Challenges in Data-to-Document Generation

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

March 21, 2019

INTRODUCTION

- Neural systems begin to move toward generating longer outputs in response to longer and more complicated inputs.
- Generated texts begin to display reference errors, inter-se ntence incoherence, and a lack of fidelity to the source material.
- Introduce a large-scale corpus of data records of basketball games paired with descriptive documents.
- Suggest a series of extractive evaluation metrics to automatically evaluate performance.

DATA-TO-TEXT DATASETS

Setting:

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

Entity: $r.e$ Value: $r.m$ Relation: $r.t$

Generated text: $\hat{y}_{1:T} = \hat{y}_1, \ldots, \hat{y}_T$

Dataset: $(\boldsymbol{s}, y_{1:T})$ $y_{1:T}$ Gold Summary of $\boldsymbol{s}$

DATA-TO-TEXT DATASETS

- Existing Datasets: WEATHERGOV and ROBOCUP

  Problem: Simple, Short generations, Machine-generated

- Proposed Datasets: ROTOWIRE and SBNATION

  Longer target texts, a larger vocabulary space, and

  to require more difficult content selection

DATA-TO-TEXT DATASETS

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 ...

EVALUATION METHODS

Shortcomings of current approaches:

- *BLEU: It primarily rewards fluent text generation, rather than generations that capture the most important information in the database*
- *Human evaluation: Less convenient*

EXTRACTION EVALUATION

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 ...

extract: $r.e, r.m$ *predict:* $r.t$

e.g., $(r.e, r.m, r.t) = (\textsc{Miami Heat}, 95, \textsc{points})$

$$t(e,m) = \{r.t : r \in \boldsymbol{s}, r.e = e, r.m = m\}$$

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

COMPARING GENERATION

Three induced metrics:

- Content Selection (CS): precision and recall of unique relations r extracted from $\hat{y}_{1:T}$ that are also extracted from $y_{1:T}$
- Relation Generation (RG): precision and number of unique relations r extracted from $\hat{y}_{1:T}$ that also appear in $\boldsymbol{s}$
- Content Ordering (CO): normalized Damerau-Levenshtein Distance between the sequences of records extracted from $y_{1:T}$ and that extracted from $\hat{y}_{1:T}$

NEURAL DATA-TO-DOCUMENT MODEL

- *Base Model*
- *Copy-based generation*
- *Reconstruction*

BASE MODEL

- *Embedding:* $\boldsymbol{r} \in \boldsymbol{s}$ *to* $\tilde{\boldsymbol{r}}$
- *One layer MLP*
- *Source data-records:* $\tilde{\mathbf{s}} = \{\tilde{\boldsymbol{r}}_j\}_{j=1}^{J}$
- *LSTM decoder with attention and input-feeding*
- *Minimize the negative log likelihood of words*

in the gold text $y_{1:T}$ *given source material* $\boldsymbol{s}$

NEURAL DATA-TO DOCUMENT MODEL

COPY: z_t *to indicate whether* $\hat{y}_t$ *is from the source* $\hat{y}_t = r.m$

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

- ***Joint Copy***

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

- ***Conditional Copy***

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

$$r(y_t) = \{r \in \boldsymbol{s} : r.m = y_t, \text{same}-\text{sentence}(r.e, r.m)\}$$

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

RECONSTRUCTION LOSS

$\boldsymbol{b}_i$: *hidden state block*

$$p(r.e, r.m \mid \boldsymbol{b}_i) \quad = \quad \mathrm{softmax}(f(\boldsymbol{b}_i))$$

Reconstruction Loss for $\boldsymbol{b}_i$:

