

Towards High Precision Text Generation

Ankur Parikh

Joint work with Thibault Sellam, Ran Tian, Xuezhi Wang, Sebastian Gehrmann, Shashi Narayan, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das.

Google Research

Text Generation

- Given a source sequence \mathbf{x} , output a target \mathbf{y}

$$\mathbf{x} = (x_1, \dots, x_S)$$

$$\mathbf{y} = (y_1, \dots, y_T)$$

- Examples:
 - **Translation:** \mathbf{x} = English sentence, \mathbf{y} = French sentence
 - **Summarization:** \mathbf{x} = Document, \mathbf{y} = One paragraph summary
 - **Data to Text:** \mathbf{x} = Structured data, \mathbf{y} = Textual description.

Encoder Decoder Paradigm

$\mathbf{x} = (x_1, \dots, x_S)$ source text

$\mathbf{y} = (y_1, \dots, y_T)$ target text

$$P(\mathbf{y}|\mathbf{x}) \approx P(\mathbf{y}|\mathbf{c}) \quad \text{where } \mathbf{c} = F_{enc}(\mathbf{x})$$

encoding function

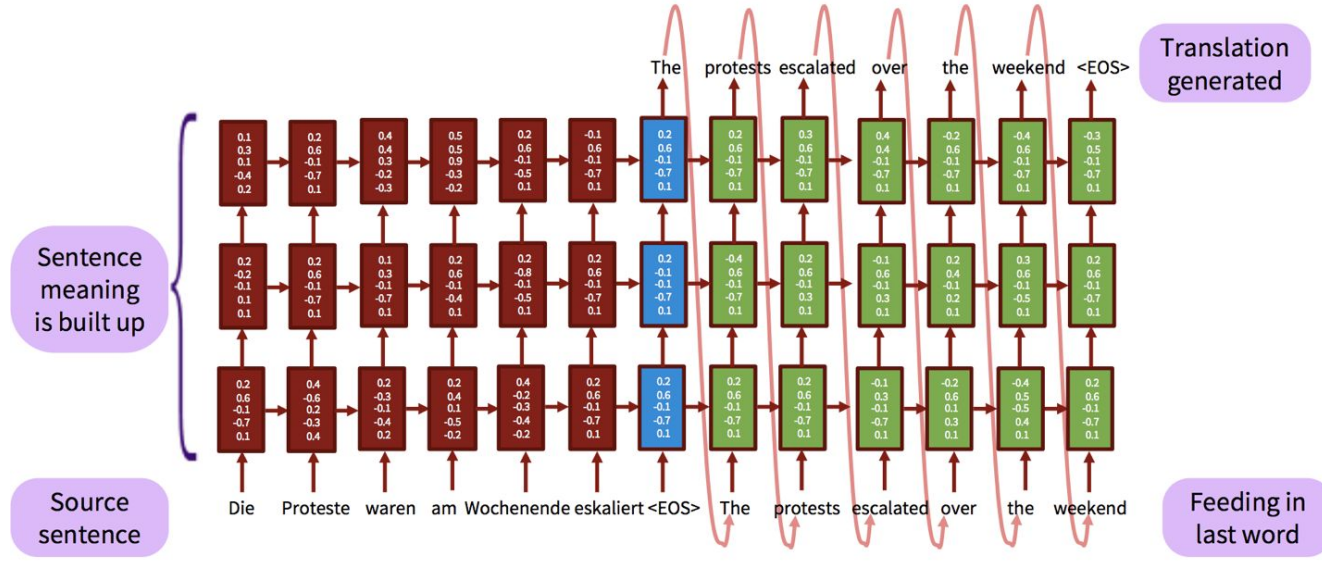
$$P(\mathbf{y}|\mathbf{c}) = \prod_{t=1}^T P(\mathbf{y}_t | \mathbf{y}_1, \dots, \mathbf{y}_{t-1}, \mathbf{c})$$

language model conditioned on the
source encoding

Encoder = some neural
network

Decoder (conditional language
model) = some neural network

Encoder Decoder Paradigm



[Sutskever et al. 2014, Bahdanau et al. 2014]

Neural Text Generation

- Neural text generation is surprisingly fluent but prone to **hallucination** which makes them unsuitable for many real world applications.
- Challenges are multifaceted:
 - Models
 - Data
 - Evaluation

Neural Text Generation

- Neural text generation is surprisingly fluent but prone to **hallucination** which makes them unsuitable for many real world applications.
- Challenges are multifaceted:
 - Models
 - Data
 - Evaluation

Motivation - Hallucination

- Neural generation models often state phrases that are unsupported or contradictory to the source data.

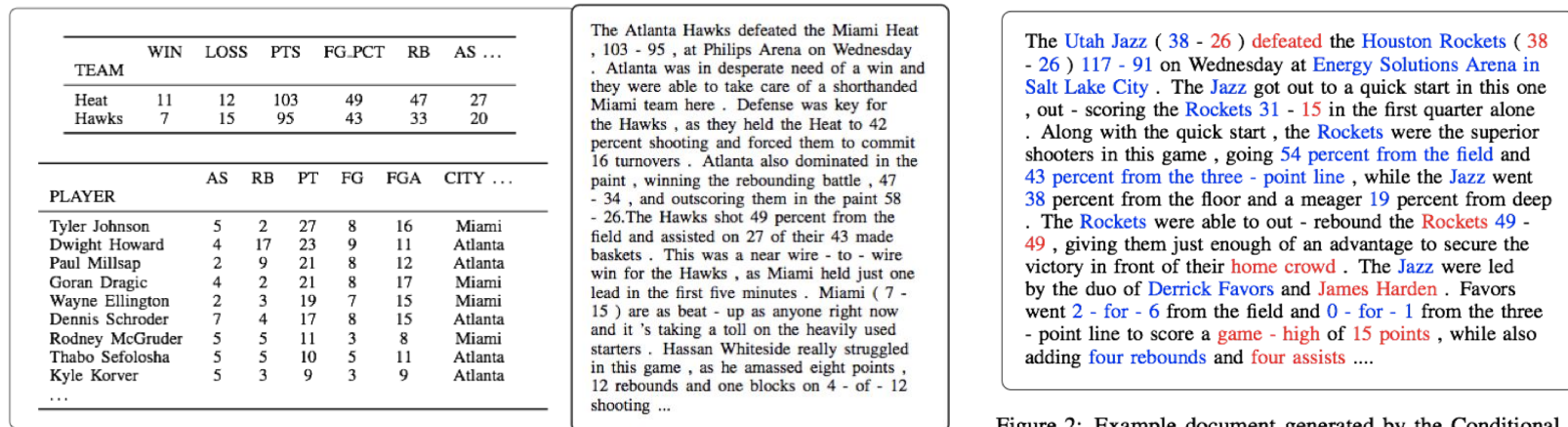


Figure 1: An example data-record and document pair from the ROTOWIRE dataset. We show a subset of the game's records (there are 628 in total), and a selection from the gold document. The document mentions only a select subset of the records, but may express them in a complicated manner. In addition to capturing the writing style, a generation system should select similar record content, express it clearly, and order it appropriately.

Figure 2: Example document generated by the Conditional Copy system with a beam of size 5. Text that accurately reflects a record in the associated box- or line-score is highlighted in blue, and erroneous text is highlighted in red.

Motivation - Hallucination

Source (Wikipedia infobox):

Frank Lino

FBI surveillance photo

Birth date October 30, 1938

Birth place Gravesend, Brooklyn, New York,
United States

https://en.wikipedia.org/wiki/Frank_Lino

Target:

Neural baseline: Frank Lino (born October 30, 1938 in Brooklyn, New York, United States) is an American criminal defense attorney.

Causes of Hallucination (Data)

- In some datasets, the target contains information that cannot be inferred by the source due to heuristic data collection.
- This makes it unclear if hallucination is caused by modeling weaknesses or data noise.

Frank Lino

FBI surveillance photo

Birth date October 30, 1938

Birth place Gravesend, Brooklyn, New York,
United States

Reference: Frank ``Curly " Lino (born October 30, 1938 Brooklyn) is a Sicilian-American Caporegime in the Bonanno crime family who later became an informant.

WIKIBIO [[Lebret et al. 2016](#)]

Causes of Hallucination (Evaluation)

BLEU prefers coverage/fluency over precision.

