On Decoding and Discourse Structure in Multi-Modal Text Generation

Nikolai Ilinykh Simon Dobnik
Department of Philosophy, Linguistics and Theory of Science
Centre for Linguistic Theory and Studies in Probability (CLASP)
University of Gothenburg, Sweden
{name.surname}@gu.se

GEM 2022, collocated with EMNLP

Overview of the paper

GÖTEBORGS UNIVERSITET

- *Decoding* (greedy, beam, sampling, etc.) is an important task in text generation.

Overview of the paper

- *Decoding* (greedy, beam, sampling, etc.) is an important task in text generation.
- In this paper, we compare texts generated by different decodings and texts produced by humans in the multi-modal task.

Overview of the paper

GÖTEBORGS UNIVERSITET

- *Decoding* (greedy, beam, sampling, etc.) is an important task in text generation.
- In this paper, we compare texts generated by different decodings and texts produced by humans in the multi-modal task.
- Evaluation methods we use are:
  - automatic evaluation (BLEU, etc)
  - human evaluation
  - referring expressions in decoding vs human texts

Overview of the paper

GÖTEBORGS UNIVERSITET

- *Decoding* (greedy, beam, sampling, etc.) is an important task in text generation.
- In this paper, we compare texts generated by different decodings and texts produced by humans in the multi-modal task.
- Evaluation methods we use are:
  - automatic evaluation (BLEU, etc)
  - human evaluation
  - referring expressions in decoding vs human texts
- Task context: image paragraph generation.

Overview of the paper

GÖTEBORGS UNIVERSITET

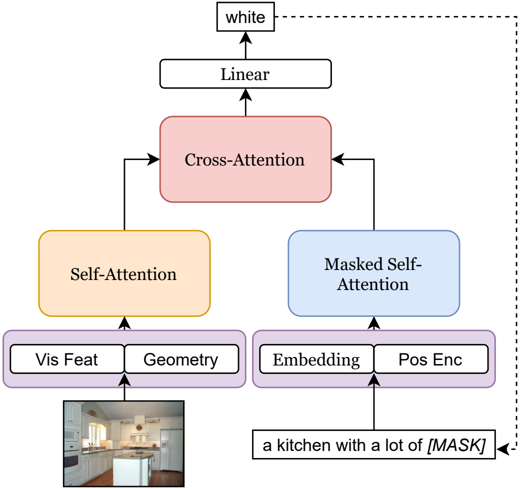
- *Decoding* (greedy, beam, sampling, etc.) is an important task in text generation.
- In this paper, we compare texts generated by different decodings and texts produced by humans in the multi-modal task.
- Evaluation methods we use are:
  - automatic evaluation (BLEU, etc)
  - human evaluation
  - referring expressions in decoding vs human texts
- Task context: image paragraph generation.
- **Primary goal**: compare discourse organisation in texts generated by different decodings and texts produced by humans for image paragraph generation.

CLASP centre for linguistic theory and studies in probability

Image paragraph example (Ilinykh et al., 2019)

1: *This is a large bedroom with two large windows, a bed, and a two person chaise lounge.*
2: *The windows have striped curtains in front of them and a curtain rod that goes over both windows.*
3: *There is a ceiling light and fan in the center of the room.*
4: *There are two large pictures above the bed and dark colored nightstands on both sides.*
5: *There are table lights on the nightstands and several plants throughout the room.*

Our model (Herdade et al., 2019)



Decodings that we evaluate

- Rule-based:
  - g: greedy
  - b2: beam, with beam size $b = 2$
  - db2: diverse Beam with group size $G = b = 2$ and $\lambda = 0.5$
- Sampling-based:
  - s$\{k\}$: top-k sampling, where $k \in \{25, 50, 75, 100\}$
  - n$\{p\}$: nucleus (top-p) sampling, where $p \in \{25, 50, 95\}$
  - st50: temperature scaling with $\tau = 0.5$

