Appendix for ToTTo: A Controlled Table-To-Text Generation Dataset

Ankur P. Parikh[♠] **Xuezhi Wang**[♠] **Sebastian Gehrmann**[♠]
Manaal Faruqui[♠] **Bhuwan Dhingra**[♣*] **Diyi Yang**[♠♢] **Dipanjan Das**[♠]
[♠] Google Research, New York, NY
[♢] Georgia Tech, Atlanta, GA
[♣] Carnegie Mellon University, Pittsburgh, PA

totto@google.com

The Appendix contains the following contents:

- Information about the variant of the PARENT metric (Dhingra et al., 2019) used for evaluation.
- More details about the baselines.
- Examples of the annotation process (Table 1).
- More examples of model errors (Table 2).
- Examples of more complex tables in our dataset (Figures 1-Figures 5).
- The dataset (development subset) is attached in a separate supplemental zip file.

1 PARENT metric

PARENT (Dhingra et al., 2019) is a metric recently proposed specifically for data-to-text evaluation that takes the table into account. We modify it to make it suitable for our dataset. Let $(\boldsymbol{x}_n, \boldsymbol{y}_n, \hat{\boldsymbol{y}}_n)$ denote one example that consists of a (source, target, prediction) tuple. PARENT is defined at an instance level as:

$$PARENT(\boldsymbol{x}_n, \boldsymbol{y}_n, \hat{\boldsymbol{y}}_n) = \frac{2 \times E_p(\boldsymbol{x}_n, \boldsymbol{y}_n, \hat{\boldsymbol{y}}_n) \times E_r(\boldsymbol{x}_n, \boldsymbol{y}_n, \hat{\boldsymbol{y}}_n)}{E_p(\boldsymbol{x}_n, \boldsymbol{y}_n, \hat{\boldsymbol{y}}_n) + E_r(\boldsymbol{x}_n, \boldsymbol{y}_n, \hat{\boldsymbol{y}}_n)}$$

$E_p(\boldsymbol{x}_n, \boldsymbol{y}_n, \hat{\boldsymbol{y}}_n)$ is the PARENT precision computed using the prediction, reference, and table (the last of which is not used in BLEU). $E_r(\boldsymbol{x}_n, \boldsymbol{y}_n, \hat{\boldsymbol{y}}_n)$ is the PARENT recall and is computed as:

$$E_r(\boldsymbol{x}_n, \boldsymbol{y}_n, \hat{\boldsymbol{y}}_n) = R(\boldsymbol{x}_n, \boldsymbol{y}_n, \hat{\boldsymbol{y}}_n)^{(1-\lambda)} R(\boldsymbol{x}_n, \hat{\boldsymbol{y}}_n)^{\lambda}$$

where $R(\boldsymbol{x}_n, \boldsymbol{y}_n, \hat{\boldsymbol{y}}_n)$ is a recall term that compares the prediction with both the reference and table. $R(\boldsymbol{x}_n, \hat{\boldsymbol{y}}_n)$ is an extra recall term that gives an additional reward if the prediction $\hat{\boldsymbol{y}}_n$ contains phrases in the table $\boldsymbol{x}_n$ that are not necessarily in the reference (λ is a hyperparameter).

In the original PARENT work, the same table $\boldsymbol{t}$ is used for computing the precision and both recall terms. While this makes sense for most existing datasets, it does not take into account the highlighted cells $\boldsymbol{t}_{highlight}$ in our task. To incorporate $\boldsymbol{t}_{highlight}$, we modify the PARENT metric so that the additional recall term $R(\boldsymbol{x}_n, \hat{\boldsymbol{y}}_n)$ uses $\boldsymbol{t}_{highlight}$ instead of $\boldsymbol{t}$ to only give an additional reward for relevant table information. The other recall and the precision term still use $\boldsymbol{t}$.

2 Baseline details

We provide some more baseline details here.

- BERT-to-BERT (Rothe et al., 2020) - Uncased model coupling both encoder and decoder as in original paper, with Adam optimizer. learning rate = 0.05, hidden size = 1024, dropout = 0.1, beam size = 4.
- Pointer Generator (See et al., 2017) - LSTM with hidden size 300, beam size=8, learning rate = 0.0003, dropout = 0.2, length penalty = 0.0, Adam optimizer.
- Content planner (Puduppully et al., 2019) - All of the original hyperparameters: content planner: LSTM with hidden size 1x600, realizer LSTM with 2x600, embedding size 600 for both, dropout=0.3, Adagrad optimizer, beam size=5.