Michael Dahlquist			BLEU
Born	December 22, 1965 Seattle, Washington	Reference: Michael Dahlquist (December 22 , 1965 -- July 14 , 2005) was a drummer in the Seattle band Silkworm.	
Died	July 14, 2005 (aged 39) Skokie, Illinois	Candidate 1: Michael Dahlquist (December 22 , 1965 -- July 14 , 2005) was a drummer in the California band Grateful dead.	0.79
Genres	Male	Candidate 2: Michael Dahlquist (December 22 , 1965 -- July 14 , 2005) was a drummer.	0.71
Occupation	Drummer	Candidate 3: Michael Dahlquist (December 22 , 1965 -- July 14 , 2005) was a drummer from Seattle Washington.	0.73
Instruments	Drums		

https://en.wikipedia.org/wiki/Michael_Dahlquist

Causes of Hallucination (Models)

Loss functions like maximum likelihood are often not suitable for stopping hallucination.

$$\mathcal{L}(\theta) = \sum_{(\mathbf{x}, \mathbf{y}) \in \mathcal{D}} p(\mathbf{y} | \mathbf{x}) = \sum_{(\mathbf{x}, \mathbf{y}) \in \mathcal{D}} \prod_{t=1}^T p(\mathbf{y}_t | \mathbf{y}_{1:t-1}, \mathbf{x})$$

Our work

Data

- ToTTo: A Controlled Table-to-Text Generation Dataset

Evaluation

- Handling Divergent Reference Texts when Evaluating Table-to-Text Generation
- BLEURT: Learning Robust Metrics for Text Generation
- Learning to Evaluate Translation Beyond English: BLEURT Submissions to the WMT Metrics 2020 Shared Task

Models

- Text Generation with Exemplar-based Adaptive Decoding
- Sticking to the Facts: Confident Decoding for Faithful Data-to-Text Generation

Machine Translation

- Consistency by Agreement in Zero-shot Neural Machine Translation
- A Multilingual View of Unsupervised Machine Translation

Focus of this Talk

Data

- ToTTo: A Controlled Table-to-Text Generation Dataset

Evaluation

- Handling Divergent Reference Texts when Evaluating Table-to-Text Generation
- BLEURT: Learning Robust Metrics for Text Generation
- Learning to Evaluate Translation Beyond English: BLEURT Submissions to the WMT Metrics 2020 Shared Task

Models

- Text Generation with Exemplar-based Adaptive Decoding
- Sticking to the Facts: Confident Decoding for Faithful Data-to-Text Generation

Machine Translation

- Consistency by Agreement in Zero-shot Neural Machine Translation
- A Multilingual View of Unsupervised Machine Translation

Focus of this Talk

Data

- ToTTo: A Controlled Table-to-Text Generation Dataset

Evaluation

- Handling Divergent Reference Texts when Evaluating Table-to-Text Generation
- BLEURT: Learning Robust Metrics for Text Generation
- Learning to Evaluate Translation Beyond English: BLEURT Submissions to the WMT Metrics 2020 Shared Task

Models

- Text Generation with Exemplar-based Adaptive Decoding
- Sticking to the Facts: Confident Decoding for Faithful Data-to-Text Generation

Machine Translation

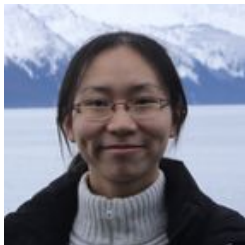
- Consistency by Agreement in Zero-shot Neural Machine Translation
- A Multilingual View of Unsupervised Machine Translation

ToTTo: A Controlled Table-To-Text Generation Dataset

EMNLP 2020



Xuezhi Wang



Sebastian
Gehrmann



Manaal Faruqui



Bhuwan Dhingra



Diyi Yang



Dipanjan Das

**TL;DR: Novel data-to-text dataset with clean references over Wikipedia.
Shows that cleaning data does not stop hallucination.**

<https://github.com/google-research-datasets/ToTTo>

Google Research

Overview - Controlled Generation Task

120K training examples

Source: Table, metadata, set of highlighted cells

Target: One sentence description

Table Title: Robert Craig (American football)

Section Title: National Football League statistics

Table Description:None

YEAR	TEAM	Rushing					Receiving				
		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

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

Overview - Controlled Generation Task

120K training examples

Source: Table, metadata, set of highlighted cells

Target: One sentence description

Table Title: Robert **Craig** (American football)

Section Title: **National Football League** statistics

Table Description: None

YEAR	TEAM	Rushing					Receiving				
		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

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

Overview - Controlled Generation Task

120K training examples

Source: Table, metadata, set of highlighted cells

Target: One sentence description

Table Title: Robert Craig (American football)

Section Title: National Football League statistics

Table Description:None

YEAR	TEAM	Rushing					Receiving				
		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

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

Overview - Controlled Generation Task

120K training examples

Source: Table, metadata, set of highlighted cells

Target: One sentence description

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

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

Overview - Controlled Generation Task

120K training examples

Source: Table, metadata, set of highlighted cells

Target: One sentence description

Table Title: Robert Craig (American football)

Section Title: National Football League statistics

Table Description:None

YEAR	TEAM	Rushing					Receiving				
		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

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

Causes of Hallucination (Data)

- In some datasets, the target contains information that cannot be inferred by the source due to heuristic data collection.
- This makes it unclear if hallucination is caused by modeling weaknesses or data noise.

Frank Lino

FBI surveillance photo

Birth date October 30, 1938

Birth place Gravesend, Brooklyn, New York,
United States

Reference: Frank ``Curly " Lino (born October 30, 1938 Brooklyn) is a Sicilian-American Caporegime in the Bonanno crime family who later became an informant.

WIKIBIO [[Lebret et al. 2016](#)]

Motivation - Data Noise

- In some datasets, the target contains information that cannot be inferred by the source due to heuristic data collection.
- This makes it unclear if hallucination is caused by modeling weaknesses or data noise.
- Asking annotators to write sentences from scratch typically results in vanilla targets that lack variety [[Gururangan et al. 2018](#), [Poliak et al. 2018](#)]

Frank Lino

FBI surveillance photo

Birth date October 30, 1938

Birth place Gravesend, Brooklyn, New York,
United States

Reference: Frank ``Curly '' Lino (born October 30, 1938 Brooklyn) is a Sicilian-American Caporegime in the Bonanno crime family who later became an informant.

WIKIBIO [[Lebret et al. 2016](#)]

Motivation - Task Definition

- Many tasks are defined as summarization which can be difficult to evaluate.
- On the other hand, tasks that are limited to verbalizing a fully specified meaning representation may not sufficiently challenge modern neural networks [[Gardent et al. 2017](#)]

TEAM	WIN	LOSS	PTS	FG.PCT	RB	AS ...
Heat	11	12	103	49	47	27
Hawks	7	15	95	43	33	20

PLAYER	AS	RB	PT	FG	FGA	CITY ...
Tyler Johnson	5	2	27	8	16	Miami
Dwight Howard	4	17	23	9	11	Atlanta
Paul Millsap	2	9	21	8	12	Atlanta
Goran Dragic	4	2	21	8	17	Miami
Wayne Ellington	2	3	19	7	15	Miami
Dennis Schroder	7	4	17	8	15	Atlanta
Rodney McGruder	5	5	11	3	8	Miami
Thabo Sefolosha	5	5	10	5	11	Atlanta
Kyle Korver	5	3	9	3	9	Atlanta
...						

The Atlanta Hawks defeated the Miami Heat , 103 - 95 , at Philips Arena on Wednesday . Atlanta was in desperate need of a win and they were able to take care of a shorthanded Miami team here . Defense was key for the Hawks , as they held the Heat to 42 percent shooting and forced them to commit 16 turnovers . Atlanta also dominated in the paint , winning the rebounding battle , 47 - 34 , and outscoring them in the paint 58 - 26.The Hawks shot 49 percent from the field and assisted on 27 of their 43 made baskets . This was a near wire - to - wire win for the Hawks , as Miami held just one lead in the first five minutes . Miami (7 - 15) are as beat - up as anyone right now and it 's taking a toll on the heavily used starters . Hassan Whiteside really struggled in this game , as he amassed eight points , 12 rebounds and one blocks on 4 - of - 12 shooting ...

Figure 1: An example data-record and document pair from the ROTOWIRE dataset. We show a subset of the game's records (there are 628 in total), and a selection from the gold document. The document mentions only a select subset of the records, but may express them in a complicated manner. In addition to capturing the writing style, a generation system should select similar record content, express it clearly, and order it appropriately.

