Text-in-Context: Token-Level Error Detection for Table-to-Text Generation

Zdeněk Kasner kasner@ufal.mff.cuni.cz
Simon Mille simon.mille@upf.edu
Ondřej Dušek odusek@ufal.mff.cuni.cz

CHARLES UNIVERSITY

Universitat Pompeu Fabra Barcelona

Inputs (i) Rotowire basketball **data**
(ii) data **descriptions** generated by neural NLG systems

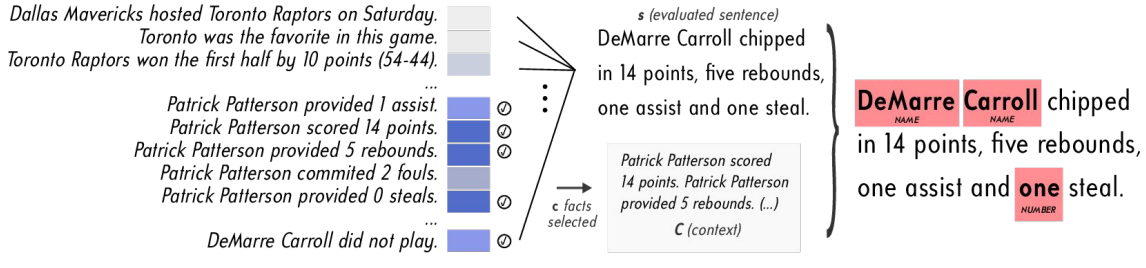
Outputs factual **errors** in the descriptions detected on **token-level**

Shared Task in Evaluating Accuracy @ INLG 2021

How-to in three steps:

① **Generate facts** from the input table with a rule-based NLG system.

② **Select the relevant subset of facts** based on the similarity of sentence embeddings (the fact vs. the checked sentence).



③ **Train RoBERTa to label the tokens with error categories** based on the selected subset of facts (=context) on human-annotated data from organizers & synthetic from RotoWire train set.

Results:

1 *out of 4 automatic metrics in the Shared Task.*

dev results - F1 score

	simple						compact					
	synth			synth+human			synth			synth+human		
EMR/c	***10***	***20***	***40***	***10***	***20***	***40***	***10***	***20***	***40***	***10***	***20***	***40***
25%	0.232	0.231	0.268	0.506	0.573	0.613	0.278	0.303	0.264	0.620	0.634	0.650
50%	0.272	0.271	0.294	0.519	0.559	0.603	0.297	0.310	0.296	0.617	0.639	0.619
75%	0.360	0.373	0.353	0.521	0.549	0.573	0.367	0.370	0.388	0.630	0.629	0.620

c = context size (# selected facts), ***EMR*** = synth data error level (% replaced entities)

test results by categories

error	recall	precision
NAME	0.750	0.846
NUMBER	0.777	0.750
WORD	0.514	0.483
CONTEXT	0.000	-
NOT_CHK.	0.000	-
OTHER	0.000	-
Overall	0.691	0.756

Presented at INLG 2021, online.

https://github.com/kasnerz/accuracySharedTask_CUNI-UPF

Supported by the Charles University projects GAUK 140320, SVV 260575, and PRIMUS/19/SCI/10, and by the European Commission via UPF under the H2020 program contract numbers 786731, 825079, 870930 and 952133.