FALL 2020 CS 395T

TEXAS
The University of Texas at Austin

Data-to-text Generation

CS 395T: Topics in Natural Language Processing

17th September 2020

Uma bharathi Govindarajan, Xuewen Yao, The University of Texas at Austin

Agenda

- Introduction & Motivation
- Rule Based Approaches
- Paper 1 : Challenges in Data-to-Document Generation
- Paper 2 : ToTTo: A Controlled Table-To-Text Generation Dataset
- Conclusion
- Discussion

Challenges in Data-to-Document Generation
A Controlled Table-To-Text Generation Dataset

Introduction & Motivation

- What to say ?
  - Content Selection

- How to say it (Surface Realization)?
  - Generation
  - Content Ordering

Overall objective

- Evaluation methods & Challenges
- Task Design and Annotation Process

Previous Rule Based Approaches

- Domain Specific (Reiter et al.(2005))
  - Rules for Sentence planning to select appropriate time phrases, special grammar rules to emulate the domain language of interest based on corpus (Document planning, Microplanning, Surface Realisation)

Table 1
Part of an input data set for SumTime-Mousam

Time	Wind dir	Wind speed 10 m	Wind speed 50 m	Gust 10 m	Gust 50 m
06:00	W	10.0	12.0	12.0	16.0
09:00	W	11.0	14.0	14.0	17.0
12:00	WSW	10.0	12.0	12.0	16.0
15:00	SW	7.0	9.0	9.0	11.0

Section 2. FORECAST 6–24 GMT, Wed 12–Jun 2002

Field	Text
WIND(KTS) 10 M	W 8–13 backing SW by mid afternoon and S 10–15 by midnight.
WIND(KTS) 50 M	W 10–15 backing SW by mid afternoon and S 13–18 by midnight.
WAVES(M) SIG HT	0.5–1.0 mainly SW swell.

SUMTIME-MOUSAM(2005)

- Data driven (Regina et al.(2005))
  - Anchor-based alignment technique to obtain records-to-text alignments, used as training data (records present in the text are positive labels, and all other records negative)

Passing

PLAYER	CP/AT	YDS	AVG	TD	INT
Brunell	17/38	192	6.0	0	0
Garcia	14/21	195	9.3	1	0
…	…	…	…	…	…

Rushing

PLAYER	REC	YDS	AVG	LG	TD
Suggs	22	82	3.7	25	1
…	…	…	…	…	…

Fumbles

PLAYER	FUM	LOST	REC	YDS
Coles	1	1	0	0
Portis	1	1	0	0
Davis	0	0	1	0
Little	0	0	1	0
…	…	…	…	…

Suggs rushed for 82 yards and scored a touchdown in the fourth quarter, leading the Browns to a 17-13 win over the Washington Redskins on Sunday. **Jeff Garcia went 14-of-21 for 195 yards and a TD** for the Browns, who didn't secure the win until **Coles fumbled** with 2:08 left. The Redskins (1-3) can pin their third straight loss on going just 1-for-11 on third downs, mental mistakes and **a costly fumble by Clinton Portis. Brunell finished 17-of-38 for 192 yards,** but was unable to get into any rhythm because Cleveland's defense shut down Portis. The Browns faked a field goal, but holder Derrick Frost was stopped short of a first down. **Brunell then completed a 13-yard pass to Coles, who fumbled** as he was being taken down and Browns safety Earl Little recovered.

Table 1: Sample target game description and example of database entries; boldface indicates correspondences between the text and the database (CP/AT: completed out of attempted, YDS: yards, AVG: average, TD: touchdown, INT: interception, REC: received, LG: longest gain, FUM: fumble).

Collective Content Selection

Discriminative decisions(Angeli et al.(2010))

- Determine which record to summarize
- Determine which fields of the record
- Determine which words to use to describe the chosen fields.
- Decision is implemented as a log-linear model with features learned from training data.
- The surface realization component performs decisions based on automatically extracted templates that are filtered with domain-specific constraints in order to guarantee fluent output.

A Simple Domain-Independent Probabilistic Approach to Generation

World state

skyCover$_1$: skyCover(time=5pm-6am,mode=50-75)
temperature$_1$: temperature(time=5pm-6am,min=44,mean=49,max=60)
...

Decisions

Record	$r_1 = \text{skyCover}_1$	$r_2 = \text{temperature}_1$	$r_3 = \text{STOP}$
Field set	$F_1 = \{\text{mode}\}$	$F_2 = \{\text{time}, \text{min}\}$	
Template	$T_1 = \langle \textit{mostly cloudy ,} \rangle$	$T_2 = \langle \textit{with a low around } [\text{min}] \textit{ .} \rangle$	

Text *mostly cloudy , with a low around 45 .*

Specific active (nonzero) features for highlighted decisions

$r_2 = \text{temperature}_1$

(R1) $[\![r_2.t = \text{temperature and } (r_1.t, r_0.t) = (\text{skyCover}, \text{START})]\!]$
$[\![r_2.t = \text{temperature and } (r_1.t) = (\text{skyCover})]\!]$
(R2) $[\![r_2.t = \text{temperature and } \{r_1.t\} = \{\text{skyCover}\}]\!]$
(R3) $[\![r_2.t = \text{temperature and } r_j.t \neq \text{temperature } \forall j < 2]\!]$
(R4) $[\![r_2.t = \text{temperature and } r_2.v[\text{time}] = \texttt{5pm-6am}]\!]$
$[\![r_2.t = \text{temperature and } r_2.v[\text{min}] = \texttt{low}]\!]$
$[\![r_2.t = \text{temperature and } r_2.v[\text{mean}] = \texttt{low}]\!]$
$[\![r_2.t = \text{temperature and } r_2.v[\text{max}] = \texttt{medium}]\!]$