Some examples of generated texts

- **ref**: there is a very large well-made bed on the left. this looks like a hotel room. two lamps are on a table on the right side. a window that you can barely see out of can be seen if you look straight ahead. there is a picture on the wall to the left and another on the right side wall.
- **greedy**: this is a living room with a white couch and love seat. there is a white couch with a white and white pillow. there is a coffee table in front of the couch. there is a glass coffee table in the middle of the room. there is a picture hanging on the wall.
- **b2**: this is a living room. there is a white couch. there is a white couch. there is a white coffee table. there is a couch.
- **s50**: a large room with a sofa for a coffee table. a four drawer rectangular coffee table sits full of glass roses loveseat. a fireplace is on the left hand wall. a large telephone is near the couch. there are two brown chairs sitting at the far end of the couch.
- **n50**: this room is very neat and space. the couches have orange felt UNK around the mantel. on the right side of the picture contains red flowers. theres a picture on the wall facing the sofa. theres a sofa on the wall.

Automatic evaluation

GÖTEBORGS UNIVERSITET

Metric	g	b2	s25	s50	s75	s100	st50	n25	n50	n95	db2
BLEU-1	**37.16**	30.79	33.82	34.57	33.84	33.91	36.48	34.11	34.36	33.61	*37.08*
BLEU-2	**23.90**	19.86	18.54	19.20	18.54	18.29	22.20	18.70	19.07	18.46	*23.85*
BLEU-3	**15.53**	13.13	10.07	10.77	10.09	9.99	13.67	10.30	10.67	10.25	*15.51*
BLEU-4	**9.54**	8.02	4.81	5.40	4.95	5.00	7.89	5.29	5.62	5.15	*9.52*
METEOR	**14.22**	12.97	12.53	12.79	12.53	12.46	14.00	12.67	12.80	12.58	*14.20*
ROUGE-L	*30.64*	**30.71**	23.86	23.77	23.56	23.29	28.48	23.15	23.75	23.79	30.55
CIDEr	*16.62*	12.30	10.48	11.30	9.78	9.54	16.51	10.54	10.76	10.56	**16.80**
WMD	39.80	39.10	38.40	38.41	38.17	38.06	**40.26**	38.28	38.34	38.33	*39.84*

CLASP centre for linguistic theory and studies in probability

Automatic evaluation

Metric	g	b2	s25	s50	s75	s100	st50	n25	n50	n95	db2
BLEU-1	**37.16**	30.79	33.82	34.57	33.84	33.91	36.48	34.11	34.36	33.61	*37.08*
BLEU-2	**23.90**	19.86	18.54	19.20	18.54	18.29	22.20	18.70	19.07	18.46	*23.85*
BLEU-3	**15.53**	13.13	10.07	10.77	10.09	9.99	13.67	10.30	10.67	10.25	*15.51*
BLEU-4	**9.54**	8.02	4.81	5.40	4.95	5.00	7.89	5.29	5.62	5.15	*9.52*
METEOR	**14.22**	12.97	12.53	12.79	12.53	12.46	14.00	12.67	12.80	12.58	*14.20*
ROUGE-L	*30.64*	**30.71**	23.86	23.77	23.56	23.29	28.48	23.15	23.75	23.79	30.55
CIDEr	*16.62*	12.30	10.48	11.30	9.78	9.54	16.51	10.54	10.76	10.56	**16.80**
WMD	39.80	39.10	38.40	38.41	38.17	38.06	**40.26**	38.28	38.34	38.33	*39.84*

Human evaluation and correlation

- We compute Pearson's, Spearman's and Kendall's correlation between human and automatic evaluation scores.
- Given an image and a paragraph describing the image, rate it in terms of:
  - correctness: does the text describe objects correctly?
  - relevance: does the text describe relevant objects?
  - composition/flow: do object descriptions naturally follow each other?

Correctness vs automatic metrics

- Correctness: does the text describe objects correctly?

