Human Language Technologies – The Baltic Perspective
A. Utka et al. (Eds.)

doi:10.3233/FAIA200606

Robust Neural Machine Translation: Modeling Orthographic and Interpunctual Variation

Toms BERGMANIS [a,b,1], Artūrs STAFANOVIČS [a], Mārcis PINNIS [a,b]
[a] *Tilde, Riga, Latvia*
[b] *Faculty of Computing, University of Latvia, Latvia*

Abstract. Neural machine translation systems typically are trained on curated corpora and break when faced with non-standard orthography or punctuation. Resilience to spelling mistakes and typos, however, is crucial as machine translation systems are used to translate texts of informal origins, such as chat conversations, social media posts and web pages. We propose a simple generative noise model to generate adversarial examples of ten different types. We use these to augment machine translation systems' training data and show that, when tested on noisy data, systems trained using adversarial examples perform almost as well as when translating clean data, while baseline systems' performance drops by 2-3 BLEU points. To measure the robustness and noise invariance of machine translation systems' outputs, we use the average translation edit rate between the translation of the original sentence and its noised variants. Using this measure, we show that systems trained on adversarial examples on average yield 50 % consistency improvements when compared to baselines trained on clean data.

Keywords. Neural machine translation, robustness, noisy data

1. Introduction

Humans exhibit resilience to orthographic variation in written text [1,2]. As a result, spelling mistakes and typos are often left unnoticed. This flexibility of ours, however, is shown to be detrimental for neural machine translation (NMT) systems, which typically are trained on curated corpora and tend to break when faced with noisy data [3,4]. Achieving NMT robustness to human blunder, however, is important when translating texts of less formal origins, such as chat conversations, social media posts and web pages with comment sections.

In this work, we propose to augment NMT system's training data with data where source sentences are corrupted with adversarial examples of different types. There have been various studies on the impact of different types and sources of noise on NMT [5,6,7]. In this work, we focus on the noise caused by orthographic variation of words, such as unintentional misspellings and deliberate spelling alternations as well as noise due to misplaced and omitted punctuation. Thus, the closest to this study is the work on

[1]Corresponding Author: Toms Bergmanis; E-mail: toms.bergmanis@tilde.lv.

black-box adversarial training of NMT systems [3,8,9], where models are trained on adversarial examples that are generated without accessing the model's parameters. Unlike the previous work, which focuses only on adversarial examples that model unintentional changes of spelling, we also model deliberate orthographic alternation, such as omission and substitution of diacritical signs. As we show in our experiments, such orthographic variation has a more substantial negative impact on MT outputs than other types of noise and thus is more important to be accounted for. Further, to overcome the lack of curated evaluation datasets as required by the previous work [4,9], we propose an automatic evaluation method that measures the noise invariance of MT outputs without relying on a reference translation. By measuring noise invariance of MT outputs, the method also allows us to assess whether MT system translation consistency improves when facing small variations in the source text.

Table 1. Noise applied to the example sentence: "*Balta jūra, zaļa zeme.*" Were possible, noise is marked in bold, otherwise it is indicated with '_'

#	Type	Examples
1	introduce extra letters	Bal**z**ta jūra, zaļa zeme.
2	delete letters	_alta jūra, zaļa zeme.
3	permute letters	Ba**tl**a jūra, zaļa zeme.
4	confuse letters	Balta jūra, **x**aļa zeme.
5	add diacritic	Balta jūra, zaļa z**ē**me.
6	sample substitute	Balta jūra, zaļa **zemi**.
7	remove punctuation	Balta jūra_ zaļa zeme_
8	add comma	Balta**,** jūra, zaļa zeme.
9	latinize	Balta j**u**ra, za**l**a zeme.
10	phonetic latinize	Balta j**uu**ra, za**lj**a zeme.

2. Methods

We propose a simple generative noise model to generate adversarial examples of ten different types. These include incidental insertion, deletion, permutation and keyboard-based confusion of letters as well as the addition of a diacritic to letters which support them (Table 1, examples 1-5). We also explicitly model the misspellings that result in another valid word (Table 1, example 6). For interpunctual variation, we consider sentences with missing punctuation and incorrectly placed commas (Table 1, examples 7-8). For deliberate orthographic changes, we support sentence-level omission and phonetic latinization of diacritical signs (Table 1, examples 9-10).

Measure of Robustness. To measure NMT robustness and noise invariance of NMT outputs, we calculate the average translation edit rate (TER) [10] between the translation of the original orthographically correct sentence and the translations of its ten noised variants for each noise type. We refer to it as tenfold noisy translation TER, or **10NT-TER**. This measure gives a score of 0 if all ten translations of a sentence with added noise match the translation of the original sentence and a score of 100 (or more) if all of them had no word in common with the translation of the original sentence.