$F_2 = \{\text{time}, \text{min}\}$

(F1) $[\![F_2 = \{\text{time}, \text{min}\}]\!]$
(F2) $[\![F_2 = \{\text{time}, \text{min}\} \text{ and } r_2.v[\text{time}] = \texttt{5pm-6am}]\!]$
(F2) $[\![F_2 = \{\text{time}, \text{min}\} \text{ and } r_2.v[\text{min}] = \texttt{low}]\!]$

$T_2 = \langle \textit{with a low around } [\text{min}] \rangle$

(W1) $[\![\text{BASE}(T_2) = \langle \textit{with a low around } [\text{min}] \rangle]\!]$
$[\![\text{COARSE}(T_2) = \langle \textit{with a } [\text{time}] \textit{ around } [\text{min}] \rangle]\!]$
(W2) $[\![\text{BASE}(T_2) = \langle \textit{with a low around } [\text{min}] \rangle \text{ and } r_2.v[\text{time}] = \texttt{5pm-6am}]\!]$
$[\![\text{COARSE}(T_2) = \langle \textit{with a } [\text{time}] \textit{ around } [\text{min}] \rangle \text{ and } r_2.v[\text{time}] = \texttt{5pm-6am}]\!]$
$[\![\text{BASE}(T_2) = \langle \textit{with a low around } [\text{min}] \rangle \text{ and } r_2.v[\text{min}] = \texttt{low}]\!]$
$[\![\text{COARSE}(T_2) = \langle \textit{with a } [\text{time}] \textit{ around } [\text{min}] \rangle \text{ and } r_2.v[\text{min}] = \texttt{low}]\!]$
(W3) $\log p_{\text{LM}}(\textit{with} \mid \textit{cloudy ,})$

Figure 3: The generation process on an example WEATHERGOV scenario. The figure is divided into two parts: The upper part of the figure shows the generation of text from the world state via a sequence of seven decisions (in boxes). Three of these decisions are highlighted and the features that govern these decisions are shown in the lower part of the figure. Note that different decisions in the generation process would result in different features being active (nonzero).

Paper 1 : Challenges in Data-to-Document Generation

Sam Wiseman, Stuart M. Shieber, Alexander M. Rush

EMNLP 2018

Summary

- A new, large-scale corpus of data records with descriptive documents
- Extractive evaluation methods for performance analysis
- Baseline results using current neural generation models and a templatized generator

Data-to-Text Datasets

Two sources of of articles summarizing NBA basketball games with corresponding box- and line-score tables

- RotoWire: professionally written, colloquial, well structured
- SBNation: fan-written summaries, larger, informal, tangential to the statistics

Dataset Notation

$(\boldsymbol{s}, y_{1:T})$ (Data, Text)

$y_{1:T}$ Human-generated summary for S

$\hat{y}_{1:T}$ Machine-generated summary for S

$\boldsymbol{s} = \{r_j\}_{j=1}^{J}$ A set of records

$r.t$ Record type (eg. points)

$r.m$ Record value (eg. 50)

$r.e$ Record entity (eg. Russell Westbrook)

s

r_j

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

r.t = points

r.e

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
...						

r.m

$y_{1:t}$

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.

Comparison with previous datasets

	RC	WG	WB	RW	SBN
Vocab	409	394	400K	11.3K	68.6K
Tokens	11K	0.9M	19M	1.6M	8.8M
Examples	1.9K	22.1K	728K	4.9K	10.9K
Avg Len	5.7	28.7	26.1	337.1	805.4
Rec. Types	4	10	1.7K	39	39
Avg Records	2.2	191	19.7	628	628

Table 1: Vocabulary size, number of total tokens, number of distinct examples, average generation length, total number of record types, and average number of records per example for the ROBOCUP (RC), WEATHERGOV (WG), WIKIBIO (WB), ROTOWIRE (RW), and SBNATION (SBN) datasets.

Evaluating Document Generation

- BLEU
  - Rewards fluent text generation rather than capture the most important information or report information in a coherent way
- Human evaluation
  - Less convenient

Propose new automatic metrics with the intuition that extracting information from documents is easier than document generation

Relation Extractive Evaluation

- Extract candidate entity (player, team, and city) and value (number and certain string) pairs r.e,r.m
- Predict the type r.t (or none) of each candidate pair.

Model $p(r.t \mid e, m; \boldsymbol{\theta})$ for each pair, with unrelated pairs $r.t = \epsilon$

$$\mathcal{L}(\boldsymbol{\theta}) = -\sum_{e,m} \log \sum_{t' \in t(e,m)} p(r.t = t' \mid e, m; \boldsymbol{\theta})$$

r.e (record.entity) = Russell Westbrook
r.m (record.value) = 50
r.t (record.type) = Points

90% accuracy on RotoWire held out
Recall ~60% of the relations by the records

Comparing Generations

- Content Selection (CS)
  - Precision and recall of unique relations from machine-generated text and human-generated text
- Relation Generation (RG)
  - Precision and # of unique relations from generation that also appear in s
- Content Ordering (CO)
  - Normalized Damerau-Levenshtein Distance between sequences of records from generation and ground truth text

Comparing with adversarial evaluation approaches which uses a black-box classifier to determine the quality of generation, this method is more interpretable

Neural Data-to-Document Models

- Standard attention-based encoder-decoder model and its extensions
  - Base Model
  - Base Model with copy-based generation
  - Base Model with training with a source reconstruction term in the loss

Base Model

- $r \in \boldsymbol{s}$ into a vector $\tilde{\boldsymbol{r}}$ ⇨ $\tilde{\boldsymbol{s}} = \{\tilde{\boldsymbol{r}}_j\}_{j=1}^{J}$
- Embedding r.t, r.e, and r.m, and then applying a 1-layer MLP
- Using an LSTM decoder with attention and input-feeding, to compute the probability of each target word, conditioned on the previous words and on *s*
- Model is trained end-to-end to minimize the negative log-likelihood of the words in the human-generated text given corresponding source material s.

