On Decoding and Discourse Structure in Multi-Modal Text Generation

UNIVERSITY OF GOTHENBURG

Nikolai Ilinykh and Simon Dobnik

Centre for Linguistic Theory and Studies in Probability, Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg, Sweden

`name.surname@gu.se`

Key questions

- What are the properties of and differences between natural and generated discourses of image descriptions?
- How does a choice of a decoding strategy affect reference to scene entities?
- What metrics can be used to evaluate discourse structure of referring?

Paragraph generation

- **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.

Decodings tested

- **G** - greedy
- **B2** - beam search with size 2
- **Sk** - sampling from top-k tokens
- **St50** - sampling with temperature $\tau = 0.5$
- **Np** - nucleus sampling, top-p of the mass
- **DB2** - diverse beam search with size 2

Do decoding strategies …

… sequence reference to entities similarly to humans?

No, the progression of noun phrases differs compared to natural sentences.

	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

… refer to the same entities as humans?

Rule-based decodings better approximate human strategies.

	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

… describe objects correctly?

Rule-based decoding strategies are more correct - but temperature can mitigate negative effect of sampling.

	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

Evaluation with humans

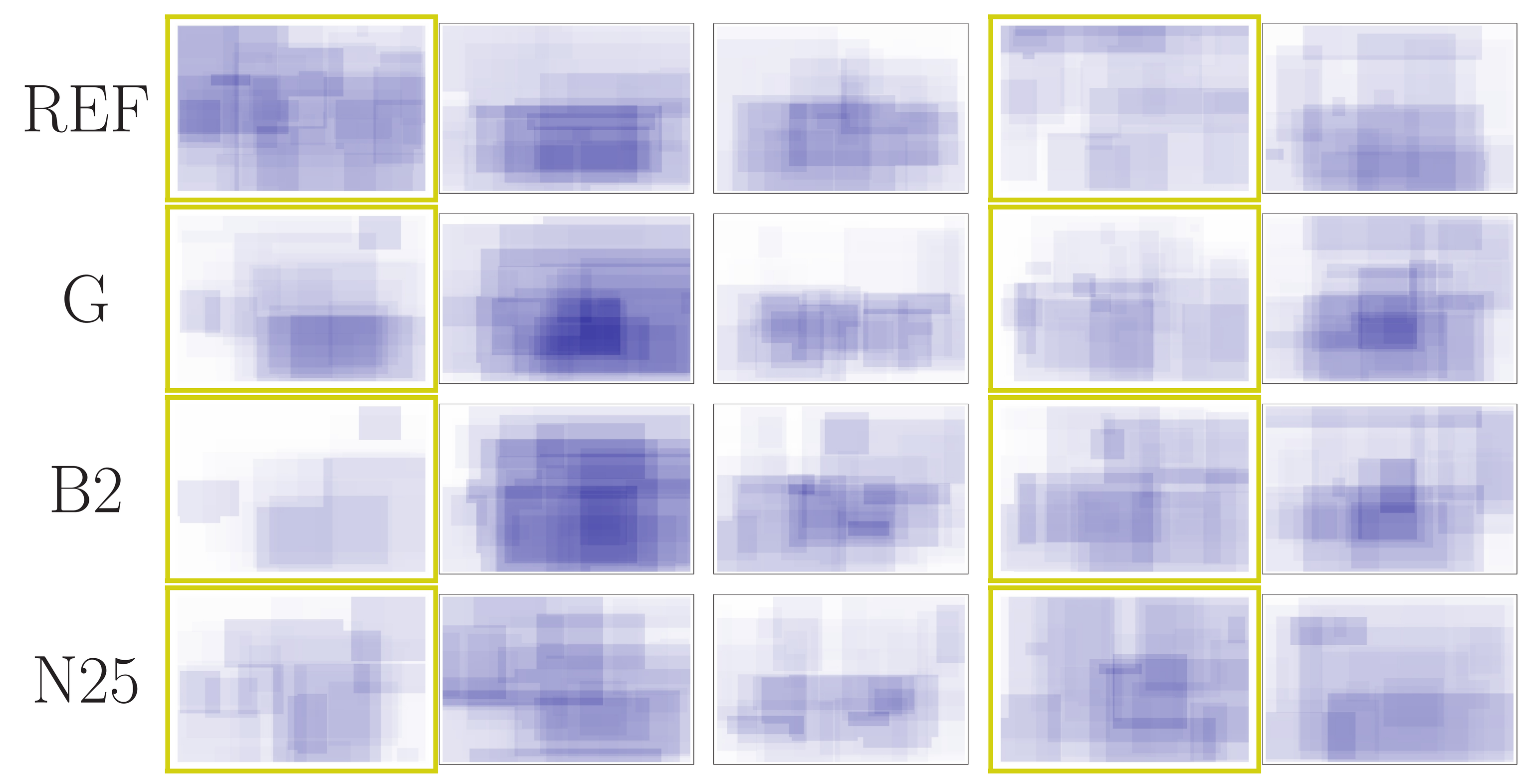
Relevance, correctness, flow

Individual decoding strategies correlate with different judgement aspects.

	g			**b2**			**s50**			**st50**			**n50**			**db2**		
	P	*S*	*K*	*P*	*S*	*K*	*P*	*S*	*K*	*P*	*S*	*K*	*P*	*S*	*K*	*P*	*S*	*K*
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

Reference and attention

Decodings "attend" images differently from humans as a narrative progresses.



CLASP centre for linguistic theory and studies in probability