# Visual Story Post-Editing

Ting-Yao Hsu, Chieh-Yang Huang, Yen-Chia Hsu, and Ting-Hao (Kenneth) Huang



PennState Carnegie Mellon University Crowd-Al Lab



# What is a Visual Storytelling (VIST) task?

- Input: A sequence of five photos
- Output: A short story describing the photo sequence.
- The VIST dataset [1] contains 20,211 photo sequences, aligning to human-written stories.

[1] Huang, et al. "Visual Storytelling". NAACL'16.

# Ok, but why automatic post-editing (APE)?

- Machine-generated stories is not good enough.
- APE leverages the user-edit data on VIST.
- APE corrects systematic errors of the model and improves story quality.

#### How?

- Learn the transformation from machine-generated story to human-edited story.
- Augment data by sorting the similarities between edited stories and the original story.
- End-to-end LSTM and Transformer are utilized.
- Two different input setting:
  - Text only (T)
  - Text and Images (T+I)

Machine-Generated Story (a): - visual storytelling

the family got together for a dinner, the food was delicious. everyone was having a great time. the meal was delicious. the kids had a great time.

#### Machine-Generated (a) -> Human-Edited Story (b):

the whole family got together for thanksgiving, the food was delicious! everyone had a lot of fun, and the kids played the entire time. visual story post-editing

#### Machine-Generated (a) -> Machine-Edited Story (c):

the family got together for a nice dinner, the food was delicious. the guys enjoyed the food since they had never eaten there before, the food was presented well, the dessert was delicious.

## Hmmm, tell me about the VIST-Edit dataset?

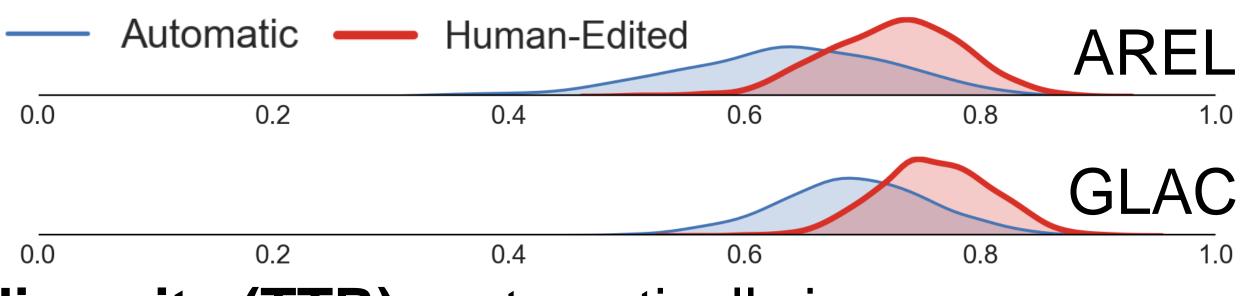
- 2,981 stories generated by GLAC [2] & AREL [3].
- 5 crowd workers from MTurk to edit each story, respectively.
- 2,981\*5 = 14,905 human-edited stories are collected.

Avg. #Token

1	Pre Post △	4.7	3.1	3.4	1.9		7.1	10.1 9.9 -0.2			7.0	43.0 41.9 -1.2
	GLAC		ADJ	ADP	ADV	CONJ	DET	NOUN	PRON	PRT	VERB	Total
	Pre	5.0	3.3	1.7	1.9	0.2	6.5	7.4	1.2	0.8	6.9	35.0
	<b>Post</b>	4.5	3.2	2.4	1.8	0.8	6.1	8.3	1.5	1.0	7.0	36.7
	$\Lambda$	-0.5	_0.1	0.7	-0.1	0.6	-0.3	0.9	0.3	0.2	0.1	1.7

. ADJ ADP ADV CONJ DET NOUN PRON PRT VERB Total

Type-Token Ratio (TTR)



- Lexical diversity (TTR) systematically increases.
- People shorten AREL's stories but lengthen GLAC's stories. (Why?)
- [2] Kim, et al. "GLACNet: GLocal Attention Cascading Networks for Multi-image Cued Story Generation."
- [3] Wang, et al. "No metrics are perfect: Adversarial reward learning for visual storytelling." ACL'18.

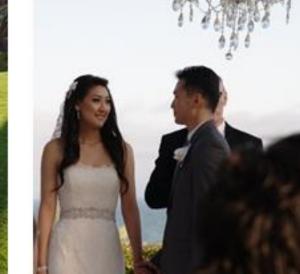
### Does it work?

