Are Current Decoding Strategies Capable of Facing the Challenges of Visual Dialogue?

Amit Kumar Chaudhary
CIMeC - University of Trento
amitkumar.chaudhar@unitn.it

Alex J. Lucassen
CIMeC - University of Trento
alex.lucassen@unitn.it

Ioanna Tsani
CIMeC - University of Trento
ioanna.tsani@unitn.it

Alberto Testoni
DISI - University of Trento
alberto.testoni@unitn.it

Overview

- The last few years have witnessed remarkable progress in developing efficient generative language models. The choice of the decoding strategy plays a crucial role in the quality of the output (see Zarrieß et al., 2021).
- We can group decoding strategies into two main classes. Decoding strategies that maximise the likelihood of the generated sequence and strategies that take random samples of the model.
- Multimodal Vision & Language dialogue systems are receiving an increasing interest from the research community for their numerous applications, but decoding strategies are usually designed and evaluated in text-only tasks.

Mutimodal Goal-Oriented Dialogue Systems

1. The generated text must be coherent with the visual context it refers to.
2. The output must be informative to solve the task.
3. The dialogue history must be coherent and it should allow the speaker to incrementally solve the task.

The GuessWhat Game

Questioner	Oracle
Is it a vase?	Yes
Is it partially visible?	No
Is it in the left corner?	No
Is it the turquoise and purple one?	Yes

Figure 1: Example of a GuessWhat game from De Vries et al. (2017).

Results

	Accuracy (%) ↑	CHAIR-i ↓	CHAIR-s ↓	% games with repetitions ↓	Vocabulary Size ↑	Rare Words ↑
Confirm-it	**51.39**	**15.09**	**28.48**	30.33	858	34
Beam Search (beam size = 3)	**47.05**	18.33	**31.08**	38.49	731	27
Nucleus Sampling ($p = 0.3$)	**46.92**	**17.96**	33.60	32.35	1016	78
Greedy Search	46.58	**17.75**	**32.97**	35.63	834	46
Typical Decoding ($\tau = 0.7$)	45.45	21.84	37.81	**16.18**	**1703**	**247**
Top-k Sampling ($k = 5$)	45.10	22.84	37.71	**14.93**	**1462**	**171**
Pure Sampling	43.13	26.55	43.23	**8.32**	**2609**	**793**

Table 1: Comparison between decoding strategies and their best-performing (in terms of accuracy) hyper-parameters. The decoding strategies are sorted by accuracy.

There exists a trade-off between informativeness / visual grounding and linguistic quality. Strategies that reach the highest accuracy and lowest hallucination rate are also the ones with more repetitions and less lexical variability. Stochastic strategies show interesting properties about lexical richness, but they clearly decrease repetitions and increase vocabulary richness by generating tokens that are not related to the source input (high hallucination rate).