Table 2. The original training data sizes and data sizes with adversarial examples included

		Train	
		Original	+ adversarial noise
Small data	English-Latvian	4.5M	9M
Large data	English-Estonian	34.9M	69.8M
	English-Latvian	45.2M	90.4M
	English-Lithuanian	22.1M	44.2M

3. Experimental Setting

Languages and Data. We conduct experiments on Estonian-English, Latvian-English and Lithuania-English language pairs. We use the Latvian-English constrained data from the WMT 2017[2] news translation shared task to train **small data systems** that we use for development and analysis of our methods. To verify that our findings also hold not only for small data settings, but also for production-grade systems that are trained on much larger data, we use large datasets from the Tilde Data Library[3] to train **large data systems**. For the validation during training and testing, we use development and test sets from the WMT news translation shared tasks. For English-Estonian, we use the data from WMT 2018, for English-Latvian – WMT 2017, and for English-Lithuanian – WMT 2019[4].

We use a simplified and production-grade data pre-processing pipelines. The simplified data pre-processing consists of the standard Moses [11] scripts for tokenization, cleaning, normalization, and truecasing, while the production grade pipeline consists of Tilde MT platform's [12] implementation of the same processes.

NMT Models. We mostly use the default configuration[5] of the Marian [13] toolkit's implementation of the Transformer model [14]. We select batch sizes dynamically so that they fit in a workspace of 9000MB. Additionally, we use delayed gradient updates [15] by setting optimizer delay to 4. We stop model training after ten consecutive evaluations with no improvement in translation quality on the development set [16].

4. Experiments

Initial Experiments. To test the effect of individual noise models on MT systems' performance, we train separate Latvian-English small data systems on original data augmented in a 1-to-1 proportion with each type of adversarial examples. All in all, we obtain ten systems trained using adversarial examples and the baseline. We test each system on the original development set and development sets that have adversarial examples of each type of noise. Table 3 summarises the results. First, we note that including adversarial examples improves the overall translation quality and especially quality on development sets containing the adversarial examples that the systems have seen during training.

[2] http://www.statmt.org/wmt17
[3] https://www.tilde.com/products-and-services/data-library
[4] http://www.statmt.org/wmt17|18|19
[5] https://github.com/marian-nmt/marian-examples/tree/master/transformer

Table 3. Latvian-English development set results in BLEU [17] points for small data systems. Rows: systems trained on original data that are 1:1 up-sampled with each type of adversarial examples. Columns: development sets with each type of adversarial examples

	original data	latinize	phonetic latinize	add diacritic	delete letters	permute letters	introduce extra letters	confuse letters	sample substitute	remove punctuation	add comma	average
baseline	21.4	9.3	7.6	20.1	19.9	19.5	20.2	19.9	20.2	16.9	21.1	17.8
latinize	21.9	21.2	15.6	20.6	20.5	20.0	20.9	20.1	20.4	17.1	21.4	20.0
phonetic latinize	21.4	15.3	21.2	20.4	20.3	19.7	20.4	19.8	20.3	16.9	21.2	19.7
add diacritic	21.7	11.2	8.8	21.6	20.7	20.5	20.7	20.2	20.5	17.3	21.4	18.6
delete letters	21.8	12.0	9.5	20.8	21.1	20.7	20.9	20.2	20.5	17.0	21.2	18.7
permute letters	22.0	12.1	9.8	21.1	21.3	21.7	21.6	20.7	20.7	17.4	21.7	19.1
introduce extra letters	21.6	11.7	10.1	20.8	20.7	20.4	21.2	20.7	20.4	17.1	21.4	18.7
confuse letters	21.7	12.8	11.0	21.1	21.0	20.9	21.3	21.2	20.8	17.2	21.3	19.1
sample substitute	21.7	10.6	8.3	20.6	20.4	20.1	20.6	20.3	21.3	17.1	21.4	18.4
remove punctuation	21.6	9.5	7.6	20.0	20.2	19.3	20.4	19.9	20.5	20.4	21.5	18.3
add comma	21.3	9.3	7.5	20.2	20.0	19.6	20.5	20.0	20.2	17.3	21.5	17.9

Second, we observe that not all diagonal elements of Table 3 contain the highest BLEU score for their respective column, suggesting existing redundancies between the noise models. Examples are MT systems trained using adversarial examples from noise models that *delete letters* and *introduce extra letters*, which both when tested on their respective adversarial example development sets come second to the MT system that was trained using adversarial examples from the noise model that *permutes letters* (21.1 vs 21.3 BLEU points and 20.9 vs 21.6 BLEU points respectively). Similarly, the MT system trained using the noise model that *adds a comma* (21.5 BLEU), shows no benefit over the system that was trained using examples from the model that *removes punctuation* (21.5 BLEU). Based on these results, we decided not to include the redundant models (*delete letters*, *introduce extra letters* and *add comma*) in further experiments.