Copying

- Introduce an additional binary variable z_t into the per-timestep target word distribution to indicates whether the target word is copied from source or generated
- Assume that target words are copied from the *value* portion of a record r

$$p(\hat{y}_t \mid \hat{y}_{1:t-1}, \boldsymbol{s}) = \sum_{z \in \{0,1\}} p(\hat{y}_t, z_t = z \mid \hat{y}_{1:t-1}, \boldsymbol{s})$$

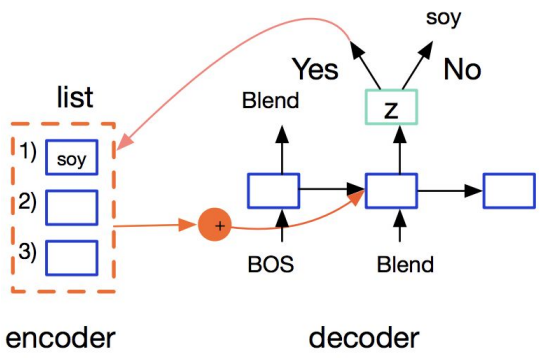


Figure 2: Recipe pointer

Yang, 2016

Joint Copy Model

- parameterize the *joint* distribution table over $\hat{y}_t, z_t$ directly
- copy and gen are functions parameterized in terms of the decoder RNN's hidden state that assign scores to words

$$p(\hat{y}_t, z_t \mid \hat{y}_{1:t-1}, \boldsymbol{s}) \propto \begin{cases} \text{copy}(\hat{y}_t, \hat{y}_{1:t-1}, \boldsymbol{s}) & z_t = 1, \hat{y}_t \in \boldsymbol{s} \\ 0 & z_t = 1, \hat{y}_t \notin \boldsymbol{s} \\ \text{gen}(\hat{y}_t, \hat{y}_{1:t-1}, \boldsymbol{s}) & z_t = 0, \end{cases}$$

Conditional Copy Model

- decompose the joint probability as:

$$p(\hat{y}_t, z_t \mid \hat{y}_{1:t-1}, \boldsymbol{s}) =$$

$$\begin{cases} p_{\text{copy}}(\hat{y}_t \mid z_t, \hat{y}_{1:t-1}, \boldsymbol{s})\, p(z_t \mid \hat{y}_{1:t-1}, \boldsymbol{s}) & z_t{=}1 \\ p_{\text{gen}}(\hat{y}_t \mid z_t, \hat{y}_{1:t-1}, \boldsymbol{s})\, p(z_t \mid \hat{y}_{1:t-1}, \boldsymbol{s}) & z_t{=}0, \end{cases}$$

where an MLP is used to model $p(z_t \mid \hat{y}_{1:t-1}, \boldsymbol{s})$.

- modify the p_{copy} portion of the loss to sum over all matched records

$$p_{\text{copy}}(y_t \mid z_t, y_{1:t-1}, \boldsymbol{s}) = \sum_{r \in r(y_t)} p(r \mid z_t, y_{1:t-1}, \boldsymbol{s})$$

Reconstruction Losses

- Utilize the hidden states of the decoder to try to reconstruct the input data
- Segment the decoder hidden states $\boldsymbol{h}_t$ into $\lceil \frac{T}{B} \rceil$ contiguous blocks of size at most B
- Single one of these hidden state blocks as b_i
- $p(r.e, r.m \mid \boldsymbol{b}_i) = \mathrm{softmax}(f(\boldsymbol{b}_i))$

$$\mathcal{L}(\boldsymbol{\theta}) = -\sum_{k=1}^{K} \min_{r \in \boldsymbol{s}} \log p_k(r \mid \boldsymbol{b}_i; \boldsymbol{\theta})$$

$$= -\sum_{k=1}^{K} \min_{r \in \boldsymbol{s}} \sum_{x \in \{e,m,t\}} \log p_k(r.x \mid \boldsymbol{b}_i; \boldsymbol{\theta}),$$

Templatized Generator

first emits a sentence about the teams playing in the game

The <team1> (<wins1>-<losses1>) defeated the <team2> (<wins2>-<losses2>) <pts1>-<pts2>.

6 highest-scoring players sentences

<player> scored <pts> points (<fgm>-<fga> FG, <tpm>-<tpa> 3PT, <ftm>-<fta> FT) to go with <reb> rebounds.

a typical end sentence

The <team1>' next game will be at home against the Dallas Mavericks, while the <team2> will travel to play the Bulls.

Results

		Development						
		RG		CS		CO	PPL	BLEU
Beam	Model	P%	#	P%	R%	DLD%		
	Gold	91.77	12.84	100	100	100	1.00	100
	Template	99.35	49.7	18.28	65.52	12.2	N/A	6.87
B=1	Joint Copy	47.55	7.53	20.53	22.49	8.28	7.46	10.41
	Joint Copy + Rec	57.81	8.31	23.65	23.30	9.02	7.25	10.00
	Joint Copy + Rec + TVD	60.69	8.95	23.63	24.10	8.84	7.22	12.78
	Conditional Copy	68.94	9.09	25.15	22.94	9.00	7.44	13.31
B=5	Joint Copy	47.00	10.67	16.52	26.08	7.28	7.46	10.23
	Joint Copy + Rec	62.11	10.90	21.36	26.26	9.07	7.25	10.85
	Joint Copy + Rec + TVD	57.51	11.41	18.28	25.27	8.05	7.22	12.04
	Conditional Copy	71.07	12.61	21.90	27.27	8.70	7.44	14.46
		Test						
	Template	99.30	49.61	18.50	64.70	8.04	N/A	6.78
	Joint Copy + Rec (B=5)	61.23	11.02	21.56	26.45	9.06	7.47	10.88
	Joint Copy + Rec + TVD (B=1)	60.27	9.18	23.11	23.69	8.48	7.42	12.96
	Conditional Copy (B=5)	71.82	12.82	22.17	27.16	8.68	7.67	14.49

Table 2: Performance of induced metrics on gold and system outputs of RotoWire development and test data. Columns indicate Record Generation (RG) precision and count, Content Selection (CS) precision and recall, Count Ordering (CO) in normalized Damerau-Levenshtein distance, perplexity, and BLEU. These first three metrics are described in Section 3.2. Models compare Joint and Conditional Copy also with addition Reconstruction loss and Total Variation Distance extensions (described in Section 4).