The **Utah Jazz** (**38 - 26**) **defeated** the **Houston Rockets** (**38 - 26**) 117 - 91 on Wednesday at **Energy Solutions Arena in Salt Lake City** . The **Jazz** got out to a quick start in this one , out - scoring the **Rockets** **31 - 15** in the first quarter alone . Along with the quick start , the **Rockets** were the superior shooters in this game , going **54 percent from the field** and **43 percent from the three - point line** , while the **Jazz** went **38 percent from the floor** and a meager **19 percent from deep** . The **Rockets** were able to out - rebound the **Rockets** **49 - 49** , giving them just enough of an advantage to secure the victory in front of their **home crowd** . The **Jazz** were led by the duo of **Derrick Favors** and **James Harden** . Favors went **2 - for - 6** from the field and **0 - for - 1** from the three - point line to score a **game - high of 15 points** , while also adding **four rebounds** and **four assists**

Figure 2: Example document generated by the Conditional Copy system with a beam of size 5. Text that accurately reflects a record in the associated box- or line-score is highlighted in **blue**, and erroneous text is highlighted in **red**.

Our Dataset

ToTTo is novel in two ways:

- **Task Design:** “Controlled generation”: Set of highlighted cells gives guidance as to what to generate.
- **Annotation process:** Annotators iteratively revise natural sentences on Wikipedia so they are faithful to the table.
- **ToTTo** statistics:
 - 120K training examples
 - 7500 dev examples
 - 7500 test examples

(Heuristic) Data Collection

Get tables from Wikipedia.

Year ↕	Title ↕	Role ↕	Notes
2003	<i>What a Girl Wants</i>	Noelle	
2004	<i>Alfie</i>	Carol	
2005	<i>The Jacket</i>	Nurse Nina	
2005	<i>Vado a messa</i>	Frances	Short film
2006	<i>Rabbit Fever</i>	Ally	
2006	<i>Factory Girl</i>	Brigid Polk	
2008	<i>Love Lies Bleeding</i>	Det. Alice Sands	Video
2010	<i>Luster</i>	Rachel	
2010	<i>Stephany + Me</i>		Video short
2010	<i>Inspired by Bret Easton Ellis</i>		Short film
2010	<i>The Lake Effect</i>	Natalie	
2012	<i>Here Comes the Night</i>	Simone	
2012	<i>Hitchcock</i>	Rita Riggs	

[Tara Summers](#)

United States presidential election in Louisiana, 1956			
Party	Candidate	Votes	%
 Republican	Dwight D. Eisenhower (inc.)	329,047	53.28%
 Democratic	Adlai Stevenson	243,977	39.51%
 Dixiecrat	<i>Unpledged electors</i>	44,520	7.21%
	Write-in	2,503	0.24%
Total votes		617,544	100%

[1956 United States presidential election in Louisiana](#)

(Heuristic) Data Collection

Use heuristics such as word overlap or hyperlinks to find sentences that may be related to the table.

Year ↕	Competition ↕	Venue ↕	Position ↕	Event ↕	Notes ↕
Representing  Germany					
1992	World Junior Championships	Seoul, South Korea	10th (semis)	100 m	11.83
1993	European Junior Championships	San Sebastián, Spain	7th	100 m	11.74
			3rd	4×100 m relay	44.60
1994	World Junior Championships	Lisbon, Portugal	12th (semis)	100 m	11.66 (wind: +1.3 m/s)
			2nd	4×100 m relay	44.78
1995	World Championships	Gothenburg, Sweden	7th (q-finals)	100 m	11.54
			3rd	4×100 m relay	43.01

After winning the German under-23 100 m title, she was selected to run at the 1995 World Championships in Athletics both individually and in the relay.

[Gabriele Becker](#)

(Heuristic) Data Collection

Heuristic sentence is too noisy to be a generation target.

Year ↕	Competition ↕	Venue ↕	Position ↕	Event ↕	Notes ↕
Representing  Germany					
1992	World Junior Championships	Seoul, South Korea	10th (semis)	100 m	11.83
1993	European Junior Championships	San Sebastián, Spain	7th	100 m	11.74
			3rd	4×100 m relay	44.60
1994	World Junior Championships	Lisbon, Portugal	12th (semis)	100 m	11.66 (wind: +1.3 m/s)
			2nd	4×100 m relay	44.78
1995	World Championships	Gothenburg, Sweden	7th (q-finals)	100 m	11.54
			3rd	4×100 m relay	43.01

After winning the German under-23 100 m title, she was selected to run at the 1995 World Championships in Athletics both individually and in the relay.

[Gabriele Becker](#)

Annotation Process

Annotators do the following:

- Highlight cells that support sentence
- Iteratively revise the sentence so that it is faithful to the table and standalone

Table Title: Gabriele Becker
Section Title: International competitions

Year	Competition	Venue	Position	Event	Notes
Representing Germany					
1992	World Junior Championships	Seoul, South Korea	10th (semis)	100 m	11.83
1993	European Junior Championships	San Sebastián, Spain	7th	100 m	11.74
			3rd	4×100 m relay	44.60
1994	World Junior Championships	Lisbon, Portugal	12th (semis)	100 m	11.66 (wind: +1.3 m/s)
			2nd	4×100 m relay	44.78
1995	World Championships	Gothenburg, Sweden	7th (q-finals)	100 m	11.54
			3rd	4×100 m relay	43.01

Original Sentence

After winning the German under-23 100 m title, she was selected to run at the 1995 World Championships in Athletics both individually and in the relay.

After Deletion

~~After winning the German under-23 100 m title, she was selected to run at the 1995 World Championships in Athletics both individually and in the relay.~~

After

Decontextualization

Gabriele Becker competed at the 1995 World Championships in both individually and in the relay.

After Grammar

Gabriele Becker competed at the 1995 World Championships in both individually and in the relay.

Dataset Statistics

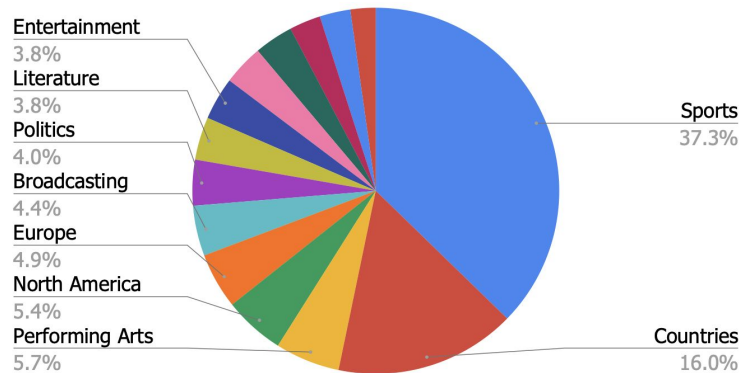
Property	Value
<i>Training set size</i>	120,761
<i>Number of target tokens</i>	1,268,268
<i>Avg Target Length (tokens)</i>	17.4
<i>Target Vocabulary size</i>	136,777
<i>Unique tables</i>	83,141
<i>Rows per table (median)</i>	16
<i>Cells per table (median)</i>	87
<i>No. of highlighted cells (median)</i>	3

Annotator Agreement

Our iterative annotation process allows us to measure annotator agreement at each stage.

