Reproducibility of *Exploring Neural Text Simplification Models*: A Review

Mohammad Arvan
Luís Pina
Natalie Parde
University of Illinois Chicago
{marvan3,luispina,parde}@uic.edu

Introduction

- We rely on empirical evidence
- Yet, reproducing research is still a challenge.

Background

- Task: Text Simplification (TS)
- TS Metrics: BLEU, SARI
- Reproducibility Metrics: CV^* or coefficient of variation
- Dataset: EW-SEW (training), TurkCorpus (val, test)
- Models: LSTM with either random or pre-trained embedding

Methods

- Data (dataset and preprocessing)
- Software artifacts (code, dependencies, released models)
- Automatic Evaluation (empirical results)

Conclusion

Our work is reproducible! Try it for yourself.

Changing the **random seed** has a *larger* impact on the performance than **critical bugs** or **design decisions**.

Take a picture to download the full paper

Object	Measurand	Sample Size	Mean	Unbiased STDEV	STDEV 95% CI	CV^*
NTS	SARI	8	30.23	0.56	[0.23, 0.89]	1.92
NTS	BLEU	13	86.07	1.64	[0.94, 2.34]	1.94
NTS w2v	SARI	7	30.22	0.96	[0.34, 1.58]	3.28
NTS w2v	BLEU	12	87.71	2.45	[1.35, 3.54]	2.85

Table 1: Precision (CV^*) and component measures (mean, standard deviation, standard deviation confidence intervals) for measured quantity values obtained in multiple measurements of the two *NTS* systems.

Measurand	Mean	Min	Max
SARI	29.24 ± 0.31	28.62	29.89
BLEU	87.9 ± 1.18	84.47	89.59

Table 2: Results of the random seed experiments on the TurkCorpus test set, with a sample size of 36. Models are trained with the same configuration, but have unique random seeds. The evaluation script by Nisioi et. al was used.

Object	Measurand	Eval. Script by	Measured Value
NTS w2v	BLEU	t1	87.04
NTS w2v	BLEU	sb2.1	87.10
NTS w2v	SARI	t1	29.70
NTS w2v †	BLEU	t1	89.43
NTS w2v †	BLEU	sb2.1	89.40
NTS w2v †	SARI	t1	29.80
NTS w2v †‡	BLEU	t1	89.12
NTS w2v †‡	BLEU	sb2.1	89.10
NTS w2v †‡	SARI	t1	29.58
NTS w2v ‡	BLEU	t1	88.01
NTS w2v ‡	BLEU	sb2.1	88.00
NTS w2v ‡	SARI	t1	29.18

Table 3: Results of the experiments tracking performance impacts for identified issues, computed for this paper using our version of the model, our output, and the evaluation script provided by Nisioi et. al and sacreBLEU. † indicates contaminated conditions, and ‡ indicates mismatched conditions.

System	BLEU ($\mu \pm$ 95% CI)
Baseline: NTS w2v	87.9 (87.9 ± 2.0)
NTS	84.6 (84.6 ± 2.9)

Table 4: Statistical significance analysis performed on Nisioi et al's released output. With $p = 0.0079$, the difference in reported results between the two variants is statistically significant.

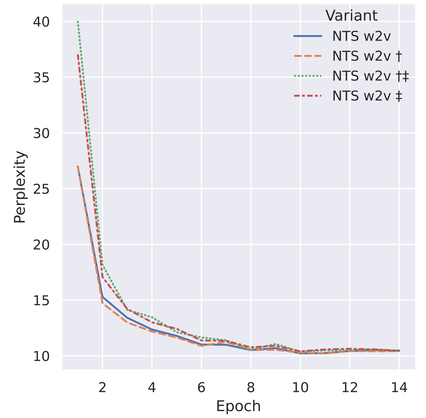

Figure 1: Validation perplexity of *NTS w2v* variants during training (lower is better). † indicates contaminated conditions, and ‡ indicates mismatched conditions.