Noise and Neural Natural Language Generation
Rubbish in, Rubbish out?

Ondřej Dušek* David M. Howcroft† Karin Sevegnani† Verena Rieser†

Research Questions

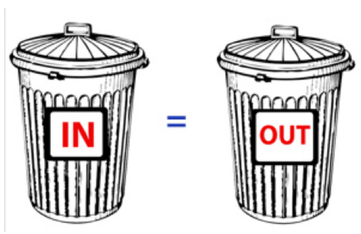
- Does noisy data matter for Neural Natural Language Generation (**NNLG**)?
- How do errors in the generated output **affect user ratings**?
- How noisy are **user ratings**?

The Problem

NNLG systems behave oddly.

→ *How much is caused by the data?*

→ *Can NNLG systems learn to ignore errors in training data by generalising away from them?*



Types of Noise Observed in the Data:

- *Semantic*: content omitted/inserted when e.g. crowdsourcing
- *Typographic*: misspellings
- *Grammatical*: disfluencies, non-standard syntax, lack of punctuation

Errors in NNLG Systems:

- *Content*: 'Hallucinations' and omissions
- Lack of *fluency*

Semantic Noise in the E2E NLG Challenge

- Original corpus has a **16.37% Semantic Error Rate**
- We **corrected MRs** to match their texts and
- ...**compared 2 NNLGs** with different semantic control mechanisms: TGen [1] & SC-LSTM [2]

Cleaned data can **improve performance by up to 97%**

→ *Semantic noise in the training data matters for NNLG!* [3]

Example MR Fixes from the E2E Challenge

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

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

NNLG errors matter to users

Manually annotated NNLG errors
...in 500 system outputs from E2E [4] and NEM [5] datasets

Different error types impact user ratings differently
Post hoc comparisons of error type (fluency vs. content) show that outputs with *fluency* errors receive lower human ratings.

	Disfluency	Content	Disf.+Cont.	No error
E2E	15%	27%	11%	47%
NEM	12%	26%	3%	59%

Table: Error percentage generated from neural systems, contained in both E2E and NEM datasets. *Content* = hallucinations + omissions.

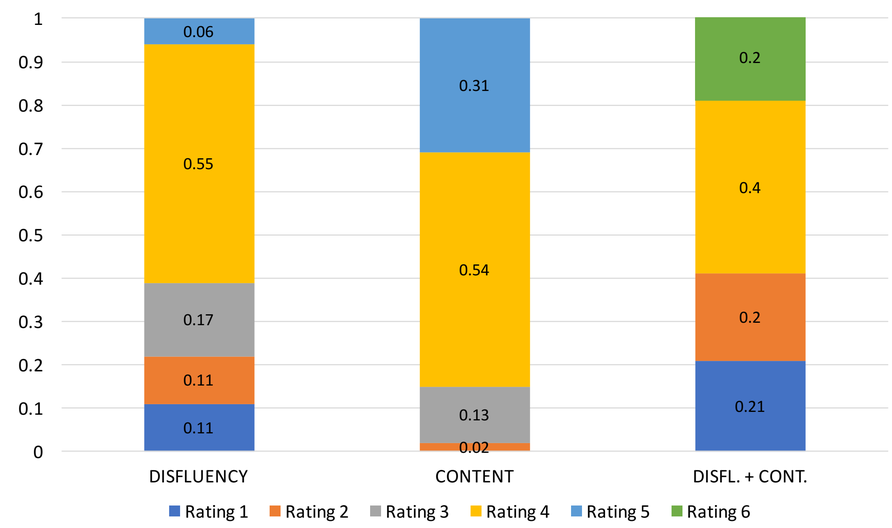


Figure: Distribution of human ratings for the different types of errors in the NEM dataset

Better Instructions, Better User Ratings

New Annotation on NEM

- More detailed instructions
- Provided examples for ratings
- Qualification test for users

System	Pearson	Spearman	MAE	RMSE
Old data on avg	0.309	0.284	0.915	1.208
New data on avg	**0.458**	0.402	0.593	0.836
Old data on med	0.287	0.273	0.915	1.237
New data on med	**0.483**	**0.422**	**0.322**	**0.702**

Table: Comparison of RatPred trainable quality estimation system [6] prediction results on the original and the improved dataset, using identical system settings.

→ **More consistent human ratings!**

Ongoing and future work

- impact of typographic and grammatical noise on NNLG
- crowdsourcing data cleaning / denoising
- adversarial training to increase robustness
- principled way to generate adversarial examples
- alternatives to ratings: reading times for evaluating fluency

References

[1] Dušek, O., & F. Jurčíček .2016. 'Sequence-to-Sequence Generation for Spoken Dialogue via Deep Syntax Trees and Strings.' *ACL*.
[2] Wen, T.-H., P.-h. Su, D. Vandyke, & S. Young. 2015. 'Semantically Conditioned LSTM-based Natural Language Generation for Spoken Dialogue Systems.' *EMNLP*.
[3] Dušek, O., D. M. Howcroft, & V. Rieser. 2019. 'Semantic Noise Matters for Neural Natural Language Generation.' *INLG*.
[4] Dušek, O., J. Novikova, & V. Rieser. 2020. 'Evaluating the state-of-the-art of End-to-End Natural Language Generation: The E2E NLG challenge'. *Computer Speech & Language*.
[5] Novikova, J., O. Dušek, A. Cercas Curry, & V. Rieser. 2017. 'Why We Need New Evaluation Metrics for NLG'. *EMNLP*
[6] Dušek, O., K. Sevegnani, I. Konstas, & V. Rieser. 2019. 'Automatic Quality Estimation for Natural Language Generation: Ranting (Jointly Rating and Ranking)'. *INLG*.