		g			**b2**			**s50**			**st50**			**n50**			**db2**		
		P	*S*	*K*	*P*	*S*	*K*	*P*	*S*	*K*	*P*	*S*	*K*	*P*	*S*	*K*	*P*	*S*	*K*
C	BLEU_1	0.14	0.13	0.09	0.11	0.12	0.09	0.02	0.01	0.0	0.06	0.11	0.07	0.22	0.19	0.13	0.19	0.18	0.12
	BLEU_2	0.12	0.08	0.06	0.15	0.15	0.12	-0.05	-0.12	-0.09	0.09	0.13	0.09	0.13	0.1	0.06	0.21	0.18	0.12
	BLEU_3	0.02	0.05	0.03	0.09	0.11	0.08	-0.09	-0.12	-0.09	0.11	0.13	0.08	0.12	0.1	0.06	0.18	0.18	0.13
	BLEU_4	-0.12	-0.02	-0.03	0.02	0.09	0.06	-0.0	-0.13	-0.1	0.1	0.14	0.09	0.22	0.15	0.11	0.17	0.22	0.16
	METEOR	-0.15	-0.13	-0.08	0.09	0.08	0.06	-0.19	-0.17	-0.13	-0.09	-0.1	-0.07	-0.16	-0.24	-0.16	-0.27	-0.27	-0.2
	ROUGE_L	0.13	0.19	0.14	0.06	0.08	0.05	-0.07	-0.09	-0.07	-0.02	0.02	0.02	0.22	0.19	0.14	0.16	0.17	0.11
	CIDEr	0.03	0.12	0.09	0.14	0.1	0.05	-0.0	-0.01	-0.01	-0.07	0.09	0.08	0.22	0.26	0.17	0.12	0.21	0.16
	WMD	-0.02	-0.03	-0.02	0.16	0.13	0.1	-0.22	-0.17	-0.12	-0.09	-0.07	-0.05	-0.22	-0.28	-0.21	-0.1	-0.09	-0.08

- No significant correlation, generally more negative scores for sampling-based decodings.

Relevance vs automatic metrics

- Relevance: does the text describe relevant objects?

		g			b2			s50			st50			n50			db2		
		P	*S*	*K*	*P*	*S*	*K*	*P*	*S*	*K*	*P*	*S*	*K*	*P*	*S*	*K*	*P*	*S*	*K*
R	BLEU_1	0.23	0.18	0.13	0.3	0.28	0.22	-0.01	-0.06	-0.03	-0.06	-0.03	-0.02	0.25	0.21	0.15	0.27	0.19	0.15
	BLEU_2	0.21	0.17	0.12	0.34	0.28	0.2	-0.04	-0.16	-0.1	-0.13	-0.15	-0.11	0.14	0.1	0.06	0.3	0.19	0.14
	BLEU_3	0.14	0.16	0.1	0.29	0.22	0.17	-0.05	-0.12	-0.07	-0.15	-0.2	-0.14	0.11	0.1	0.07	0.27	0.21	0.16
	BLEU_4	0.01	0.1	0.07	0.26	0.24	0.18	0.04	-0.1	-0.04	-0.12	-0.16	-0.12	0.19	0.11	0.08	0.2	0.22	0.16
	METEOR	-0.21	-0.18	-0.13	0.14	0.12	0.09	-0.21	-0.22	-0.16	-0.22	-0.32	-0.22	-0.05	-0.09	-0.06	-0.26	-0.24	-0.19
	ROUGE_L	0.18	0.15	0.11	0.22	0.23	0.16	0.06	0.02	0.02	-0.19	-0.22	-0.15	0.19	0.16	0.11	0.28	0.21	0.15
	CIDEr	0.02	0.15	0.1	0.33	0.17	0.12	-0.06	-0.17	-0.1	-0.15	-0.18	-0.11	0.23	0.23	0.17	0.16	0.19	0.14
	WMD	-0.0	0.0	-0.0	0.2	0.16	0.1	-0.14	-0.09	-0.06	-0.12	-0.14	-0.1	-0.09	-0.09	-0.06	-0.14	-0.12	-0.09

- Occasional significant correlation: positive for rule-based and negative for sampling-based decodings.