References

Bhuwan Dhingra, Manaal Faruqui, Ankur Parikh, Ming-Wei Chang, Dipanjan Das, and William W Cohen. 2019. Handling divergent reference texts when evaluating table-to-text generation. In *Proc. of ACL*.

[*] Work done during an internship at Google.

Ratish Puduppully, Li Dong, and Mirella Lapata. 2019. Data-to-text generation with content selection and planning. In *Proc. of AAAI*.

Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2020. Leveraging pre-trained checkpoints for sequence generation tasks. In *Proc. of TACL*.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proc. of ACL*.

Original	After Deletion	After Decontextualization	Final
He was the first president of the Federal Supreme Court (1848–1850) and president of the National Council in 1850–1851.	He was the first president of the Federal Supreme Court (1848–1850) ~~and president of the National Council in 1850–1851~~.	Johann Konrad Kern was the first president of the Federal Supreme Court from 1848 to 1850.	Johann Konrad Kern was the first president of the Federal Supreme Court from 1848 to 1850.
He later raced a Nissan Pulsar and then a Mazda 626 in this series, with a highlight of finishing runner up to Phil Morriss in the 1994 Australian Production Car Championship.	He ~~later~~ raced a Nissan Pulsar and then a Mazda 626 ~~in this series, with a highlight of~~ finishing runner up ~~to Phil Morriss~~ in the 1994 Australian Production Car Championship.	Murray Carter raced a Nissan Pulsar and finished as a runner up in the 1994 Australian Production Car Championship.	Murray Carter raced a Nissan Pulsar and finished as runner up in the 1994 Australian Production Car Championship.
On July 6, 2008, Webb failed to qualify for the Beijing Olympics in the 1500 m after finishing 5th in the US Olympic Trials in Eugene, Oregon with a time of 3:41.62.	On July 6, 2008, Webb ~~failed to qualify for the Beijing Olympics in the 1500 m after~~ finishing 5th in the ~~US~~ Olympic Trials in Eugene, Oregon with a time of 3:41.62.	On July 6, 2008, Webb finishing 5th in the Olympic Trials in Eugene, Oregon with a time of 3:41.62.	On July 6, 2008, Webb finished 5th in the Olympic Trials in Eugene, Oregon, with a time of 3:41.62.
Out of the 17,219 inhabitants, 77 percent were 20 years of age or older and 23 percent were under the age of 20.	~~Out of the~~ 17,219 inhabitants ~~, 77 percent were 20 years of age or older and 23 percent were under the age of 20~~.	Rawdat Al Khail had a population of 17,219 inhabitants.	Rawdat Al Khail had a population of 17,219 inhabitants.

Table 1: Examples of annotation process. Deletions are indicated in red strikeouts, while added named entities are indicated in underlined blue. Significant grammar fixes are denoted in orange.

ID	Reference	Decoder output (w/ metadata)		w/o metadata
		Full table	Subtable	Subtable
1	in the 1939 currie cup, western province lost to transvaal by 176 in cape town.	northern transvaal and western province were drawn by 16 in 1989 and in 1995, western province were defeated by the sharks in durban.	the first currie cup was played in 1939 in transvaal1 at newlands, with western province winning 176.	the first scottish cup was played in 1939 in transvaal1, where transvaal1 defeated western province 176 in the final at newlands in cape town.
2	a second generation of microdrive was announced by ibm in 2000 with increased capacities at 512 mb and 1 gb.	the microdrive models formed 512 megabyte and 1 gigabyte in 2000.	there were 512 microdrive models in 2000: 1 gigabyte.	cortete's production was 512 megabyte.
3	the 1956 grand prix motorcycle racing season consisted of six grand prix races in five classes: 500cc, 350cc, 250cc, 125cc and sidecars 500cc.	the 1966 grand prix motorcycle racing season consisted of seven grand prix races in five classes: 500cc, 350cc, 250cc, 125cc and sidecars 500cc.	the 1956 grand prix motorcycle racing season consisted of eight grand prix races in five classes: 500cc, 350cc, 250cc, 125cc and sidecars 500cc.	the 1955 grand prix motorcycle racing season consisted of eight grand prix races in five classes: 500cc, 350cc, 250cc, 125cc and sidecars 500cc.
4	in travis kelce's last collegiate season, he set personal career highs in receptions (45), receiving yards (722), yards per receptions (16.0) and receiving touchdowns (8).	during the 2011 season, travis kelceum caught 76 receptions for 1,612 yards and 14 touchdowns.	travis kelce finished the 2012 season with 45 receptions for 722 yards (16.0 avg.) and eight touchdowns.	kelce finished the 2012 season with 45 catches for 722 yards (16.0 avg.) and eight touchdowns.
5	in the 2012 film pizza bagel, michael pillarella portrays tommy.	in 2012, groff played the role of tommy in the film pizza bagel.	in 2012, pillarella appeared as tommy in the film pizza bagel.	harris played the role of tommy in the 2012 film pizza bagel.
6	the album shari addison placed at no. 176 on the billboard 200 along with no. 5 on the gospel albums.	shari addison's " 5", reached number 176 on the billboard 200.	shari addison charted at number 176 on the us chart and at number 5 on the us billboard 200.	the shari addison peaked at number 176 on the billboard 200 chart.
7	the bnp secured their best general election result in oldham west and royton where nick griffin secured 16.4% of the votes.	bnp results ranged from 278 to 6,552 votes.	in the british national party election, nick griffin placed third with 16.4% of the vote.	in oldham west and royton, nick griffin won 16.4% of the vote.