Annotation Stage	BLEU-4
<i>After Deletion</i>	82.9
<i>After Decontextualization</i>	72.56
<i>Final (After Grammar)</i>	68.98

Linguistic Phenomena



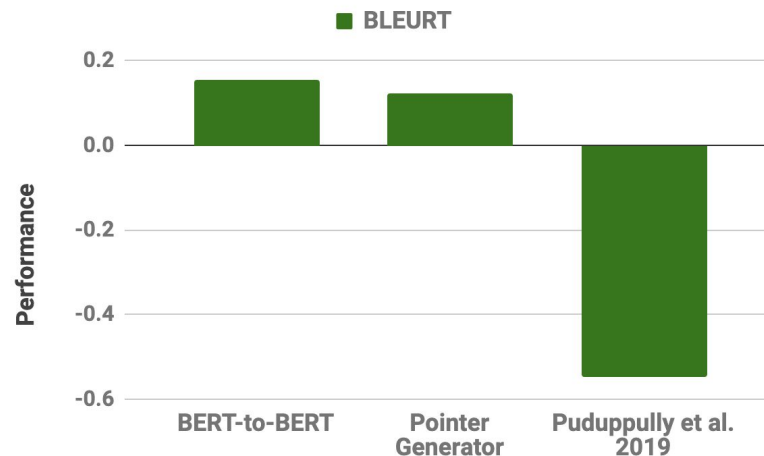
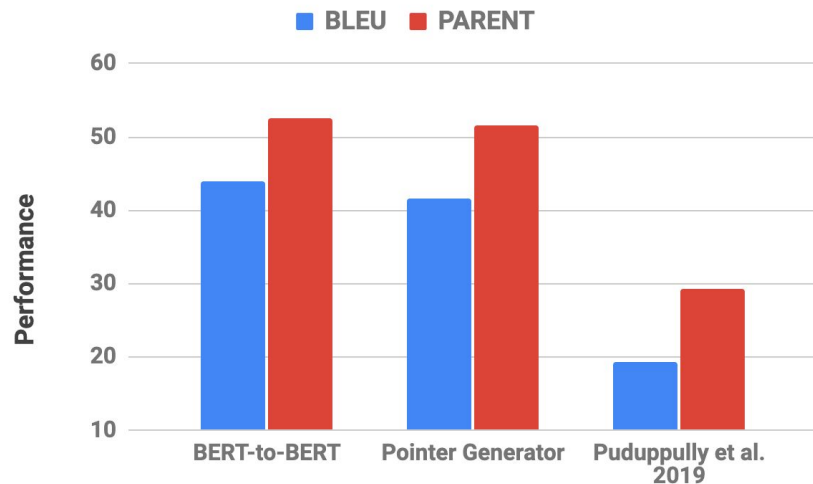
Annotation Stage	Percentage
Require reference to page title	82%
Require reference to section title	19%
Require reference to table description	3%
Reasoning (logical, numerical, temporal etc.)	21%
Comparison across rows / columns / cells	13%
Require background information	12%

Baselines

- BERT-to-BERT [[Rothe et al. 2019](#)] - BERT initialized encoder-decoder model
- Pointer Generator [[See et al. 2017](#)] - Seq2Seq with Copy mechanism
- [[Puduppully et al. 2019](#)] - Content planning mechanism for data-to-text

Baseline Results

- BLEU [[Papineni et al. 2002](#)] - n-gram metric
- PARENT [[Dhingra et al. 2019](#)] - n-gram metric for data-to-text
- BLEURT [[Sellam et al. 2020](#)] - learnt evaluation metric

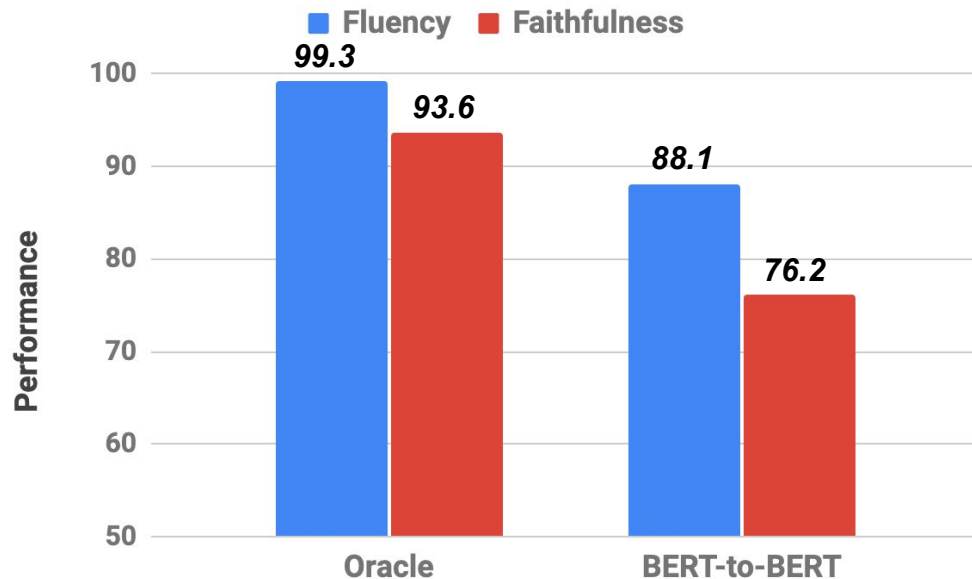


Baseline Results - Human Evaluation

Compared to the human oracle, baseline is

- Reasonably fluent
- However, considerably less faithful.

Evidence that neural models hallucinate even in the presence of clean data.



Model Shortcomings - Rare topics

Table Title: Microdrive

Section Title: Microdrive models by timeline

Table Description: Date of release of large sizes

1999:	170 megabyte, 340 megabyte (IBM)
2000:	512 megabyte, 1 gigabyte (IBM)
2003:	2 gigabytes, 4 gigabytes (Hitachi)
2004:	2.5 and 5 gigabytes (Seagate)
2005:	6 gigabytes (Hitachi), 8 gigabytes (Seagate)
2006:	8 gigabytes (Hitachi)

Prediction: there were **512 microdrive models in 2000: 1 gigabyte.**

Reference: a second generation of microdrive was announced by ibm in 2000 with increased capacities at 512 mb and 1 gb.

Model Shortcomings - Reasoning

Table Title: Travis Kelce

Section Title: Collegiate statistics

Table Description: None

Travis Kelce			Receiving			
Year	Team	G	Rec	Yds	Avg	TD
2009	Cincinnati	11	1	3	3.0	0
2011	Cincinnati	11	13	150	11.5	2
2012	Cincinnati	13	45	722	16.0	8
Career		35	59	875	14.8	10

Prediction: *kelce finished the 2012 season with 45 receptions for 722 yards (16.0 average) and eight touchdowns.*

Reference: *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).*

Summary

Try out our dataset!

<https://github.com/google-research-datasets/ToTTo>

Feel free to contact:

totto@google.com

Focus of this Talk

Data

- ToTTo: A Controlled Table-to-Text Generation Dataset

Evaluation

- Handling Divergent Reference Texts when Evaluating Table-to-Text Generation
- **BLEURT: Learning Robust Metrics for Text Generation**
- Learning to Evaluate Translation Beyond English: BLEURT Submissions to the WMT Metrics 2020 Shared Task

Models

- Text Generation with Exemplar-based Adaptive Decoding
- Sticking to the Facts: Confident Decoding for Faithful Data-to-Text Generation

Machine Translation

- Consistency by Agreement in Zero-shot Neural Machine Translation
- A Multilingual View of Unsupervised Machine Translation

BLEURT: Learning Robust Metrics for Text Generation

ACL 2020



Thibault Sellam



Dipanjan Das



TL;DR: State of the art learnt metric with fast domain/task adaptation.

<https://github.com/google-research/bleurt>

Google Research

Metrics are a Bottleneck to Generation Progress

Prediction

pete kmetovic from stanford university **took third place at stanford university.**

the 1956 grand prix motorcycle racing season consisted of **eight** grand prix races in **six** classes: 500cc, 350cc, 250cc, 125cc and sidecars 500cc.

kelce finished the 2012 season with 45 receptions for 722 yards (16.0 average) and eight touchdowns.

Reference

the eagles first pick, and third overall, was pete kmetovic, a halfback from stanford university.

the 1956 grand prix motorcycle racing season consisted of six grand prix races in five classes: 500cc, 350cc, 250cc, 125cc and sidecars 500cc.

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).

Comments

incorrect and word overlap low

incorrect but word overlap high

Correct, but reference is more informative

Learnt Metrics

- These types of task-oriented complex relationships can only be captured by learnt metrics.
- Naive learnt metric: Fine-Tune BERT on human ratings data.

BERT [Devlin et al. 2018]



+

*Human Ratings
Data*

—	—	0.1
—	—	0.7
...		
—	—	0.4

=

**Learnt
Metric**

- **Problem:** Brittle, requires lots of fine-tuning data for every new dataset/task.

BLEURT

- Additional pertaining step based on synthetic data.
- Makes model robust to train/test skew and enables fast adaptation to other domains.



- State of the art results on WMT 2017, 2018, 2019 and WebNLG

Generating Synthetic Data

- Goal is to generate pairs (x, x') that resemble reference, prediction pairs.

Reference

*I bought soy milk from the store on
Tuesday.*

Predictions

*I bought ~~soy~~ milk from the store on
Tuesday.*

*I bought soy milk from the store on
Wednesday.*

*I bought soy milk from the grocery store
on Tuesday.*

*I bought soy milk from the grocery store
on Tuesday on Tuesday on Tuesday.*

- Model errors typically do not resemble existing paraphrasing datasets.
- Generate synthetic data instead.