Human Evaluation

	# Supp.	# Cont.	Order Rat.
Gold	2.04	0.70	5.19
Joint Copy	1.65	2.31	3.90
Joint Copy + Rec	2.33	1.83	4.43
Joint Copy + Rec +TVD	2.43	1.16	4.18
Conditional Copy	3.05	1.48	4.03

Table 3: Average rater judgment of number of box score fields supporting (left column) or contradicting (middle column) a generated sentence, and average rater Likert rating for the naturalness of a summary's ordering (right column). All generations use B=1.

Qualitative Example

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.

Conclusion

- Explored the challenges of neural data-to-document generation by:
  - introducing a new dataset
  - proposing metrics for automatically evaluating content selection, generation, and ordering
  - ideas in copying and reconstruction (neural models) improved the results, but still a significant gap between them and templated systems.

Future work

- approaches to process the source records in a more sophisticated way
- incorporate semantic or reference-related constraints in generation models
- condition on facts/records that are not as explicit in the box- and line-scores.

Recent Work

- Data-to-Text Generation with Content Selection and Planning
  - Ratish Puduppully and Li Dong and Mirella Lapata
  - AAAI 2019

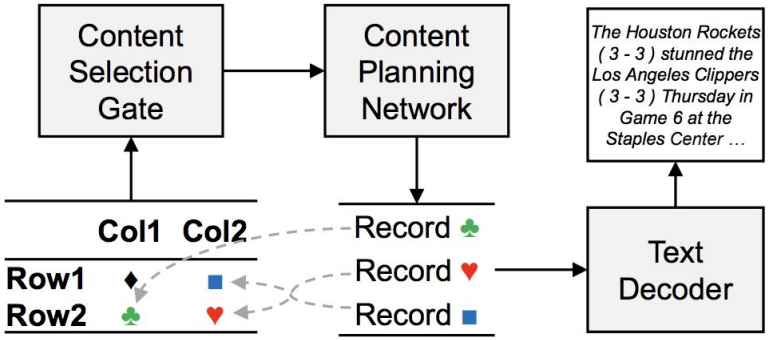


Figure 1: Block diagram of our approach.

Model	RG #	RG P%	CS P%	CS R%	CO DLD%	BLEU
TEMPL	**54.23**	**99.94**	26.99	**58.16**	14.92	8.46
WS-2017	23.72	74.80	29.49	36.18	15.42	14.19
NCP+JC	34.09	87.19	32.02	47.29	17.15	14.89
NCP+CC	34.28	87.47	**34.18**	51.22	**18.58**	**16.50**

Table 5: Automatic evaluation on ROTOWIRE test set using relation generation (RG) count (#) and precision (P%), content selection (CS) precision (R%) and recall (R%), content ordering (CO) in normalized Damerau-Levenshtein distance (DLD%), and BLEU.

Recent Work

- Learning Neural Templates for Text Generation
  - Sam Wiseman, Stuart M. Shieber, Alexander M. Rush
  - August 2018

Source Entity: Cotto

type[coffee shop], rating[3 out of 5], food[English], area[city centre], price[moderate], near[The Portland Arms]

System Generation:

Cotto is a coffee shop serving English food in the moderate price range. It is located near The Portland Arms. Its customer rating is 3 out of 5.

Neural Template:

The ___ ___ …	is a / is an / is an expensive …	___	providing / serving / offering …	___
food / cuisine / foods …	in the / with a / and has a …	___	price range / price bracket / pricing …	.	It's / It is / The place is …
located in the / located near / near …	___	.	Its customer rating is / Their customer rating is / Customers have rated it …	___	.

Figure 1: An example template-like generation from the E2E Generation dataset (Novikova et al., 2017). Knowledge base x (top) contains 6 records, and $\hat{y}$ (middle) is a system gen-

Recent Work

- Data-to-text Generation with Entity Modeling
  - Ratish Puduppully, Li Dong, Mirella Lapata
  - ACL 2019

RW	RG		CS		CO	BLEU
	#	P%	P%	R%	DLD%	
TEMPL	**54.23**	**99.94**	26.99	**58.16**	14.92	8.46
WS-2017	23.72	74.80	29.49	36.18	15.42	14.19
NCP+CC	34.28	87.47	34.18	51.22	18.58	**16.50**
ENT	30.11	92.69	**38.64**	48.51	**20.17**	16.12

Recent Work

- Enhanced Transformer Model for Data-to-Text Generation
  - Li Gong, Josep Crego, Jean Senellart
  - EMNLP-IJCNLP 2019

Model	RG		CS		CO	BLEU
	#	P%	P%	R%	DLD%	
GOLD	23.32	94.77	100	100	100	100
TEMPL	54.29	99.92	26.61	59.16	14.42	8.51
WS-2017	23.95	75.10	28.11	35.86	15.33	14.57
NCP-2019	33.88	87.51	33.52	51.21	18.57	16.19
DATA-TRANS	23.31	79.81	36.90	43.06	22.75	20.60
+DATA_GEN	22.59	**82.49**	**39.48**	42.84	23.32	19.76
+DATA_SEL	**26.94**	79.54	35.27	**47.49**	22.22	19.97
+BOTH	24.24	80.52	37.33	44.66	23.04	20.22

Table 2: Automatic evaluation on ROTOWIRE development set using relation generation (RG) count (#) and precision (P%), content selection (CS) precision (P%) and recall (R%), content ordering (CO) in normalized Damerau-Levenshtein distance (DLD%), and BLEU.