Table 2: Decoder output examples from BERT-to-BERT Books models on the development set. The "subtable with metadata" model achieves the highest BLEU. Red indicates model errors and blue denotes interesting reference language not in the model output.

Table Title: Robert Craig (American football)
Section Title: National Football League statistics
Table Description: *None*

	Rushing						Receiving				
YEAR	TEAM	ATT	YDS	AVG	LNG	TD	NO.	YDS	AVG	LNG	TD
1983	SF	176	725	4.1	71	8	48	427	8.9	23	4
1984	SF	155	649	4.2	28	4	71	675	9.5	64	3
1985	SF	214	1,050	4.9	62	9	92	1,016	11.0	73	6
1986	SF	204	830	4.1	25	7	81	624	7.7	48	0
1987	SF	215	815	3.8	25	3	66	492	7.5	35	1
1988	SF	310	1,502	4.8	46	9	76	534	7.0	22	1
1989	SF	271	1,054	3.9	27	6	49	473	9.7	44	1
1990	SF	141	439	3.1	26	1	25	201	8.0	31	0
1991	RAI	162	590	3.6	15	1	17	136	8.0	20	0
1992	MIN	105	416	4.0	21	4	22	164	7.5	22	0
1993	MIN	38	119	3.1	11	1	19	169	8.9	31	1
Totals	—	**1,991**	**8,189**	**4.1**	**71**	**56**	**566**	**4,911**	**8.7**	**73**	**17**

Target sentence: Craig finished his eleven NFL seasons with 8,189 rushing yards and 566 receptions for 4,911 receiving yards.

Figure 1: ToTTo example with numerical reasoning about table cells.

Table Title: Ken Fujita
Section Title: Club statistics
Table Description: *None*

Club performance			League		Cup		League Cup		Total	
Season	Club	League	Apps	Goals	Apps	Goals	Apps	Goals	Apps	Goals
Japan			League		Emperor's Cup		J.League Cup		Total	
1998	Júbilo Iwata	J1 League	0	0	0	0	0	0	0	0
2001	Ventforet Kofu	J2 League	35	4	3	0	2	0	40	4
2002			33	5	2	0			35	5
2003			39	9	1	0			40	9
2004			28	2	1	0			29	2
2005			41	10	2	0			43	10
2006		J1 League	26	2	3	1	1	0	30	3
2007			32	2	1	0	7	0	40	2
2008		J2 League	38	3	1	0			39	3
2009			50	2	2	0			52	2
2010			32	2	1	0			33	2
Country	**Japan**		**354**	**41**	**15**	**1**	**10**	**0**	**379**	**42**
Total			**354**	**41**	**15**	**1**	**10**	**0**	**379**	**42**

Target sentence: After 2 years blank, Ken Fujita joined the J2 League club Ventforet Kofu in 2001.

Figure 2: ToTTo example with complex table structure and temporal reasoning.