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

TEMPLATED GENERATOR

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

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

RESULT

		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

Data-to-Text Generation with Content Selection and Planning

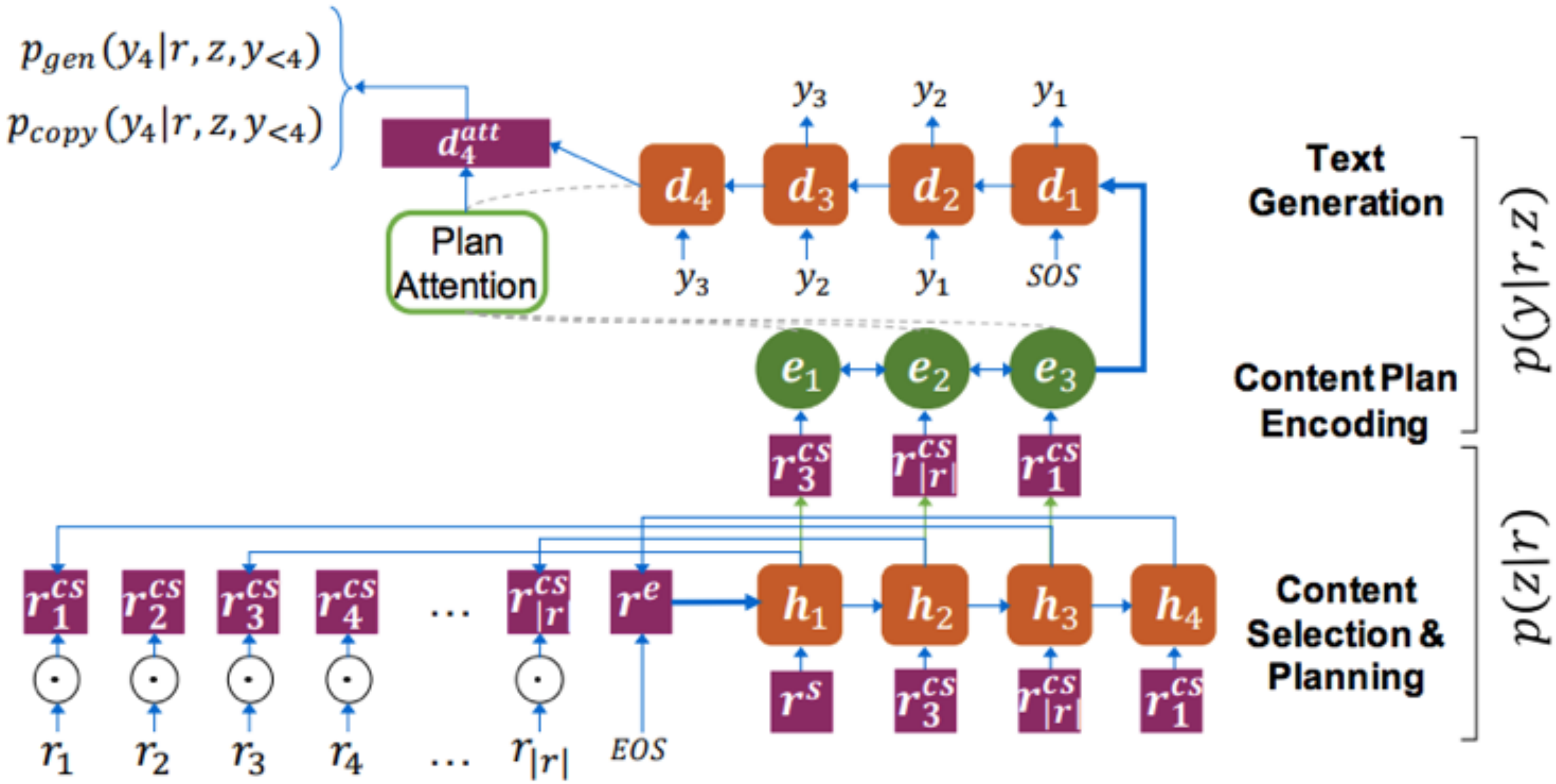
Trend	Dataset	Best Method	Paper title
	RotoWire (Relation Generation)	🏆 Neural Content Planning + conditional copy	Data-to-Text Generation with Content Selection and Planning
	Rotowire (Content Selection)	🏆 Neural Content Planning + conditional copy	Data-to-Text Generation with Content Selection and Planning
	RotoWire (Content Ordering)	🏆 Neural Content Planning + conditional copy	Data-to-Text Generation with Content Selection and Planning
	RotoWire	🏆 Neural Content Planning + conditional copy	Data-to-Text Generation with Content Selection and Planning

Ratish Puduppully and Li Dong and Mirella Lapata
Institute for Language, Cognition and Computation
School of Informatics, University of Edinburgh

INTRODUCTION

- Data-to-text generation: what to say? in what order? how to say it?
- Neural Model Problem: end-to-end, without modeling what to say and in what order
- Propose a neural network architecture which incorporates content selection and planning without sacrificing end-to-end training

