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

Mohammad Arvan, Luís Pina, Natalie Parde

{marvan3,luispina,parde}@uic.edu

Department of Computer Science
University of Illinois at Chicago

Introduction

- We rely on empirical evidence
- Yet, reproducing research is still a challenge.
- ‘Exploring Neural Text Simplification Models’ by Nisioi et al. 2017

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)

Data Reproducibility

- Original dataset is no longer available
- The released source code does not contain preprocessing steps
- Preprocessed dataset is included in the repository

Software Artifacts Reproducibility

- Released artifacts are of high quality
- 5/5 ML Completeness Checklist
- All important dependencies have been deprecated for years. (Python 2.7, Torch7, OpenNMT)

Software Artifacts Reproducibility (cont.)

- Bugs/issues affecting NTS w2v models
- 2/3 Have been confirmed by the authors

Source Code Issues

Issue 1 Data Contamination

Issue 2 Mismatched Embedding

Issue 3 Zero Embedding Weight

Reproducibility of Automatic Evaluation

- Follow a similar training procedure
- We evaluate three new outputs
    1. Provided by Nisioi et al. (2017)
    2. Generated by running the trained model provided by Nisioi et al. (2017)
    3. Generated by training and running a model based using a modified version of the source code provided by Nisioi et al. (2017)

Results

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.

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 2: 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.

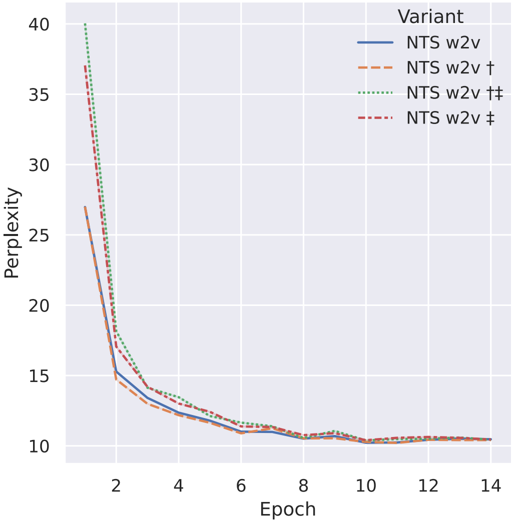

Figure 1: Validation perplexity of *NTS w2v* variants during training (lower is better). † indicates contaminated conditions, and ‡ indicates mismatched conditions.

System	BLEU (μ ± 95% CI)
Baseline: NTS w2v	87.9 (87.9 ± 2.0)
NTS	84.6 (84.6 ± 2.9)

Table 3: Statistical significance analysis performed on Nisioi et al. released output. With $p = 0.0079$, the difference in reported results between the two variants is statistically significant.

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

Table 4: 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.

Discussions

- We do not see enough evidence to justify the performance gains are coming from design decisions
- On the other hand, changing random seed seems to cause the observed variation
- We find the resilience to bugs in neural networks quite alarming.

Discussions (cont.)

- The unavailability of full runtime environment will render most research obsolete
- It was quite challenging to get code to a running state.
- We have taken steps to ensure reproducibility of our work

Thank you!

More information available in our paper.