We, however, also recognize that the performance gains caused by the remaining noise models are numerically small (+0.5 BLEU) when compared against the next best performing MT system. For this reason, we use bootstrap re-sampling [18] to test if the performance gains of MT systems trained on adversarial examples generated by the noise models that *add a diacritic*, *confuse letters*, and perform *sample substitution* are statistically significant if compared against a system that is trained on adversarial examples generated by the noise model that *permutes letters*. Tests confirm that all differences are indeed significant at $p < 0.05$. Based on these tests, we include these models in our final experiments.

Large Data Systems. To test the effect of the seven productive noise models on MT system translation quality, we train Estonian-English, Latvian-English and Lithuanian-English large data MT systems. For systems trained using adversarial examples, we augment the original data with another copy of the data in which each type of noise is applied

Table 4. Test set results in BLEU points for large data MT systems

		original data	latinize	phonetic latinize	add diacritic	permute letters	confuse letters	sample substitute	remove punctuation	**average**
ET-EN	**baseline**	22.5	17.0	-	20.8	20.4	20.3	20.7	18.1	20.0
	+ adversarial noise	22.6	**22.5**	-	**22.4**	**22.2**	**22.0**	**21.8**	**21.7**	**22.2**
LV-EN	**baseline**	19.0	10.8	8.2	18.0	17.6	17.7	18.2	18.2	16.0
	+ adversarial noise	19.4	**18.8**	**18.9**	**19.2**	**19.0**	**18.9**	**19.0**	**18.6**	**19.0**
LT-EN	**baseline**	20.0	14.6	-	18.6	18.3	18.3	18.7	17.6	18.0
	+ adversarial noise	20.3	**19.3**	-	**20.0**	**19.9**	**19.7**	**19.7**	**20.5**	**19.9**

Table 5. Examples of noise in Latvian language input data causing widely different English language translations

Orig.	Twitter lietotāji nespēja noticēt, dzirdot Bairona Makdonalda neiejūtīgos komentārus.
Ref.	Twitter users did not hold back when they heard how insensitive Byron Macdonlad was being.
Src.	Twitter **leitotāji** nespēja noticēt, dzirdot Bairona Makdonalda neiejūtīgos komentārus.
Hyp.	Twitter's lieutenants couldn't believe it by hearing Byron McDonald's insensitive comments.
Src.	Twitter lietotāji nespēja noticēt, **dzidrot** Bairona Makdonalda neiejūtīgos komentārus.
Hyp.	Twitter users could not believe by clarifying Byron McDonald's insensitive comments.
Src.	**Twtiter** lietotāji nespēja noticēt, dzirdot Bairona Makdonalda neiejūtīgos komentārus.
Hyp.	Twtiter users couldn't believe it when they heard Byron McDonald's insensitive comments.
Src.	Twitter **liettoāji** nespēja noticēt, dzirdot Bairona Makdonalda neiejūtīgos komentārus.
Hyp.	The Twitter countryside couldn't believe it by hearing Byron McDonald's insensitive comments.

at an equal proportion. The evaluation results of these systems on the test data sets are provided in Table 4. First, we observe that the performance of the baseline MT systems and systems trained using adversarial examples on the original test data sets is comparable, suggesting that using adversarial examples does not harm the translation of clean data. Next, we observe, that when tested on noisy data systems that are trained using adversarial examples perform only slightly worse (about -0.4 BLEU on average) than when translating clean data, while baseline systems show an average performance drop of about 2 BLEU points.

Robustness and Noise Invariance. Besides measuring the changes in translation quality caused by noisy input data, we also would like to measure the robustness and noise invariance of the MT systems. We motivate this by the observation that often small perturbations in input data lead to widely different MT outputs (see Table 5). Desideratum, however, is that MT system outputs are noise invariant to an extent at least that unintentional changes in input data do not affect the meaning of the translation output. Results (see Table 6) of our experiments show that using adversarial examples in training im-

proves the robustness and noise invariance of the MT systems measured in 10NT-TER (see Section 2) on average by 0.1 10NT-TER points or in relative terms an average consistency improvement of about 50 %.