Composition vs automatic metrics

- Composition: do object descriptions naturally follow each other?

		g			b2			s50			st50			n50			db2		
		P	*S*	*K*	*P*	*S*	*K*	*P*	*S*	*K*	*P*	*S*	*K*	*P*	*S*	*K*	*P*	*S*	*K*
F	BLEU_1	0.41	0.37	0.27	0.42	0.4	0.31	-0.22	-0.24	-0.19	0.01	0.0	0.01	0.13	0.08	0.06	0.32	0.32	0.24
	BLEU_2	0.39	0.36	0.28	0.38	0.29	0.23	-0.18	-0.27	-0.21	-0.01	-0.04	-0.03	0.07	0.05	0.03	0.32	0.31	0.22
	BLEU_3	0.29	0.32	0.23	0.35	0.25	0.19	-0.22	-0.25	-0.18	0.01	-0.0	0.0	0.12	0.07	0.05	0.3	0.3	0.22
	BLEU_4	0.15	0.24	0.18	0.23	0.2	0.14	-0.01	-0.17	-0.12	0.03	0.05	0.03	0.19	0.06	0.04	0.22	0.24	0.17
	METEOR	-0.07	-0.07	-0.08	0.12	0.09	0.06	-0.11	-0.14	-0.1	-0.0	-0.06	-0.03	-0.12	-0.2	-0.16	-0.01	-0.01	-0.01
	ROUGE_L	0.31	0.29	0.22	0.24	0.24	0.18	-0.06	-0.1	-0.08	-0.08	-0.07	-0.05	0.16	0.12	0.09	0.28	0.29	0.19
	CIDEr	0.16	0.27	0.2	0.36	0.27	0.21	-0.35	-0.37	-0.28	0.01	0.02	0.02	0.02	0.1	0.07	0.13	0.29	0.23
	WMD	0.04	0.03	0.02	0.19	0.22	0.14	-0.14	-0.16	-0.12	-0.03	-0.02	-0.03	-0.02	-0.03	-0.03	0.1	0.05	0.03

- Strong positive correlation between rule-based decodings and BLEU metric, the same is observed for some other metrics. Strong negative correlation between sampling and the CIDEr metric.

Non-grounded and grounded evaluation

GÖTEBORGS UNIVERSITET

	ref	g	b2	s25	s50	s75	s100	st50	n25	n50	n95	db2
s1	2.9	0.9	0.3	1.0	1.1	1.1	1.0	0.9	1.1	1.1	1.1	0.9
s2	1.6	1.7	1.5	1.7	1.7	1.8	1.8	1.7	1.8	1.8	1.7	1.8
s3	1.4	1.6	1.4	1.7	1.8	1.8	1.8	1.6	1.8	1.8	1.8	1.6
s4	1.3	1.6	1.4	1.8	1.8	1.8	1.9	1.6	1.8	1.9	1.8	1.6
s5	1.2	1.7	1.4	1.8	1.8	1.8	1.7	1.6	1.7	1.8	1.7	1.7

Table: Average number of noun phrases generated by different inference algorithms.

CLASP centre for linguistic theory and studies in probability

Non-grounded and grounded evaluation

GÖTEBORGS UNIVERSITET

	ref	g	b2	s25	s50	s75	s100	st50	n25	n50	n95	db2
s1	2.9	0.9	0.3	1.0	1.1	1.1	1.0	0.9	1.1	1.1	1.1	0.9
s2	1.6	1.7	1.5	1.7	1.7	1.8	1.8	1.7	1.8	1.8	1.7	1.8
s3	1.4	1.6	1.4	1.7	1.8	1.8	1.8	1.6	1.8	1.8	1.8	1.6
s4	1.3	1.6	1.4	1.8	1.8	1.8	1.9	1.6	1.8	1.9	1.8	1.6
s5	1.2	1.7	1.4	1.8	1.8	1.8	1.7	1.6	1.7	1.8	1.7	1.7

Table: Average number of noun phrases generated by different inference algorithms.

	g	b2	s25	s50	s75	s100	st50	n25	n50	n95	db2
s1	0.18	0.10	0.10	0.13	0.11	0.08	0.14	0.12	0.10	0.13	0.18
s2	0.17	0.17	0.13	0.13	0.13	0.14	0.17	0.12	0.15	0.13	0.17
s3	0.13	0.12	0.10	0.10	0.09	0.11	0.11	0.10	0.11	0.10	0.13
s4	0.10	0.09	0.10	0.09	0.08	0.09	0.09	0.09	0.09	0.09	0.10
s5	0.10	0.10	0.07	0.07	0.08	0.07	0.08	0.07	0.07	0.08	0.10

Table: Dice similarity coefficient between the set of objects described in reference texts and texts generated by different decoding algorithms.