Paper 2 : ToTTo: A Controlled Table-To-Text Generation Dataset

Ankur P. Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui,
Bhuwan Dhingra, Diyi Yang, Dipanjan Das

Goals :

- Formulate a controlled generation task
- Dataset construction process

Table Title: Cristhian Stuani
Section Title: International goals
Table Description: As of 25 March 2019 (Uruguay score listed first, score column indicates score after each Stuani goal)

No.	Date	Venue	Opponent	Score	Result	Competition
1.	10 September 2013	Estadio Centenario, Montevideo, Uruguay	Colombia	2-0	2-0	2014 FIFA World Cup qualification
2.	13 November 2013	Amman International Stadium, Amman, Jordan	Jordan	2-0	5-0	2014 FIFA World Cup qualification
3.	31 May 2014	Estadio Centenario, Montevideo, Uruguay	Northern Ireland	1-0	1-0	Friendly
4.	5 June 2014		Slovenia	2-0	2-0	

Final Text: On 13 November 2013 Cristhian Stuani netted the second in a 5 – 0 win in Jordan.

Contribution :

- Benchmark dataset for conditional text generation
- Baseline evaluation

Related work - What is different ?

Dataset	Train Size	Domain	Target Quality	Target Source	Content Selection
Wikibio (Lebret et al., 2016)	583K	Biographies	Noisy	Wikipedia	Not specified
Rotowire (Wiseman et al., 2017)	4.9K	Basketball	Noisy	Rotowire	Not specified
WebNLG (Gardent et al., 2017b)	25.3K	15 DBPedia categories	Clean	Annotator Generated	Fully specified
E2E (Novikova et al., 2017)	50.6K	Restaurants	Clean	Annotator Generated	Partially specified
LogicNLG (Chen et al., 2020)	28.5K	Wikipedia (open-domain)	Clean	Annotator Generated	Columns via entity linking
ToTTo	**120K**	**Wikipedia (open-domain)**	**Clean**	**Wikipedia (Annotator Revised)**	**Annotator highlighted**

Table 2: Comparison of popular data-to-text datasets. ToTTo combines the advantages of annotator-generated and fully natural text through a revision process.

Dataset Definition

$$\boldsymbol{t} = \{\boldsymbol{c}_j\}_{j=1}^{\tau}$$

- String
- Row or Column Header
- Row and Column Position
- Number of rows and columns(cell spans)

$$\boldsymbol{s} = (s_1, \ldots, s_\eta)$$

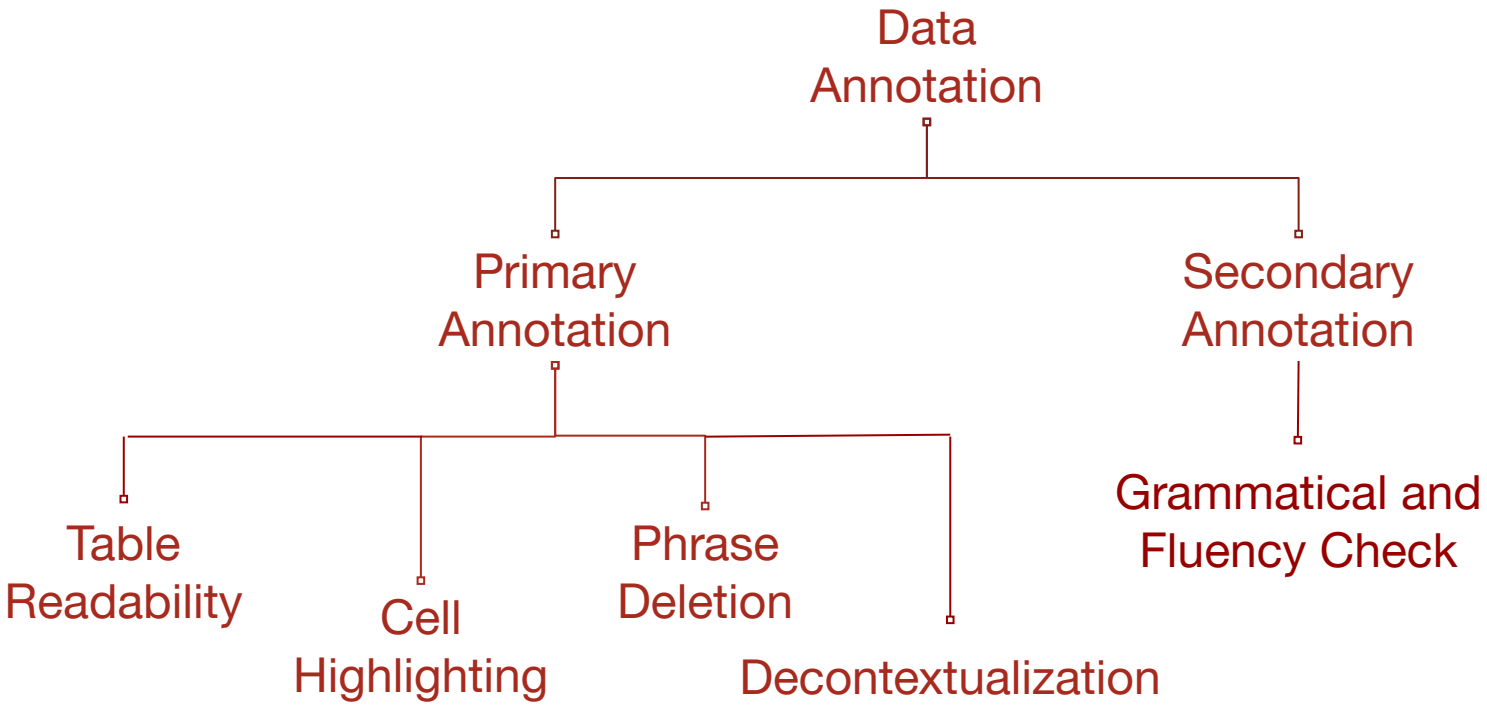
$$\boldsymbol{m} = (m_{\text{page-title}}, m_{\text{section-title}}, m_{\text{section-text}})$$