Generating Synthetic Data

- Strategy 1: Randomly mask text and employ BERT to fill masks

*I bought [MASK] milk from [MASK] store
on Tuesday.*



*I bought **almond** milk from **a** store on Tuesday.*

*I [MASK] [MASK] [MASK] from the
store on Tuesday.*



*I **stole some oranges** from the store on Tuesday.*

- Strategy 2: Use back-translation.

*I bought soy milk from the store
on Tuesday.*



*Mardi, j'ai acheté du lait de soja au
magasin.*



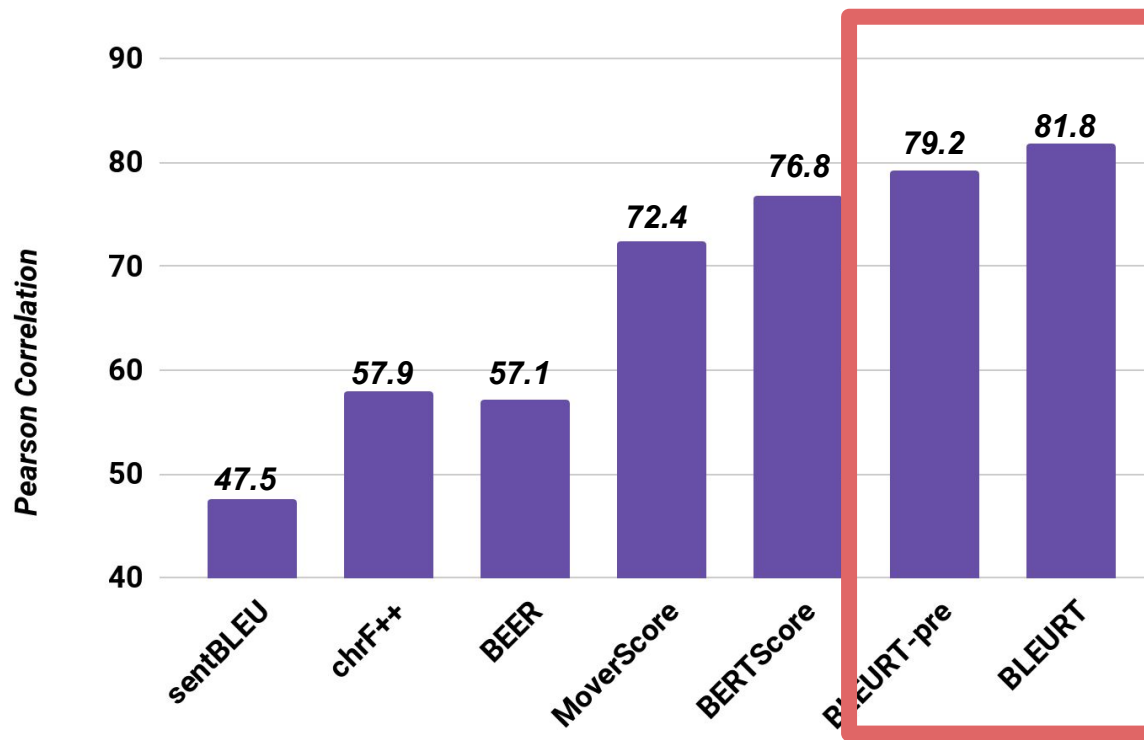
*Tuesday I bought soy milk
at the store.*

Weak Supervision Signals

- Use a variety of easily computed metrics for a multi-task objective

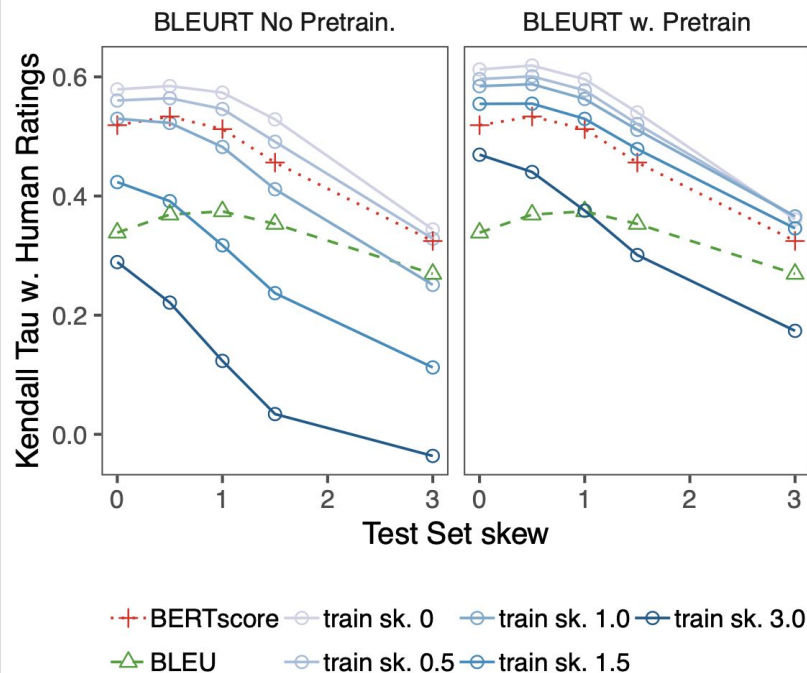
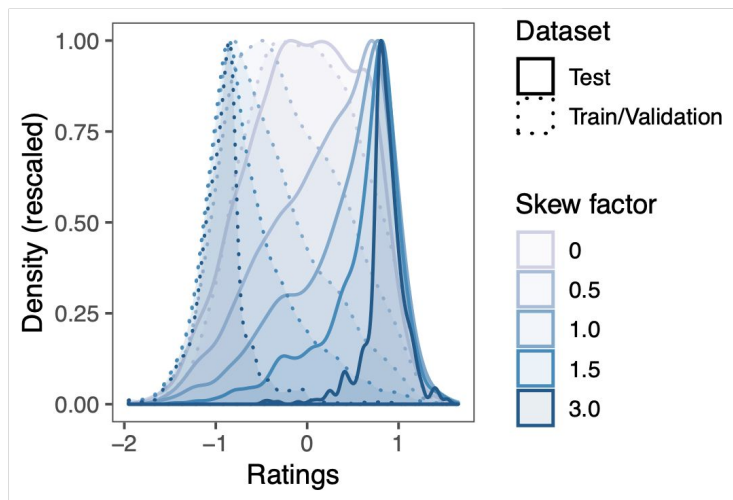
$$\left[\begin{array}{c} BLEU(x, x') \\ ROUGE(x, x') \\ BERT-SCORE(x, x') \\ Entail(x, x') \\ \cdot \\ \cdot \\ \cdot \\ BackTransProb(x, x') \end{array} \right]$$

