Tackling Hallucinations in Neural Chart Summarization

Saad Obaid ul Islam
saadob12@gmail.com

Iza Škrjanec
skrjanec@coli.uni-saarland.de

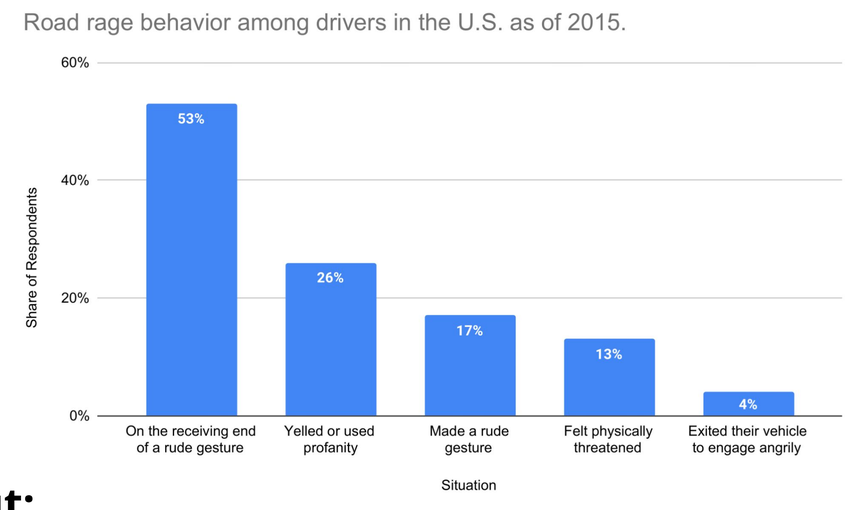
Ondrej Dušek
odusek@ufal.mff.cuni.cz

Vera Demberg
vera@coli.uni-saarland.de

NLI preprocessing & input format adjustments alleviate hallucinations

Problem and Task

Input:



Output:

This statistic shows the road rage behavior of drivers in the United States as of 2015. Four percent of the drivers said they have been on the receiving end of a rude gesture. The survey was conducted online and all the participants had a valid U.S. driving license.

Hallucinations
= generated text not grounded in the input

Intrinsic Hallucinations
= verifiable from the input

Extrinsic Hallucinations
= not verifiable from the input

Our Contributions

1. Showing that providing **more context** and **reducing long-distance dependencies** in the linearized input format is important.
2. **NLI cleaning** step to **remove ungrounded** information in the training and test data.

1. Context & Distance in Input Format

Obeid & Hoque: xlabel1 | xvalue1 | x | chart-type | ylabel1 | yvalue1 | y | chart-type ... xlabel2 | xvalue2 ...

- no title, repetitive: 22 + 13 hallucinations in 50 sents.

Kantharaj et al: title yvalue1 yvalue2 ... xvalue1 xvalue2

- no x-y labels, long-dist. deps.: 4 + 11 hallu. in 50 sents.

Ours: **title xlabel - ylabel xvalue1 yvalue1, xvalue2 yvalue2... xvalueN yvalueN**

- adding title = biggest improvement
- adding x-y labels = minor improvements
- title + x-y labels + pairing labels & values = best

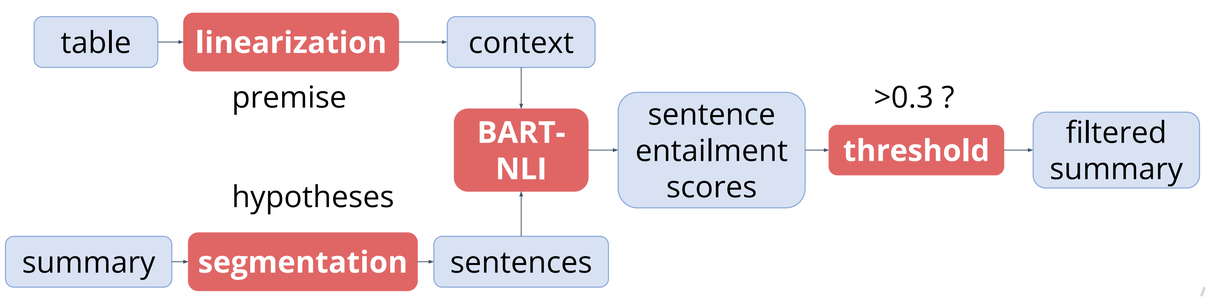
2. Cleaning Noisy Training and Test Data

Why?: 20/50 references contained ungrounded info in C2T-Small dataset

Hypothesis: Ungrounded info in training data → hallucinations in system outputs

Proof: Autochart dataset + noise → 27/50 outputs with hallucinations

Approach: **Filter data using NLI entailment**



Results

Model	BLEU	ROUGE-2	PPL	Log. Agree.	Log. Contra.	NUBIA
C2T-Small Data						
Obeid & Hoque	18.5	-	-	-	-	-
T5 + Obeid & Hoque	26.1	33.5	7.4	5.5	67.8	35.4
T5 + Ours[1]	33.9	**44.8**	7.5	**33.2**	22.3	**46.9**
T5 + Ours + NLI[2]	**34.2**	43.7	**7.1**	33.1	**10.2**	44.5
C2T-Big Data						
T5 Kantharaj et al.	37.0	50.6	10.0	34.5	22.9	53.5
T5 + Ours	39.8	**55.0**	**8.2**	39.3	21.3	**55.6**
T5 + Ours + NLI	**42.2**	50.7	**8.2**	**40.3**	**15.1**	53.5

Error analysis: [1]0 + 18 / [2]0 + 4 hall. in 50 sents.

Human Evaluation

Model	Values Correct	Has Outside Info	Informative	Coherent	Fluent
C2T-Small Data, 50 samples					
T5 + Ours	56.00%	38.00%	3.80/5	3.81/5	3.88/5
T5 + Ours + NLI	*76.00%	*17.00%	3.60/5	3.91/5	3.96/5

*significant difference

Key takeaways and Discussion

- More context & less long-distance deps → less intrinsic hall.
- Ungrounded info in training data → hallucinations in output
- NLI filtering → significantly less hallucination

- Gold-standard datasets have ungrounded info
- Automatic metrics do not measure hallucinations well

https://github.com/WorldHellow/Hallucinations-C2T

Presented at INLG 2023, Prague, Czechia.

Supported by ERC NG-NLG (101039303) & EUIN-ACTION from NORFACE Governance (462-19-010, GL950/2-1). Using resources provided by LINDAT/CLARIAH-CZ (Czech Ministry of Education LM2018101).