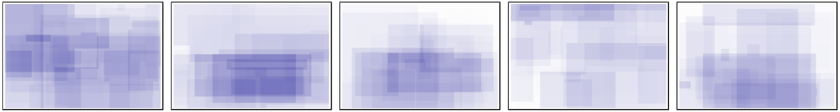
The effect of linking on grounded evaluation

	g	b2	s25	s50	s75	s100	st50	n25	n50	n95	db2
s1	69.5	72.7	46.5	46.3	43.2	46.1	66.5	51.1	45.2	50.5	69.5
s2	65.6	65.1	47.2	49.0	43.8	46.6	58.8	44.7	50.3	47.7	65.5
s3	61.6	59.5	43.6	46.9	40.5	45.1	53.1	40.1	40.6	44.7	60.7
s4	55.4	57.6	43.7	42.7	44.7	41.0	52.5	45.0	43.7	38.3	55.7
s5	60.5	57.4	47.6	43.2	43.4	43.7	53.3	39.2	38.9	44.4	59.3

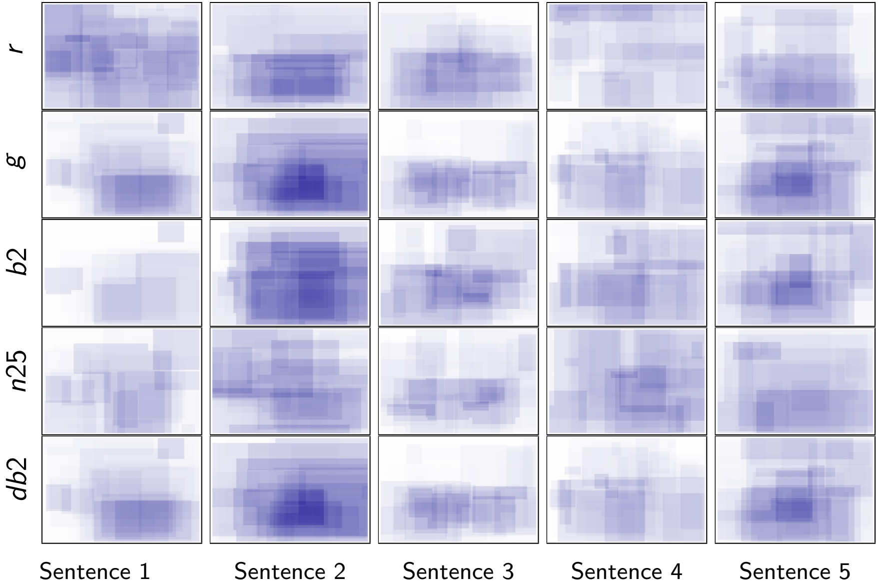
Table: Average proportion of successful linking (in percent) between noun phrases in generated texts and image objects.

Attentional structure of discourse in references

For more, see Dobnik et al. (2022).



Sentence 1 Sentence 2 Sentence 3 Sentence 4 Sentence 5


r
g
b2
n25
db2
Sentence 1
Sentence 2
Sentence 3
Sentence 4
Sentence 5

Conclusions

GÖTEBORGS UNIVERSITET

- Image paragraph generation is challenging; one of the reasons is the problem of modelling *discourse organisation*.
- There is a disagreement between automatic metrics and human evaluation concerning discourse in texts generated by different decoding methods. e,g, no single decoding is a good fit for everything.
- A large discrepancy between referring expressions.
- Future work must develop better models and evaluation metrics to learn and generate better discourses in image paragraphs.

CLASP centre for linguistic theory and studies in probability

References I

GÖTEBORGS UNIVERSITET

Simon Dobnik, Nikolai Ilinykh, and Aram Karimi. 2022. What to refer to and when? reference and re-reference in two language-and-vision tasks. In *Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*, pages 146–159, Dublin, Ireland. SEMDIAL.

Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. 2019. Image captioning: Transforming objects into words. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Nikolai Ilinykh, Sina Zarrieß, and David Schlangen. 2019. Tell me more: A dataset of visual scene description sequences. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 152–157, Tokyo, Japan. Association for Computational Linguistics.

CLASP centre for linguistic theory and studies in probability