$$\boldsymbol{d} = (\boldsymbol{t}, \boldsymbol{m}, \boldsymbol{s}) \qquad \boldsymbol{D} = \{\boldsymbol{d}_n\}_{n=1}^{N}$$

Dataset Collection (Wikipedia)

- Number matching
  - overlap with a non-date number of at least 3 non-zero digits
- Cell matching
  - tokens matching at least 3 distinct cell contents from the same row in the table
- Hyperlinks

Annotation Process



Annotation Example

Table Title: Cristhian Stuani
Section Title: International goals
Table Description: As of 25 March 2019 (Uruguay score listed first, score column indicates score after each Stuani goal)

No.	Date	Venue	Opponent	Score	Result	Competition
1.	10 September 2013	Estadio Centenario, Montevideo, Uruguay	Colombia	2-0	2-0	2014 FIFA World Cup qualification
2.	13 November 2013	Amman International Stadium, Amman, Jordan	Jordan	2-0	5-0	2014 FIFA World Cup qualification
3.	31 May 2014	Estadio Centenario, Montevideo, Uruguay	Northern Ireland	1-0	1-0	Friendly
4.	5 June 2014		Slovenia	2-0	2-0	

Original Text: On 13 November 2013, he netted the Charruas' second in their 5 – 0 win in Jordan for the playoffs first leg, finishing Nicolas Lodeiro's cross at close range.
Text after Deletion: On 13 November 2013, he netted the second in their 5 – 0 win in Jordan.
Text after Decontextualization: On 13 November 2013, Cristhian Stuani netted the second in 5 – 0 win in Jordan.
Final Text: On 13 November 2013 Cristhian Stuani netted the second in a 5 – 0 win in Jordan.

Examples

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.

Table 3: 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.

Dataset Analysis

Types	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%

Table 6: Distribution of different linguistic phenomena among 100 randomly chosen sentences.

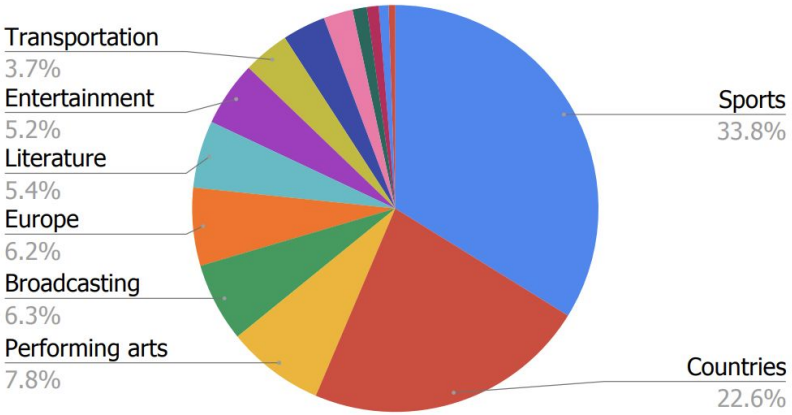


Figure 1: Topic distribution of our dataset.

Dataset splits

Property	Value
Training set size	120,761
Number of target tokens	1,268,268
Avg Target Length (tokens)	17.4
Target vocabulary size	136,777
Unique Tables	83,141
Rows per table (Median/Avg)	16 / 32.7
Cells per table (Median/Avg)	87 / 206.6
No. of Highlighted Cell (Median/Avg)	3 / 3.55
Development set size	7,700
Test set size	7,700

Table 4: ToTTo dataset statistics.

$$\boldsymbol{D}_{\text{train}} := \{\boldsymbol{d} : \boldsymbol{h}(\boldsymbol{d}) \notin (\boldsymbol{h}(\boldsymbol{D}_{\text{dev}}) \cup \boldsymbol{h}(\boldsymbol{D}_{\text{test}})) \text{ or } \text{count}\left(\boldsymbol{h}(\boldsymbol{d}), \boldsymbol{D}_{\text{orig-train}}\right) > \kappa\}.$$

For a given d,
h(d) - header values
h(D) - set of header values for a given Dataset

$$\boldsymbol{D}_{\text{test-overlap}} := \{\boldsymbol{d} : \boldsymbol{h}(\boldsymbol{d}) \in \boldsymbol{h}(\boldsymbol{D}_{\text{train}})\}$$
$$\boldsymbol{D}_{\text{test-nonoverlap}} := \{\boldsymbol{d} : \boldsymbol{h}(\boldsymbol{d}) \notin \boldsymbol{h}(\boldsymbol{D}_{\text{train}})\}$$

Experiments

Given : A table t and related metadata m (page title, section title, table section text), a set of highlighted cells $t_{highlight}$, produce the final sentence s_{final}.