MODEL

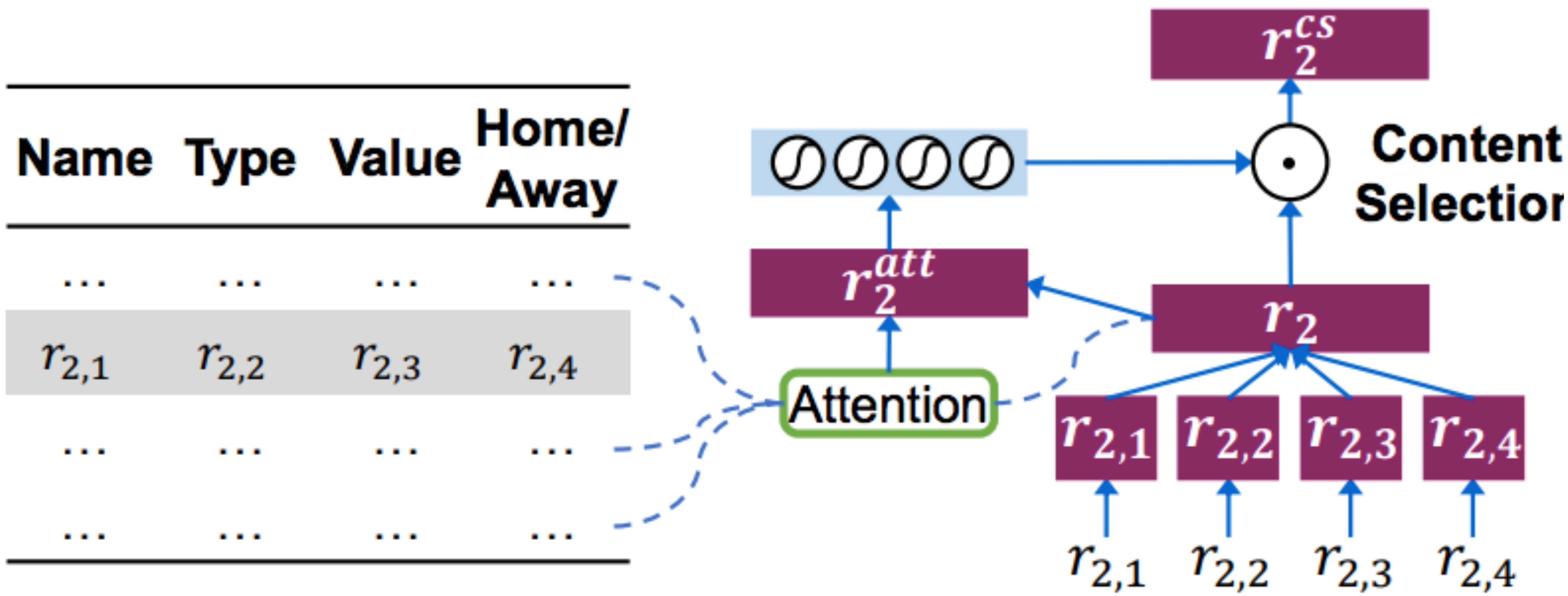


$\{r_{j,k}\}_{k=1}^{4}$

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

z: content plan

CONTENT SELECTION GATE



$$\alpha_{j,k} \propto \exp(\mathbf{r}_j^\intercal \mathbf{W}_a \mathbf{r}_k)$$

$$\mathbf{c}_j = \sum_{k \neq j} \alpha_{j,k} \mathbf{r}_k$$

$$\mathbf{r}_j^{att} = \mathbf{W}_g[\mathbf{r}_j; \mathbf{c}_j]$$

