Cleaning E2E training data fixes up to 97% NLG semantic errors

Semantic Noise Matters
for Neural Natural Language Generation

Ondřej Dušek* David M. Howcroft† Verena Rieser†

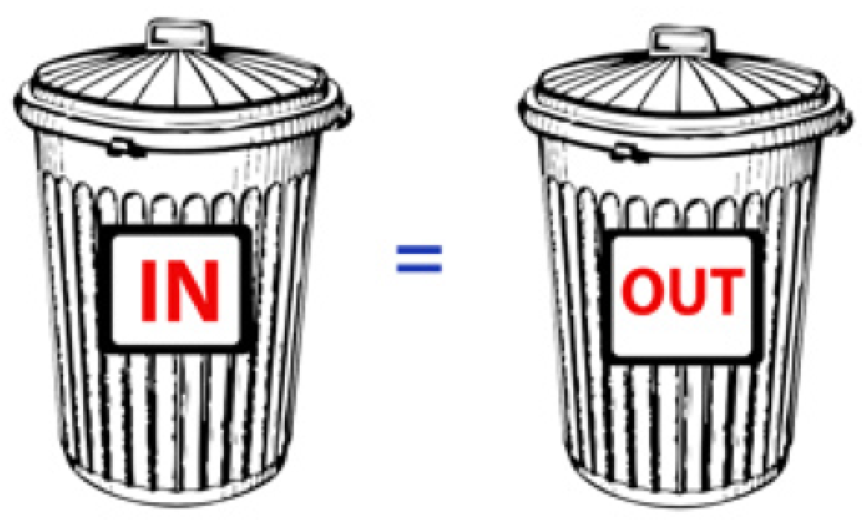
HERIOT WATT UNIVERSITY

CHARLES UNIVERSITY

Research Questions

- Does noisy data matter for Neural Natural Language Generation (**NNLG**)?
- Can NNLG systems learn to ignore errors in training data by generalising away from them?

The Problem: Noisy Training Data



Crowdworkers introduce more noise than expected

- Insertions
- Deletions
- Alterations

Measured by calculating the **Semantic Error Rate**

$$\text{SER} = \frac{\#\text{added} + \#\text{missing} + \#\text{wrong value}}{\#\text{slots}}$$

End-to-End Generation Challenge Corpus (E2E)

- collected via crowdsourcing
- used by 17 teams in the E2E challenge
- used in 13 published papers since

11–17% SER in the E2E dataset

- approx. 40% of references include ≥ 1 error

Fixing the E2E NLG Challenge Dataset

We cleaned the data! (a little goes a long way)

- our heuristic script for SER also provides corrections
- good accuracy but not perfect
  - SER 4.2%; 19.5% of references with errors
- some cleaned MRs from TRAIN&DEV overlapped TEST
  - these instances were removed
  - systems trained on cleaned data can be evaluated on original TEST
- cleaned data: fewer instances, more distinct MRs
  - more challenging for training

Data statistics

Dataset	Part	MRs	Refs	SER(%)
Original	TRAIN	4,862	42,061	17.69
	DEV	547	4,672	11.42
	TEST	630	4,693	11.49
Cleaned	TRAIN	8,362	33,525	(0.00)
	DEV	1,132	4,299	(0.00)
	TEST	1,358	4,693	(0.00)

Table: # of distinct MRs, # of reference texts, and SER as measured by our script.

Example MR Fixes by Our Script

Original MR and an accurate reference

MR name[Cotto], eatType[coffee shop], food[English], priceRange[less than £20], customer_rating[low], area[riverside], near[The Portland Arms]

Reference At the riverside near The Portland Arms, Cotto is a coffee shop that serves English food at less than £20 and has low customer rating.

Example corrections

Reference: Cotto is a coffee shop that serves English food in the city centre. They are located near the Portland Arms and are low rated.
Correction: removed price range; changed area
Reference: Cotto is a cheap coffee shop with one-star located near The Portland Arms.
Correction: removed area

A faulty correction

Reference: Located near The Portland Arms in riverside, the Cotto coffee shop serves English food with *a price range of $20* and a low customer rating.
Correction: incorrectly(!) removed price range
– our script's slot patterns are not perfect

Impact on Neural NLG Systems

Cleaned data can **reduce errors by up to 97%**

Results

System	TRAIN	BLEU	NIST	A	M	V	SER	
Seq2Seq	Original	63.37	7.71	0.06	15.77	0.11	15.94	-94%
	Cleaned added	64.40	7.96	0.01	13.08	0.00	13.09	
	Cleaned missing	66.28	8.52	0.14	2.26	0.22	2.61	
	Cleaned	65.87	8.64	0.20	0.56	0.21	0.97	
TGen	Original	66.41	8.55	0.14	4.11	0.03	4.27	-97%
	Cleaned added	66.23	8.55	0.04	3.04	0.00	3.09	
	Cleaned missing	67.00	8.68	0.06	0.44	0.03	0.53	
	Cleaned	66.24	8.68	0.10	0.02	0.00	0.12	

Table: A = % instances with added slots, M = missed slots, V = wrong values

- fixing missing slots has the biggest effect
- results confirmed by manual analysis
  - SER script 99.93% accurate on system outputs
- SC-LSTM is more affected by noise and works poorly on E2E data in general, likely due to reliance on delexicalization

Conclusions and Future Work

Semantic noise matters

- Crowdsourced datasets are noisy, so clean your data!

Get our data & code here:

`https://github.com/tuetschek/e2e-cleaning`

What's next?

- continuing to improve data & checking effects on diversity

D.Howcroft@hw.ac.uk authors listed alphabetically
*Charles University, Prague, Czech Republic
†Heriot-Watt University, Edinburgh, United Kingdom