Table Title: Shuttle America
Section Title: Fleet
Table Description: *As of January 2017, the Shuttle America fleet consisted of the following aircraft:*

Aircraft	Total	Orders	Passengers				Operated For	Notes
			F	Y+	Y			
Embraer E170	5	—	6	16	48	70	United Express	transferred to Republic Airline
	14	—	9	12		69	Delta Connection Delta Shuttle	2 planes on wet lease from Republic Airline
Embraer E175	16	—	12	12	52	76		
Total	**35**	**—**						

Target sentence: Shuttle America operated the E-170 and the larger E-175 aircraft for Delta Air Lines,.

Figure 3: ToTTo example with rare topics and complex table structure.

Table Title: Pune - Nagpur Humsafar Express
Section Title: Schedule
Table Description: *None*

Train Number	Station Code	Departure Station	Departure Time	Departure Day	Arrival Station	Arrival Time	Arrival Day
11417	PUNE	Pune Junction	22:00 PM	Thu	Nagpur Junction	13:30 PM	Fri
11418	NGP	Nagpur Junction	15:00 PM	Fri	Pune Junction	08:05 AM	Sat

Target sentence: The 11417 Pune - Nagpur Humsafar Express runs between Pune Junction and Nagpur Junction.

Figure 4: ToTTo example with rare topic.

Table Title: Montpellier
Section Title: Climate
Table Description: *None*

Climate data for Montpellier (1981–2010 averages)													
Month	**Jan**	**Feb**	**Mar**	**Apr**	**May**	**Jun**	**Jul**	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**	**Year**
Record high °C (°F)	21.2 (70.2)	22.5 (72.5)	27.4 (81.3)	30.4 (86.7)	35.1 (95.2)	37.2 (99.0)	37.5 (99.5)	36.8 (98.2)	36.3 (97.3)	31.8 (89.2)	27.1 (80.8)	22.0 (71.6)	37.5 (99.5)
Average high °C (°F)	11.6 (52.9)	12.8 (55.0)	15.9 (60.6)	18.2 (64.8)	22.0 (71.6)	26.4 (79.5)	29.3 (84.7)	28.9 (84.0)	25.0 (77.0)	20.5 (68.9)	15.3 (59.5)	12.2 (54.0)	19.9 (67.8)
Daily mean °C (°F)	7.2 (45.0)	8.1 (46.6)	10.9 (51.6)	13.5 (56.3)	17.3 (63.1)	21.2 (70.2)	24.1 (75.4)	23.7 (74.7)	20.0 (68.0)	16.2 (61.2)	11.1 (52.0)	8.0 (46.4)	15.1 (59.2)
Average low °C (°F)	2.8 (37.0)	3.3 (37.9)	5.9 (42.6)	8.7 (47.7)	12.5 (54.5)	16.0 (60.8)	18.9 (66.0)	18.5 (65.3)	15.0 (59.0)	11.9 (53.4)	6.8 (44.2)	3.7 (38.7)	10.4 (50.7)
Record low °C (°F)	−15 (5)	−17.8 (0.0)	−9.6 (14.7)	−1.7 (28.9)	0.6 (33.1)	5.4 (41.7)	8.4 (47.1)	8.2 (46.8)	3.8 (38.8)	−0.7 (30.7)	−5 (23)	−12.4 (9.7)	−17.8 (0.0)
Average precipitation mm (inches)	55.6 (2.19)	51.8 (2.04)	34.3 (1.35)	55.5 (2.19)	42.7 (1.68)	27.8 (1.09)	16.4 (0.65)	34.4 (1.35)	80.3 (3.16)	96.8 (3.81)	66.8 (2.63)	66.7 (2.63)	629.1 (24.77)
Average precipitation days	5.5	4.4	4.7	5.7	4.9	3.6	2.4	3.6	4.6	6.8	6.1	5.6	57.8
Average snowy days	0.6	0.7	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.7	2.4
Average relative humidity (%)	75	73	68	68	70	66	63	66	72	77	75	76	70.8
Mean monthly sunshine hours	142.9	168.1	220.9	227.0	263.9	312.4	339.7	298.0	241.5	168.6	148.8	136.5	2,668.2
Source #1: Météo France													
Source #2: Infoclimat.fr (humidity and snowy days, 1961–1990)													

Target sentence: Extreme temperatures of Montpellier have ranged from −17.8 °C recorded in February and up to 37.5 °C (99.5 °F) in July.

Figure 5: ToTTo example with interesting reference language.