$$\mathbf{g}_j = \text{sigmoid}\left(\mathbf{r}_j^{att}\right)$$

$$\mathbf{r}_j^{cs} = \mathbf{g}_j \odot \mathbf{r}_j$$

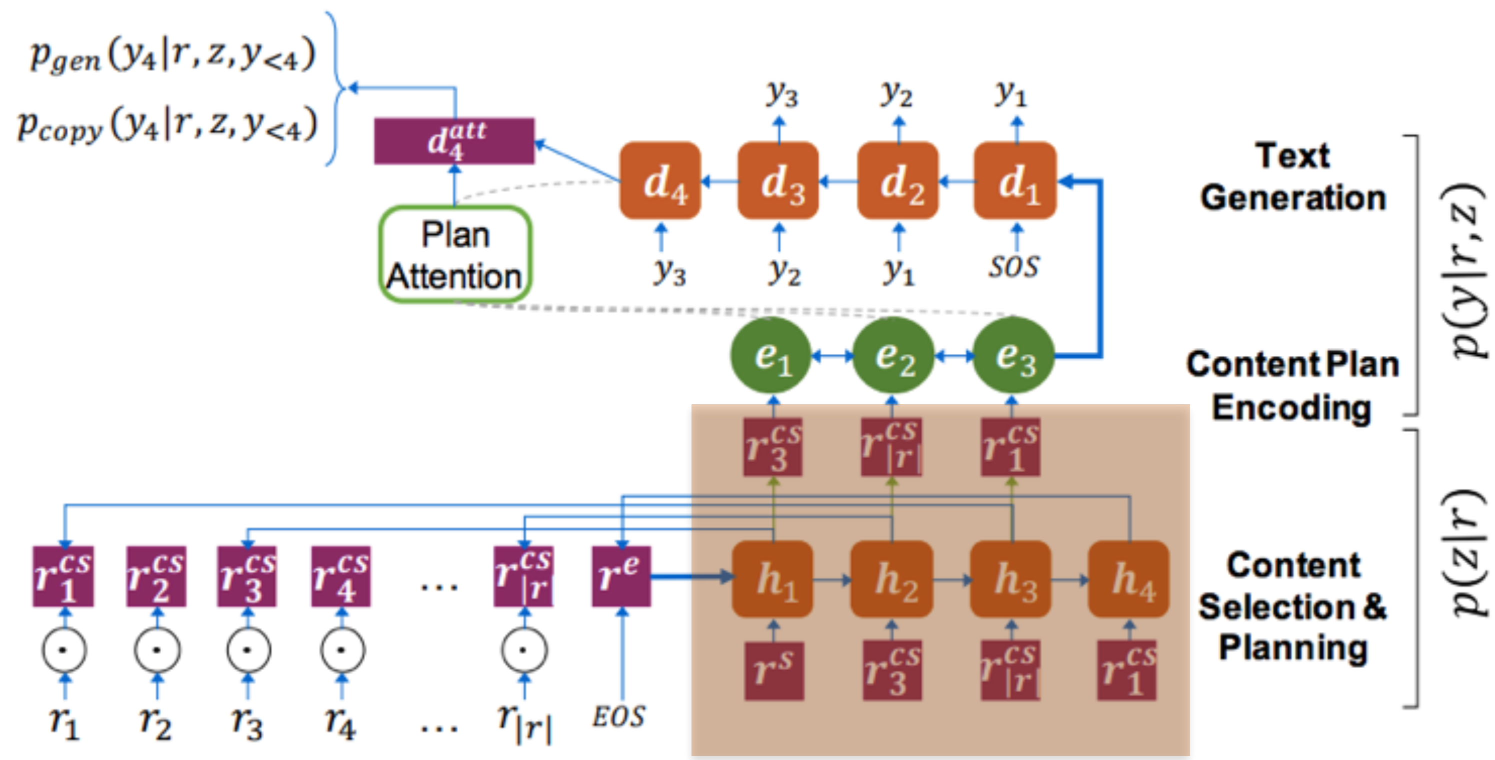
Content Planning

what to say & in what order

$$z = z_1 \ldots z_{|z|}$$

$$z_k \in \{r_j\}_{j=1}^{|r|}$$

$$p(z|r) = \prod_{k=1}^{|z|} p(z_k|z_{<k}, r)$$



$$p(z_k = r_j|z_{<k}, r) \propto \exp(\mathbf{h}_k^\intercal \mathbf{W}_c \mathbf{r}_j^{cs})$$