Focus	Coherence	Share	Human	Grounded	Detailed
3.487	3.751	3.763	3.746	3.602	3.761
3.433	3.705	3.641	3.656	3.619	3.631
3.542	3.693	3.676	3.643	3.548	3.672
3.551	3.800	<b>3.771</b>	3.751	3.631	3.810
3.497	3.734	3.746	3.742	3.573	3.755
3.592	3.870	3.856	3.885	3.779	3.878
		G	ELAC		
Focus	Coherence	Share	Human	Grounded	Detailed
3.878	3.908	3.930	3.817	3.864	3.938
3.717	3.773	3.863	3.672	3.765	3.795
3.734	3.759	3.786	3.622	3.758	3.744
3.894	3.896	3.864	3.848	3.751	3.897
3.815	3.872	3.847	3.813	3.750	3.869
4.003	4.057	4.072	3.976	3.994	4.068
	3.487 3.433 3.542 3.551 3.497 3.592  Focus 3.878 3.717 3.734 3.894 3.815	3.487       3.751         3.433       3.705         3.542       3.693         3.551       3.800         3.497       3.734         3.592       3.870    Focus Coherence         3.878       3.908         3.717       3.773         3.734       3.759         3.894       3.896         3.815       3.872	3.487       3.751       3.763         3.433       3.705       3.641         3.542       3.693       3.676         3.551       3.800       3.771         3.497       3.734       3.746         3.592       3.870       3.856         Focus       Coherence       Share         3.878       3.908       3.930         3.717       3.773       3.863         3.734       3.759       3.786         3.894       3.896       3.864         3.815       3.872       3.847	3.487       3.751       3.763       3.746         3.433       3.705       3.641       3.656         3.542       3.693       3.676       3.643         3.551       3.800       3.771       3.751         3.497       3.734       3.746       3.742         3.592       3.870       3.856       3.885         GLAC         Focus       Coherence       Share       Human         3.878       3.908       3.930       3.817         3.717       3.773       3.863       3.672         3.734       3.759       3.786       3.622         3.894       3.896       3.864       3.848         3.815       3.872       3.847       3.813	3.487       3.751       3.763       3.746       3.602         3.433       3.705       3.641       3.656       3.619         3.542       3.693       3.676       3.643       3.548         3.551       3.800       3.771       3.751       3.631         3.497       3.734       3.746       3.742       3.573         3.592       3.870       3.856       3.885       3.779         GLAC         Focus       Coherence       Share       Human       Grounded         3.878       3.908       3.930       3.817       3.864         3.734       3.759       3.786       3.622       3.758         3.894       3.896       3.864       3.848       3.751         3.815       3.872       3.847       3.813       3.750

**AREL** 

Human Evaluation











Generated By GLAC

**Edited By LSTM** 

(Text-Only)

the wedding was a beautiful event. the bride and groom were very happy, they had a great time at the reception. the couple were so excited, the bride and groom were very happy.

the wedding was a beautiful event. the bride and groom were very happy, the weather was perfect. all of the guests had a great time. everyone was dancing.

**Generated By AREL** 

**Edited By LSTM** 

(Text-Only)

we had a great time at the wedding today. the bride and groom were very happy to be married. the bride and groom were very happy to be married. the bride and groom pose for a picture, at the end of the wedding, the bride and groom pose for a picture.

the wedding was held in a beautiful church. the bride and groom walked down the aisle. they were very happy to be married. the couple looked so lovely together, the bride and groom danced the night away at the reception.

# Auto evaluation is still hard

- **Auto evaluation scores** Human judgements
- Low correlation between auto evaluation scores and human judgments.

	Refere				
	BLEU4	METEOR	ROUGE	Skip-Thoughts	Human Rating
AREL	0.93	0.91	0.92	0.97	3.69
AREL Edited By LSTM(T)	0.21	0.46	0.40	0.76	3.81

	Reference: Human-Written Stories							
	BLEU4	METEOR	ROUGE	Skip-Thoughts				
GLAC	0.03	0.30	0.26	0.66				
GLAC Edited By Human	0.02	0.28	0.24	0.65				

	Spearman rank-order correlation $\rho$						
	Data Includes	BLEU4	METEOR	ROUGE	<b>Skip-Thoughts</b>		
1	AREL	.110	.099	.063	.062		
2	LSTM-Edited AREL	.106	.109	.067	.205		
3	1+2	.095	.092	.059	.116		
4	GLAC	.222	.203	.140	.151		
<b>5</b>	LSTM-Edited GLAC	.163	.176	.138	.087		
<b>6</b>	4+5	.196	.194	.148	.116		
7	1+4	.091	.086	.059	.088		
8	2+5	.089	.103	.067	.101		
9	1+2+4+5	.090	.096	.069	.094		

Spearman rank-order correlation between the auto eval scores and human judgment.