Table 6. Robustness and noise invariance of large data MT systems measured in 10NT-TER (see Section 2)

		latinize	phonetic latinize	add diacritic	permute letters	confuse letters	sample substitute	remove punctuation	**average**
ET-EN	**baseline**	0.27	-	0.13	0.14	0.15	0.13	0.29	0.19
	+ adversarial noise	**0.06**	-	**0.05**	**0.06**	**0.07**	**0.08**	**0.16**	**0.08**
LV-EN	**baseline**	0.51	0.63	0.16	0.17	0.17	0.12	0.28	0.29
	+ adversarial noise	**0.16**	**0.14**	**0.07**	**0.09**	**0.10**	**0.09**	**0.15**	**0.11**
LT-EN	**baseline**	0.39	-	0.19	0.21	0.20	0.15	0.39	0.25
	+ adversarial noise	**0.12**	-	**0.08**	**0.09**	**0.10**	**0.09**	**0.29**	**0.13**

5. Conclusions

We have proposed a simple generative noise model for the generation of adversarial examples for training data augmentation of NMT systems. Our results demonstrate that NMT systems that are trained using adversarial examples are more resilient to noisy input data. We show that while for the baseline NMT systems, noisy inputs cause a substantial drop in the translation quality (a drop of 2-3 BLEU points), for the systems that are trained using adversarial examples translation quality changes comparatively little (an average of -0.4 BLEU). In terms of translation robustness, systems trained on adversarial examples on average yield 50% consistency improvement when compared to baselines trained on clean data. Methods proposed here will be useful for achieving NMT robustness to orthographic and interpunctual variation in input data. This will be especially beneficial in use cases where NMT systems are used to translate texts of informal origins, such as chat conversations, social media posts and web pages with comment sections.

6. Acknowledgments

This research has been supported by the European Regional Development Fund within the joint project of SIA TILDE and University of Latvia “Multilingual Artificial Intelligence Based Human Computer Interaction” No. 1.1.1.1/18/A/148.

References

[1] Rawlinson GE. The significance of letter position in word recognition. University of Nottingham; 1976.

[2] McCusker LX, Gough PB, Bias RG. Word recognition inside out and outside in. Journal of Experimental Psychology: Human Perception and Performance. 1981;7(3):538.

[3] Belinkov Y, Bisk Y. Synthetic and Natural Noise Both Break Neural Machine Translation. In: International Conference on Learning Representations; 2018. .

[4] Michel P, Neubig G. MTNT: A Testbed for Machine Translation of Noisy Text. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing; 2018. p. 543–553.

[5] Carpuat M, Vyas Y, Niu X. Detecting Cross-Lingual Semantic Divergence for Neural Machine Translation. In: Proceedings of the First Workshop on Neural Machine Translation. Vancouver: Association for Computational Linguistics; 2017. p. 69–79.

[6] Khayrallah H, Koehn P. On the Impact of Various Types of Noise on Neural Machine Translation. In: Proceedings of the 2nd Workshop on Neural Machine Translation and Generation; 2018. p. 74–83.

[7] Zhou S, Zeng X, Zhou Y, Anastasopoulos A, Neubig G. 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); 2019. p. 565–571.

[8] Sperber M, Niehues J, Waibel A. Toward robust neural machine translation for noisy input sequences. In: International Workshop on Spoken Language Translation (IWSLT); 2017. .

[9] Vaibhav V, Singh S, Stewart C, Neubig G. 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); 2019. p. 1916–1920.

[10] Snover M, Dorr B, Schwartz R, Micciulla L, Makhoul J. A study of translation edit rate with targeted human annotation. In: Proceedings of association for machine translation in the Americas. vol. 200; 2006. .

[11] Koehn P, Hoang H, Birch A, Callison-Burch C, Federico M, Bertoldi N, et al. Moses: Open source toolkit for statistical machine translation. In: Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions. Association for Computational Linguistics; 2007. p. 177–180.

[12] Pinnis M, Vasiļjevs A, Kalniņš R, Rozis R, Skadiņš R, Šics V. Tilde MT Platform for Developing Client Specific MT Solutions. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018); 2018. .

[13] Junczys-Dowmunt M, Grundkiewicz R, Dwojak T, Hoang H, Heafield K, Neckermann T, et al. Marian: Fast neural machine translation in C++. arXiv preprint arXiv:180400344. 2018.

[14] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Advances in neural information processing systems; 2017. p. 5998–6008.

[15] Bogoychev N, Heafield K, Aji AF, Junczys-Dowmunt M. Accelerating Asynchronous Stochastic Gradient Descent for Neural Machine Translation. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics; 2018. p. 2991–2996. Available from: https://www.aclweb.org/anthology/D18-1332.

[16] Prechelt L. Early stopping-but when? In: Neural Networks: Tricks of the trade. Springer; 1998. p. 55–69.

[17] Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics; 2002. p. 311–318.

[18] Koehn P. Statistical significance tests for machine translation evaluation. In: Proceedings of the 2004 conference on empirical methods in natural language processing; 2004. p. 388–395.