BLEURT Results - WMT2017

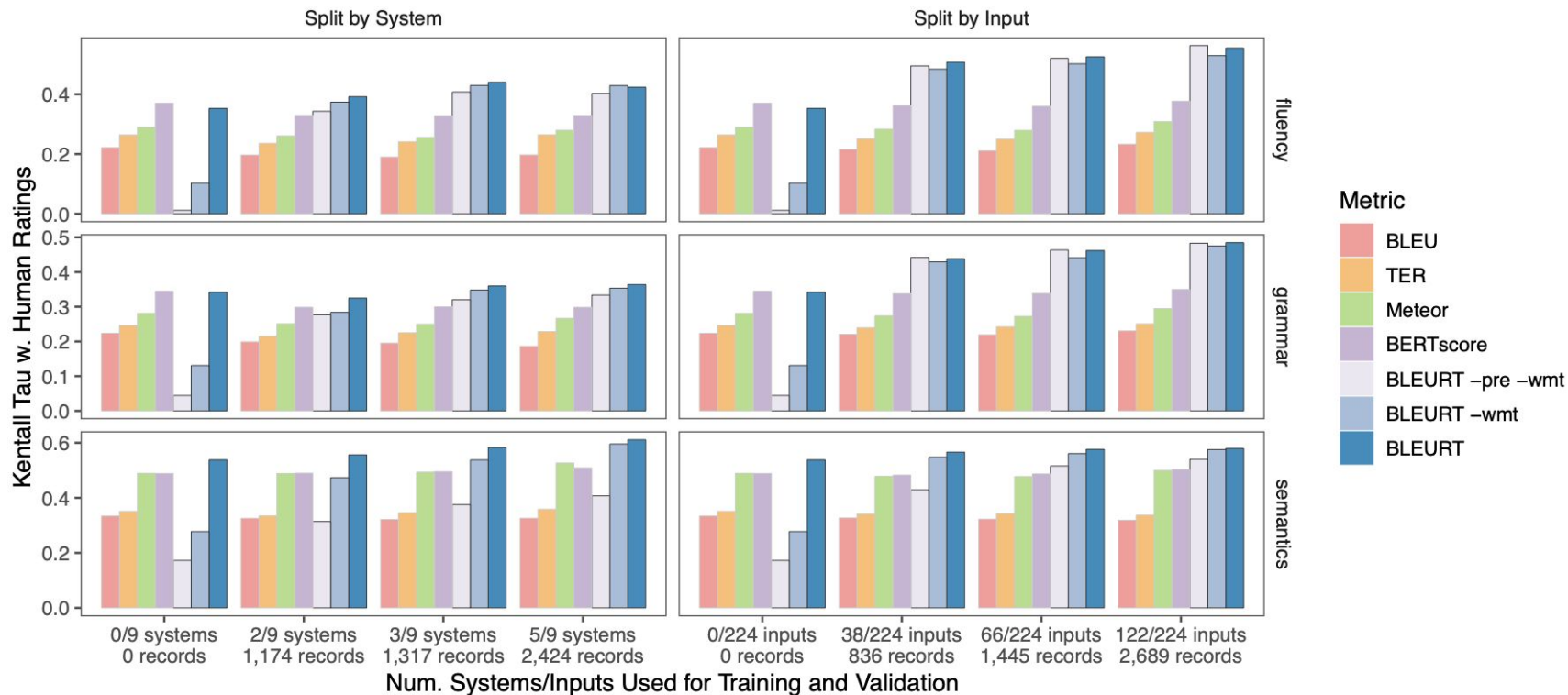


Robustness

- Skew training set toward lower ratings, Skew test toward higher ratings
- Synthetic pretraining makes BLEURT more robust.



WebNLG Results



Focus of this Talk

Data

- ToTTo: A Controlled Table-to-Text Generation Dataset

Evaluation

- Handling Divergent Reference Texts when Evaluating Table-to-Text Generation
- BLEURT: Learning Robust Metrics for Text Generation
- Learning to Evaluate Translation Beyond English: BLEURT Submissions to the WMT Metrics 2020 Shared Task

Models

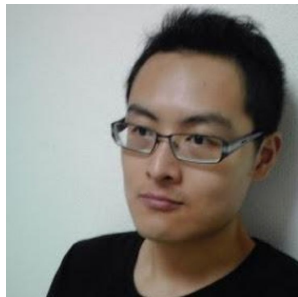
- Text Generation with Exemplar-based Adaptive Decoding
- **Sticking to the Facts: Confident Decoding for Faithful Data-to-Text Generation**

Machine Translation

- Consistency by Agreement in Zero-shot Neural Machine Translation
- A Multilingual View of Unsupervised Machine Translation

Sticking to the Facts: Confident Decoding for Faithful Data-to-Text Generation

[arXiv](#)



Ran Tian



Shashi Narayan



Thibault Sellam



TL;DR: Learnt per-token confidence score that reduces hallucination by 50% on Wikibio dataset

Reducing Hallucination

Source (Wikipedia infobox):

Frank Lino

FBI surveillance photo

Birth date October 30, 1938

Birth place Gravesend, Brooklyn, New York,
United States

https://en.wikipedia.org/wiki/Frank_Lino

Target:

Reference: Frank ``Curly '' Lino (born October 30, 1938 Brooklyn) is a Sicilian-American Caporegime in the Bonanno crime family who later became an informant.

Baseline ([See et al. 2017](#)): Frank Lino (born October 30, 1938 in Brooklyn, New York, United States) is an American criminal defense attorney.

Our model ([Tian et al. 2019](#)): Frank Lino (born October 30, 1938 in Brooklyn, New York, United States) is an American.

Intuition

Michael Eric Dyson

Name	Michael Eric Dyson
Birth date	23 October 1958
Birth place	Detroit, Michigan, USA
Nationality	United States
Education	Knoxville College

Michael Eric Dyson (born October 23, 1958) is an American academic, author and radio host.

Intuition

Michael Eric Dyson

Name	Michael Eric Dyson
Birth date	23 October 1958
Birth place	Detroit, Michigan, USA
Nationality	United States
Education	Knoxville College

Michael Eric Dyson (born October 23, 1958) is an American academic, author and radio host.

Templatic words do not convey source information and can be generated by an unconditioned language model

Intuition

Michael Eric Dyson

Name	Michael Eric Dyson
Birth date	23 October 1958
Birth place	Detroit, Michigan, USA
Nationality	United States
Education	Knoxville College

Michael Eric Dyson (born October 23, 1958) is an American academic, author and radio host.

Faithful content words require attention to the source.

Intuition

Michael Eric Dyson

Name	Michael Eric Dyson
Birth date	23 October 1958
Birth place	Detroit, Michigan, USA
Nationality	United States
Education	Knoxville College

Michael Eric Dyson (born October 23, 1958) is an American academic, author and radio host.

Hallucinations are typically content words that are not closely associated with the source.

Confidence Score

Michael Eric Dyson

Name	Michael Eric Dyson
Birth date	23 October 1958
Birth place	Detroit, Michigan, USA
Nationality	United States
Education	Knoxville College

Michael Eric Dyson (born October 23, 1958) is an American academic, author and radio host.

Confidence score should high if the token is:

- a templatic word (blue)
- *or*
- Is supported by the source (green)

Confidence score should be low if:

- not a templatic word *and* not supported by the source.

Confidence Score - Determining if Word is Templatic

Michael Eric Dyson

Name	Michael Eric Dyson
Birth date	23 October 1958
Birth place	Detroit, Michigan, USA
Nationality	United States
Education	Knoxville College

Michael Eric Dyson (born October 23, 1958) is an American academic, author and radio host.

Word is considered templatic if it has high probability under a base language model that is not conditioned on the source:

$$P_B(y_t \mid \mathbf{y}_{<t})$$

Confidence Score - Determining if Word is Supported by the Source

Michael Eric Dyson

Name	Michael Eric Dyson
Birth date	23 October 1958
Birth place	Detroit, Michigan, USA
Nationality	United States
Education	Knoxville College

Michael Eric Dyson (born *October 23, 1958*) is an *American* academic, author and radio host.


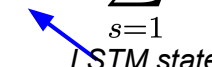
Word is supported by the source if it has a high attention score:

$$A_t := \frac{\|a_t\|}{\frac{1}{2}(\|a_t\| + \|h_t\| + \|v_t\|)}$$

where:

$$P(y_t \mid \mathbf{y}_{<t}, \mathbf{x}) = \frac{\exp(\mathbf{v}_t^\top \mathbf{e}_{y_t})}{\sum_{y \in V} \exp(\mathbf{v}_t^\top \mathbf{e}_y)}$$

$$\mathbf{v}_t = \mathbf{a}_t + \mathbf{h}_t = \sum_{s=1}^S \alpha_{s,t} \mathbf{s}_s + \mathbf{h}_t$$

 attention vector
 LSTM state

Confidence Score - Determining if Word is Supported by the Source

Michael Eric Dyson

Name	Michael Eric Dyson
Birth date	23 October 1958
Birth place	Detroit, Michigan, USA
Nationality	United States
Education	Knoxville College

Michael Eric Dyson (born October 23, 1958) is an American academic, author and radio host.

Confidence score should high if the token is:

- a templatic word (blue)
- or*
- Is supported by the source (green)

In math:

$$C_t(y_t) := \underbrace{A_t} + \underbrace{(1 - A_t)P_B(y_t \mid \mathbf{y}_{<t})}$$

Using and Learning the Confidence Score

Confidence score depends on learned parameters.

- How do we train it jointly with the rest of the model?
- How do we use it at inference time?

$$C_t(y_t) := A_t + (1 - A_t)P_B(y_t \mid \mathbf{y}_{<t})$$

Learning Confidence Score in Training

If we knew the confidence score before training the other model parameters then we could use it to clean noisy references to remove unsupported phrases.

Michael Eric Dyson

Name	Michael Eric Dyson
Birth date	23 October 1958
Birth place	Detroit, Michigan, USA
Nationality	United States
Education	Knoxville College

Michael Eric Dyson (born October 23, 1958) is an American academic, author and radio host.

