on overall performance and robustness. Our robustness metrics reveal a clear trend of subword regularization being much more robust to input perturbations than standard BPE. Furthermore, we identify a strong correlation between robustness and consistency in these models indicating that consistency can be used to estimate robustness on data sets or domains lacking reference translations.

6 Acknowledgements

We thank the anonymous reviewers for their comments and suggestions.

References

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Alexandre Berard, Ioan Calapodescu, Marc Dymetman, Claude Roux, Jean-Luc Meunier, and Vassilina Nikoulina. 2019. Machine translation of restaurant reviews: New corpus for domain adaptation and robustness. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 168–176, Hong Kong. Association for Computational Linguistics.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Yong Cheng, Lu Jiang, and Wolfgang Macherey. 2019. Robust neural machine translation with doubly adversarial inputs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4324–4333, Florence, Italy. Association for Computational Linguistics.

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.

Javid Ebrahimi, Daniel Lowd, and Dejing Dou. 2018. On adversarial examples for character-level neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 653–663, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*.

Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2017. Sockeye: A toolkit for neural machine translation. *CoRR*, abs/1712.05690.

Vladimir Karpukhin, Omer Levy, Jacob Eisenstein, and Marjan Ghazvininejad. 2019. Training on synthetic noise improves robustness to natural noise in machine translation. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 42–47, Hong Kong, China. Association for Computational Linguistics.

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. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the Tenth Machine Translation Summit*, volume 5, pages 79–86. Citeseer.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. 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.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Hairong Liu, Mingbo Ma, Liang Huang, Hao Xiong, and Zhongjun He. 2019. Robust neural machine translation with joint textual and phonetic embedding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3044–3049, Florence, Italy. Association for Computational Linguistics.

Paul Michel, Xian Li, Graham Neubig, and Juan Pino. 2019. 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. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 543–553, Brussels, Belgium. Association for Computational Linguistics.

Nicolas Papernot, Patrick D. McDaniel, Ian J. Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. 2016. Practical black-box attacks against deep learning systems using adversarial examples. *CoRR*, abs/1602.02697.

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.

Ofir Press and Lior Wolf. 2017. Using the output embedding to improve language models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 157–163, Valencia, Spain. Association for Computational Linguistics.

Ivan Provilkov, Dmitrii Emelianenko, and Elena Voita. 2019. Bpe-dropout: Simple and effective subword regularization. *CoRR*, abs/1910.13267.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers), pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, volume 200.

Matthias Sperber, Jan Niehues, and Alex Waibel. 2017. Toward robust neural machine translation for noisy input sequences. In *Proceedings of the 14th International Workshop on Spoken Language Translation*, pages 1715–1725, Tokyo, Japan.

Vaibhav Vaibhav, Sumeet Singh, Craig Stewart, and Graham Neubig. 2019. Improving robustness of machine translation with synthetic noise. 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 1916–1920, Minneapolis, Minnesota. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008.

Qizhe Xie, Zihang Dai, Eduard H. Hovy, Minh-Thang Luong, and Quoc V. Le. 2019. Unsupervised data augmentation for consistency training. *CoRR*, abs/1904.12848.