<u>Objective</u>: $f : x \rightarrow y$ where $x = (t, m, t_{highlight})$ and $y = s_{final}$

Three models and two version(Full table, Subtable):

- BERT-to-BERT
- Pointer-Generator
- Seq2Seq model with explicit content selection

Models

- BERT-to-BERT(Rothe et al.)
  - A transformer encoder - decoder architecture, pre trained with Books Corpus

- Pointer-Generator(See et al.)
  - A Seq2Seq model with attention and copy mechanism

BERT-to-BERT

Pointer Gen

- Seq2Seq model with explicit content selection([Puduppully et al.](#))

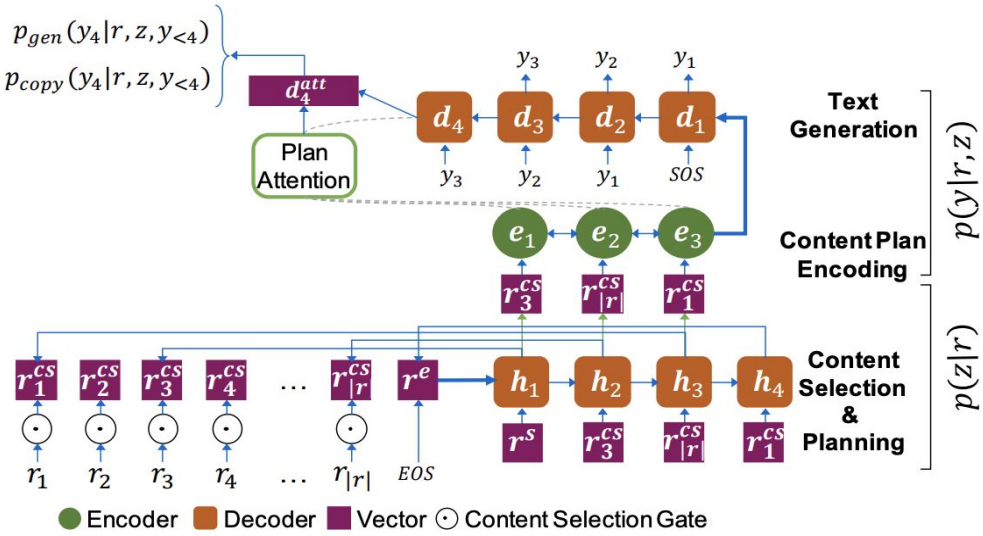


Figure 2: Generation model with content selection and planning; the content

$$p(y|r) = \sum_z p(y,z|r) = \sum_z p(z|r)p(y|r,z)$$

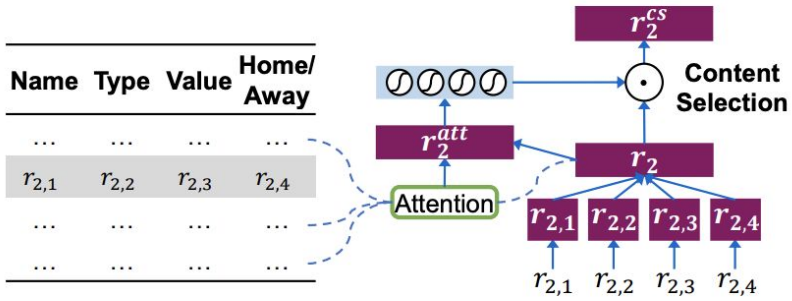


Figure 3: Content selection mechanism.

Text-to-Text Pre-Training for Data-to-Text Tasks(Kale)

The data-to-text task is cast in the text-to-text framework by representing the structured data as a flat string

Table Title: Cristhian Stuani
Section Title: International goals

No.	Date	Venue	Opponent	Result
2	13 November 2013	Amman International Stadium, Amman, Jordan	Jordan	5-0

```
<page_title> Cristhian Stuani </page_title>
<section_title> International goals </section_title>
<table> <cell> 2. <col_header> No. </col_header> </cell>
<cell> 13 November 2013 <col_header> Date </col_header>
</cell> <cell> Amman International Stadium, Amman,
Jordan <col_header> Venue </col_header> </cell> <cell>
Jordan <col_header> Opponent </col_header> </cell>
<cell> 5-0 <col_header> Result </col_header> </cell>
</table>
```

On 13 November 2013 Cristhian Stuani netted the second in a 5–0 win in Jordan.

[T5 based]

TaBERT: Pre training for Joint Understanding of Textual and Tabular Data (Yin et al.)