However, confidence score is being jointly learned with the rest of the parameters.....

Learning Confidence Score in Training

Use a variational bayes objective to subsample confidence subsequences in training while learning confidence score.

x:

Salome Jens

Jens in 1962

Birth date May 8, 1935

Birth place Milwaukee,
Wisconsin, U.S.

Occupation Actress

Years active 1956 -- present

Confidence Score:



Random Sub-sequence

z: salome jens (born may 8 , 1935) is an american and television actress $P(z | x)$

$Q(z | y, x)$

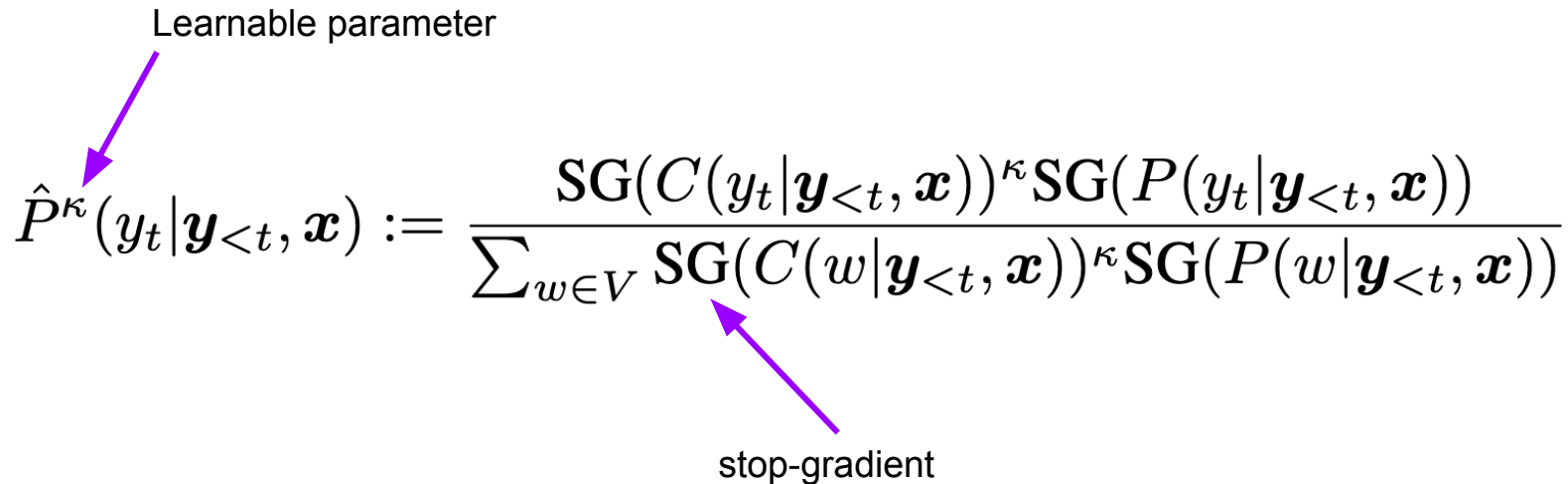
Confidence Score at Test Time

Based on [[Braverman et al. 2018](#)]. Add a one parameter model to the likelihood to re-calibrate output probabilities with the confidence score.

Learnable parameter

$$\hat{P}^{\kappa}(y_t | \mathbf{y}_{<t}, \mathbf{x}) := \frac{\text{SG}(C(y_t | \mathbf{y}_{<t}, \mathbf{x}))^{\kappa} \text{SG}(P(y_t | \mathbf{y}_{<t}, \mathbf{x}))}{\sum_{w \in V} \text{SG}(C(w | \mathbf{y}_{<t}, \mathbf{x}))^{\kappa} \text{SG}(P(w | \mathbf{y}_{<t}, \mathbf{x}))}$$

stop-gradient



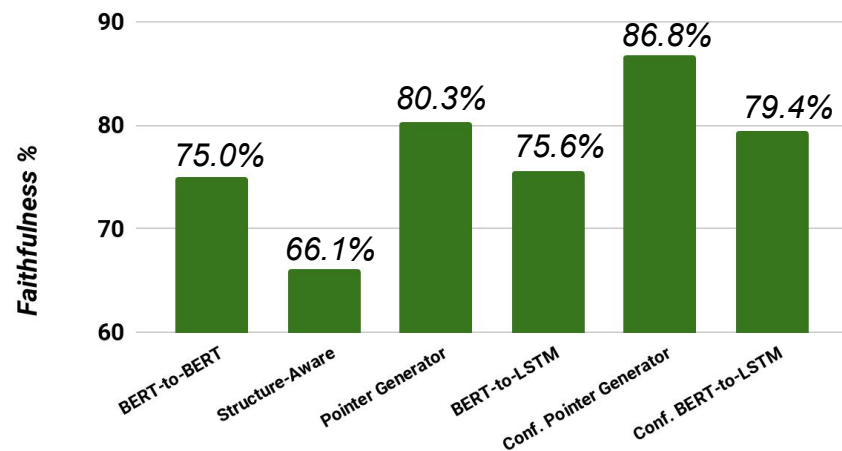
Experimental Results

- Datasets
 - Wikibio [[Lebret et al. 2016](#)]
 - WebNLG [[Gardent et al. 2017](#)]
- Models (with and without *confident decoding*)
 - Pointer Generator
 - BERT encoder + LSTM decoder
- Evaluation metrics
 - BLEU
 - PARENT [[Dhingra et al. 2019](#)]
 - Human evaluation

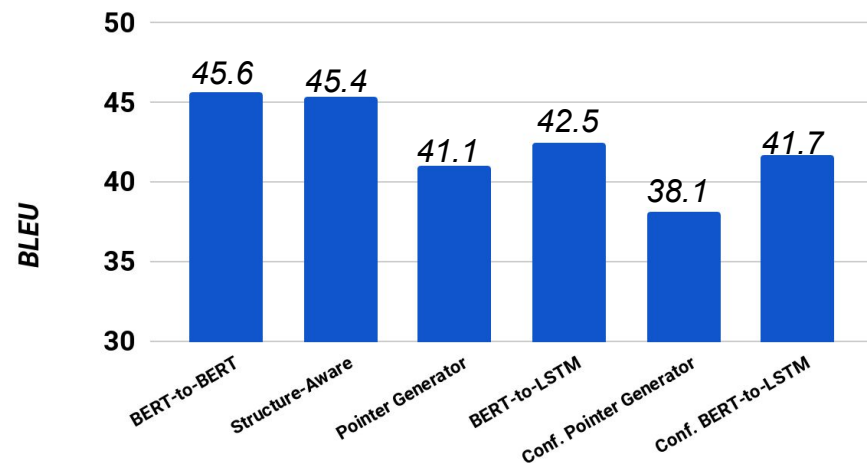
Results - Wikibio Dataset

- Confidence decoding significantly reduces hallucination. BLEU doesn't capture faithfulness