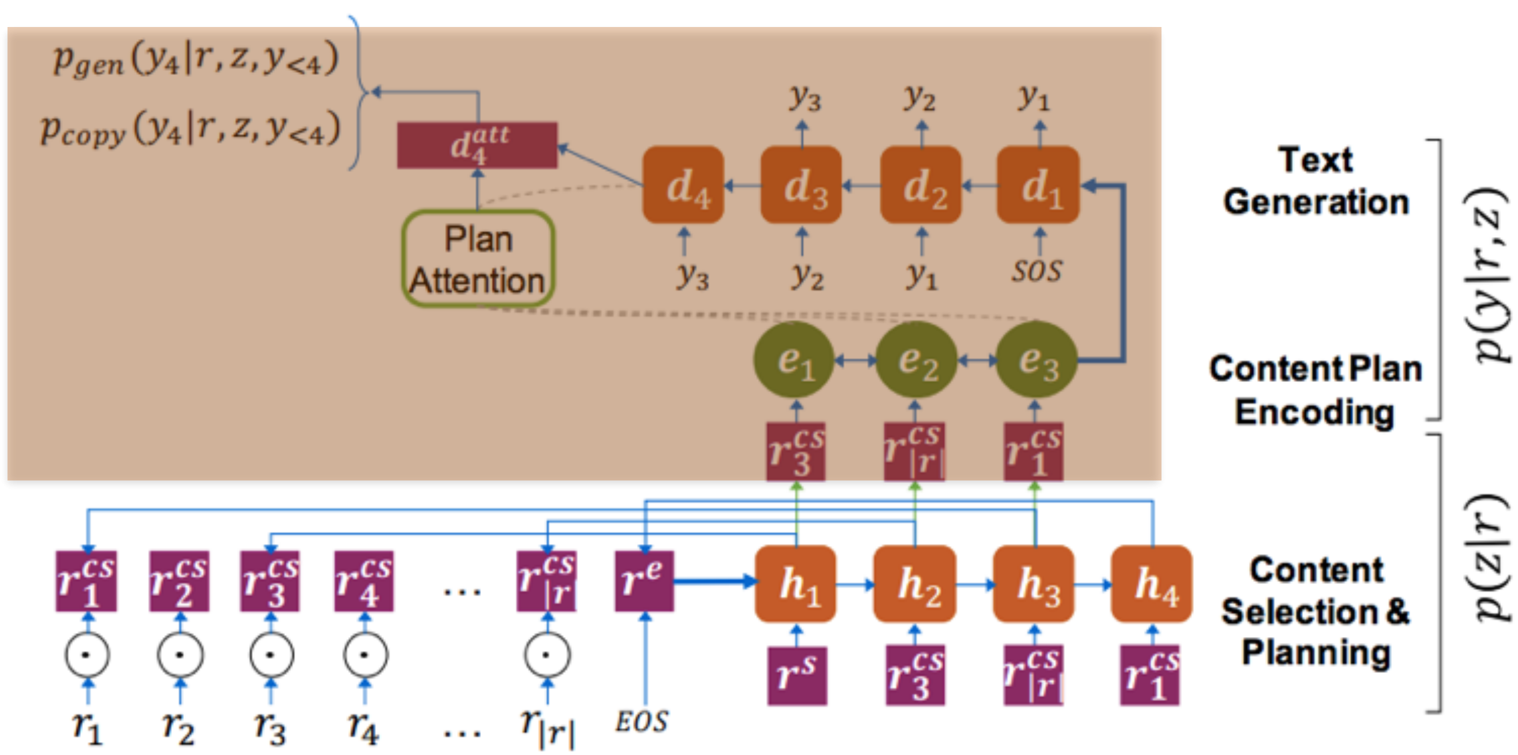
TEXT GENERATION

$$\beta_{t,k} \propto \exp(\mathbf{d}_t^\intercal \mathbf{W}_b \mathbf{e}_k)$$

$$\mathbf{q}_t = \sum_k \beta_{t,k} \mathbf{e}_k$$

$$\mathbf{d}_t^{att} = \tanh(\mathbf{W}_d[\mathbf{d}_t; \mathbf{q}_t])$$

$$p_{gen}(y_t|y_{<t}, z, r) = \text{softmax}_{y_t}(\mathbf{W}_y \mathbf{d}_t^{att} + \mathbf{b}_y)$$



TRAINING AND INFERENCE

$$\max \sum_{(r,z,y)\in\mathcal{D}} \log p\left(z|r\right) + \log p\left(y|r,z\right)$$

$$\hat{z} = \arg\max_{z'} p(z'|r)$$

$$\hat{y} = \arg\max_{y'} p(y'|r,\hat{z})$$

RESULT

Model	RG #	RG P%	CS P%	CS R%	CO DLD%	BLEU
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
ED+JC	22.98	76.07	27.70	33.29	14.36	13.22
ED+CC	21.94	75.08	27.96	32.71	15.03	13.31
NCP+JC	33.37	87.40	32.20	48.56	17.98	14.92
NCP+CC	33.88	87.51	33.52	51.21	18.57	16.19
NCP+OR	21.59	89.21	88.52	85.84	78.51	24.11

Model	RG #	RG P%	CS P%	CS R%	CO DLD%	BLEU
ED+CC	21.94	75.08	27.96	32.71	15.03	13.31
CS+CC	24.93	80.55	28.63	35.23	15.12	13.52
CP+CC	33.73	84.85	29.57	44.72	15.84	14.45
NCP+CC	33.88	87.51	33.52	51.21	18.57	16.19
NCP	34.46	—	38.00	53.72	20.27	—

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**

TEMPL: template-based
ED: Encoder-decoder
JC: Joint Copy
CC: Conditional Copy
NCP: Neural Content Planning
RG: Relation Generation
CS: Content Selection
CO: Content Ordering
NCP: Neural Content Planning
CP: Content Planning
OR: Oracle content plans

CONCLUSION

- Evaluation metrics are still not sound enough. Information extraction itself has inaccuracies.

- In the future, we can learn more detail-oriented plans involving inference over multiple facts and entities

- Future work on this task might include approaches that process or attend to the source records in a more sophisticated way

- There are very few data-to-text tasks and datasets explored, more tasks can be put forward.