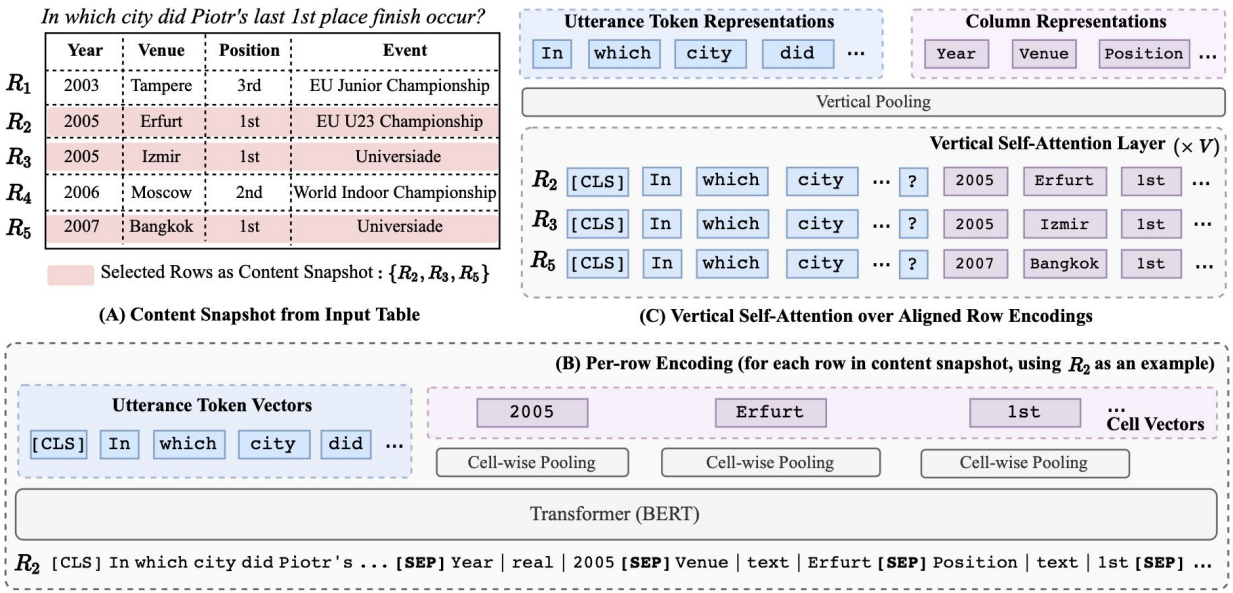


Figure 1: Overview of TaBERT for learning representations of utterances and table schemas with an example from WikiTableQuestions[3]. **(A)** A content snapshot of the table is created based on the input NL utterance. **(B)** Each row in the snapshot is encoded by a Transformer (only R_2 is shown), producing row-wise encodings for utterance tokens and cells. **(C)** All row-wise encodings are aligned and processed by V vertical self-attention layers, generating utterance and column representations.

TaBERT(May 2020)

Evaluation Metrics

- BLEU
- PARENT (Precision And Recall of Entailed Ngrams from the Table)
- Human Evaluation
  - Fluency
  - Faithfulness
  - Covered Cells
  - Coverage with Respect to Reference

Handling Divergent Reference Texts when Evaluating Table-to-Text Generation

PARENT

- Entailment Probability - probability that presence of n-gram ‘g’ in a text is correct given associated table
- Entailed Precision - fraction of n-grams in generated text to be correct if it occurs in reference or high probability being entailed by the table
- Entailed Recall - generated text match reference and cover information from table $E_r(\boldsymbol{x}_n, \boldsymbol{y}_n, \hat{\boldsymbol{y}}_n) = R(\boldsymbol{x}_n, \boldsymbol{y}_n, \hat{\boldsymbol{y}}_n)^{(1-\lambda)} R(\boldsymbol{x}_n, \hat{\boldsymbol{y}}_n)^{\lambda}$

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

Results

Model	Overall		Overlap Subset		Nonoverlap Subset	
	BLEU	**PARENT**	**BLEU**	**PARENT**	**BLEU**	**PARENT**
BERT-to-BERT (Books+Wiki)	**44.0**	**52.6**	**52.7**	**58.4**	**35.1**	**46.8**
BERT-to-BERT (Books)	43.9	**52.6**	**52.7**	**58.4**	34.8	46.7
Pointer-Generator	41.6	51.6	50.6	58.0	32.2	45.2
Puduppully et al. (2019)	19.2	29.2	24.5	32.5	13.9	25.8

Data Format	**BLEU**	**PARENT**
subtable w/ metadata	43.9	52.6
subtable w/o metadata	36.9	42.6
full table w/ metadata	26.8	30.7
full table w/o metadata	20.9	22.2

Human Evaluation

	Model	Fluency (%)	Faithfulness (%)	Covered Cells (%)	Less/Neutral/More Coverage w.r.t. Ref
Overall	*Oracle*	*99.3*	*93.6*	*94.8*	*18.3 / 61.7 / 20.0*
	BERT-to-BERT (Books)	88.1	76.2	89.0	49.2 / 36.2 / 14.5
	BERT-to-BERT (Books+Wiki)	87.3	73.6	87.3	53.9 / 32.9 / 13.2
Overlap	*Oracle*	*99.6*	*96.5*	*95.5*	*19.8 / 62.8 / 17.4*
	BERT-to-BERT (Books)	89.6	78.7	92.1	42.0 / 43.7 / 14.3
	BERT-to-BERT (Books+Wiki)	89.8	81.1	91.0	47.8 / 39.2 / 13.1
Non-overlap	*Oracle*	*99.1*	*91.4*	*94.3*	*17.0 / 60.9 / 22.1*
	BERT-to-BERT (Books)	86.9	74.2	86.4	55.5 / 29.8 / 14.7
	BERT-to-BERT (Books+Wiki)	84.8	66.6	83.8	60.1 / 26.6 / 13.3

Table 9: Human evaluation over references (to compute *Oracle*) and model outputs. For Fluency, we report the percentage of outputs that were completely fluent. In the last column $X/Y/Z$ means X% and Z% of the candidates were deemed to be less and more informative than the reference respectively and Y% were neutral.

Decoder Output

ID	Reference	Decoder output (w/ metadata)		w/o metadata
		Full table	Subtable	Subtable
1	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.
2	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.
3	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.
4	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.
5	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.
6	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.
7	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.

Conclusion & Challenges

Presents a controlled generation task and annotation process for a large English table-to-text dataset

- Hallucination - Reference targets are faithful to the source
- Rare topics - Struggle with generalization
- Diverse table structure - Difficult to make inferences
- Numerical reasoning - Still a challenge
- Evaluation metrics - can the current metrics capture all these

Interesting Reference

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 6: ToTTo example with interesting reference language.

Rare Topics

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 5: TOTTO example with rare topic.

Numerical Reasoning

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 2: ToTTo example with numerical reasoning about table cells.

Complex Table Structure

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 3: ToTTo example with complex table structure and temporal reasoning.

Discussion

- Thoughts on the task formulation ? Is it really indicating content selection when you highlight the selected cells ?
- Is noisy or clean data really needed - Does it model the real scenario or will it fail ?
- Other methods to select table and sentence (sentences with reference “ as shown in Table 1”)

Discussion ctd.

- Which among the challenges needs to be addressed first ? - Hallucination?

- Model Performance at different stages of annotation - in terms of BLEU score


YOU
BELONG
HERE
cns.utexas.edu/diversity
CNS Diversity and Inclusion

Thank you!