Faithfulness



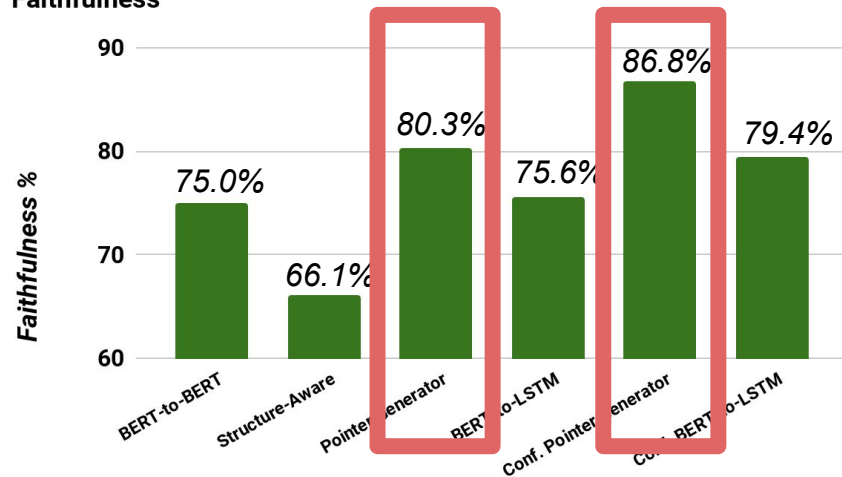
BLEU



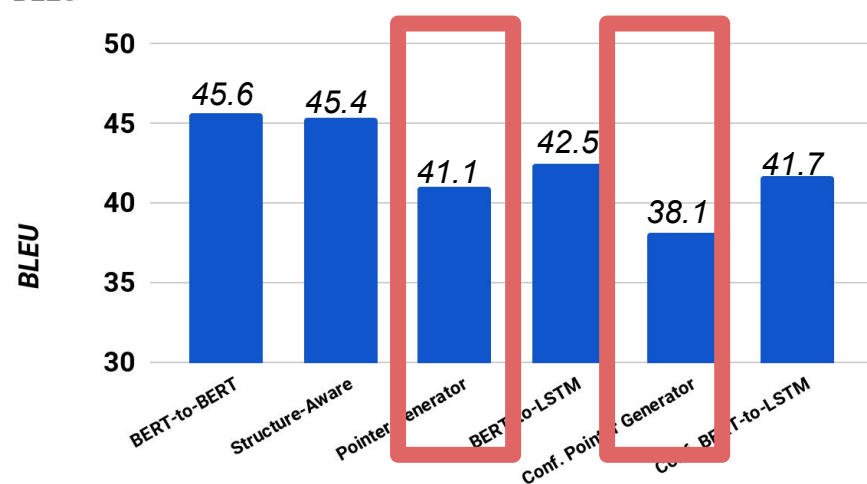
Results - Wikibio Dataset

- Confidence decoding significantly reduces hallucination. BLEU doesn't capture faithfulness

Faithfulness



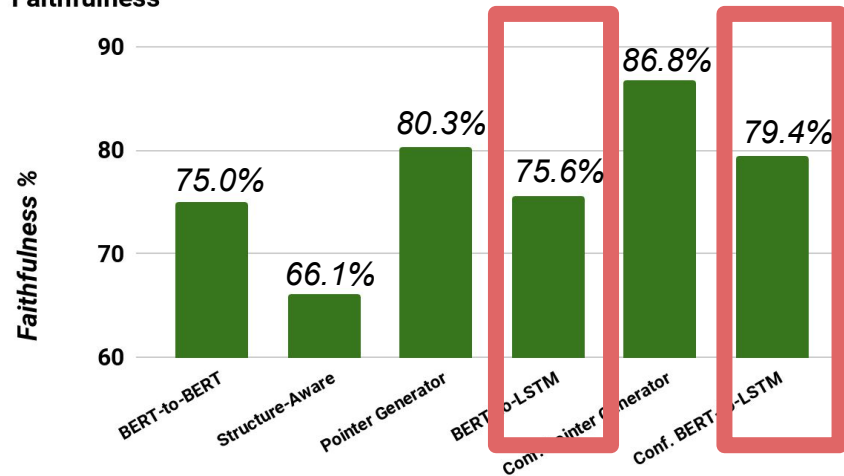
BLEU



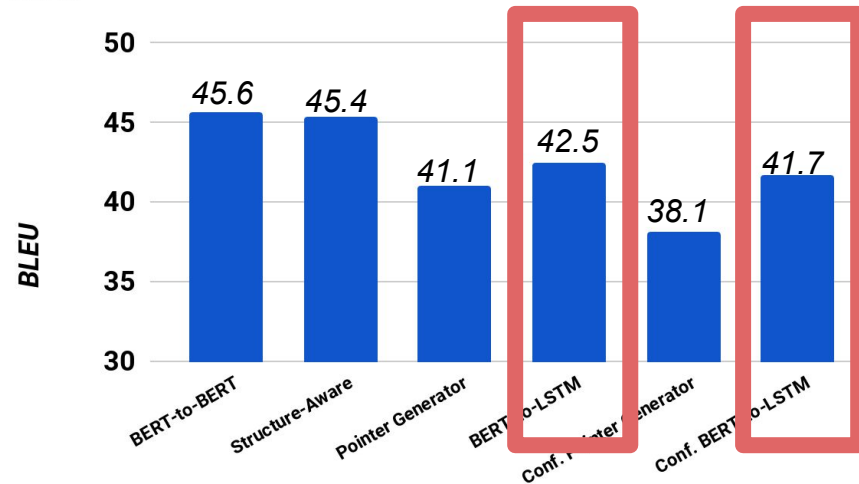
Results - Wikibio Dataset

- Confidence decoding significantly reduces hallucination. BLEU doesn't capture faithfulness

Faithfulness

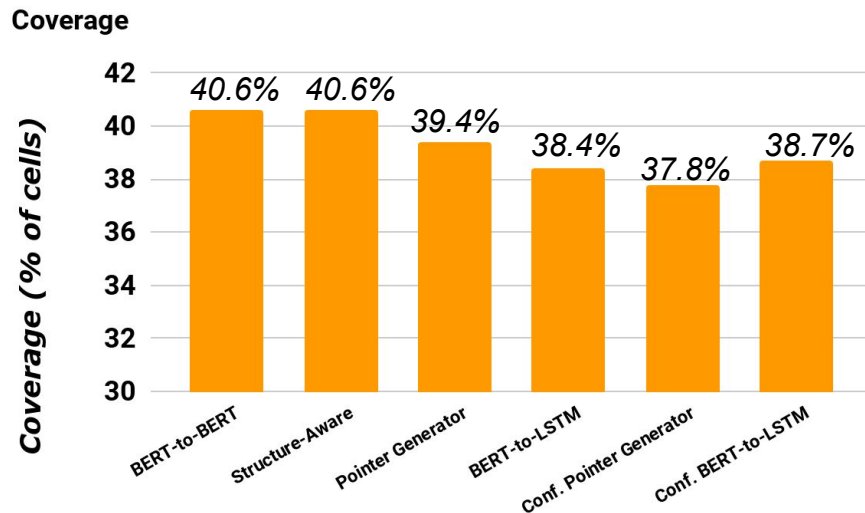
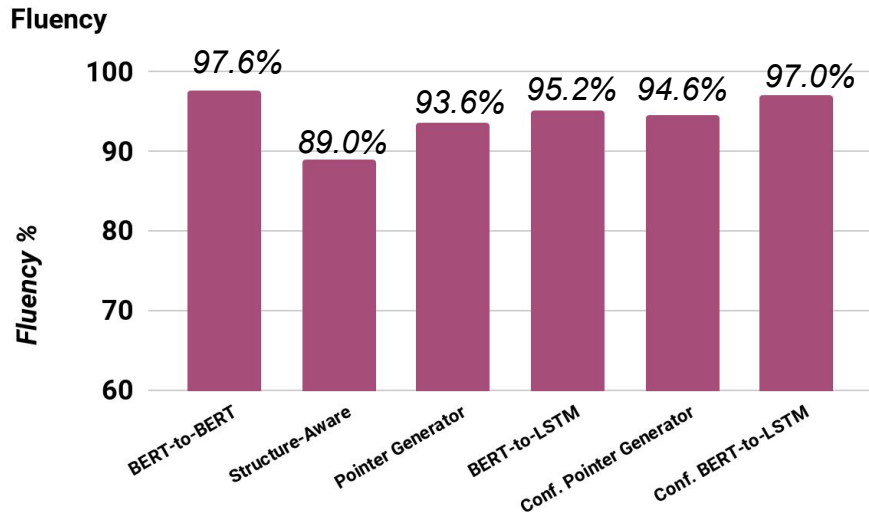


BLEU



Results - Wikibio Dataset

- Fluency is similar across methods
- Confidence decoding leads to minor coverage drop (can be increased using length penalty but at a cost to fluency)



Focus of this Talk

Data

- ToTTo: A Controlled Table-to-Text Generation Dataset

Evaluation

- Handling Divergent Reference Texts when Evaluating Table-to-Text Generation
- BLEURT: Learning Robust Metrics for Text Generation
- Learning to Evaluate Translation Beyond English: BLEURT Submissions to the WMT Metrics 2020 Shared Task

Models

- Text Generation with Exemplar-based Adaptive Decoding
- Sticking to the Facts: Confident Decoding for Faithful Data-to-Text Generation

Machine Translation

- Consistency by Agreement in Zero-shot Neural Machine Translation
- A Multilingual View of Unsupervised Machine Translation

Conclusion

- In this talk, we presented a multi-faceted approach to tackling hallucination in text generation models from the perspectives of data, evaluation, and models.
- Some links:
 - ToTTo Dataset: <https://github.com/google-research-datasets/ToTTo>
 - BLEURT: <https://github.com/google-research/bleurt>

Thank you!