In-Depth Analysis

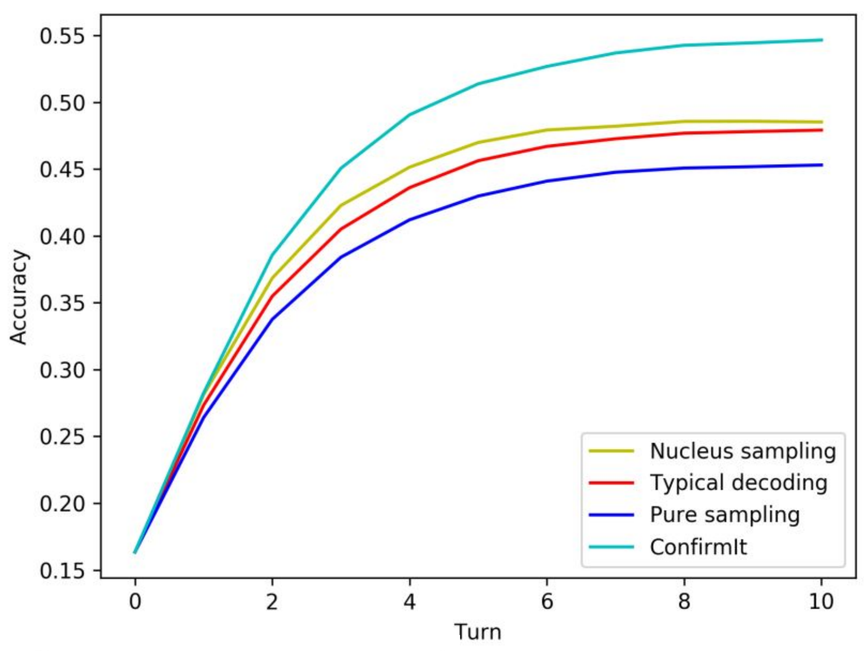

Figure 2: Incremental accuracy per dialogue turn for four different decoding strategies for dialogues of length 10. Confirm-it clearly outperforms the other strategies in generating more effective questions that incrementally identify the target object.

	Human Accuracy (%) ↑
Confirm-it	**72.5**
Typical Sampling ($\tau = 0.7$)	68.0
Nucleus Sampling ($p = 0.3$)	67.5
Pure Sampling	59.5

Table 2: Human evaluation accuracy results. We asked human annotators to play the guessing game by reading dialogues generated by different decoding strategies.

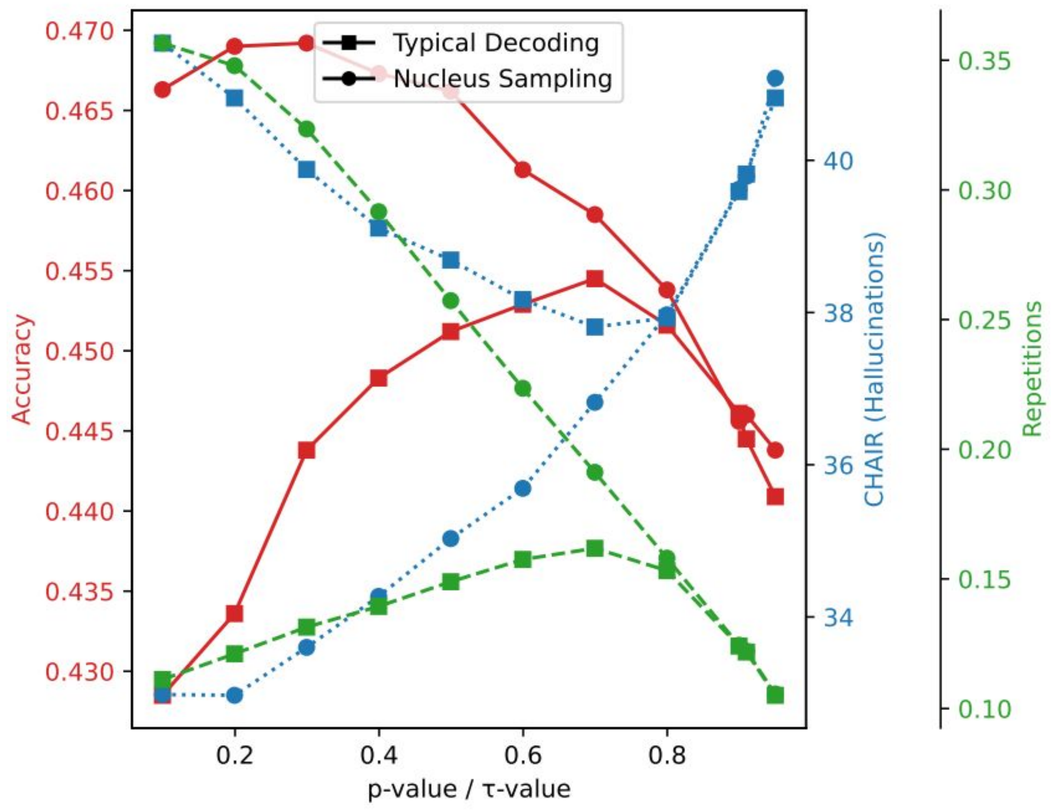

Figure 3: Different hyper-parameter values and their effect on the accuracy, hallucinations, and repetitions in typical decoding and nucleus sampling. Hyper-parameters play a crucial role in balancing the different metrics analyzed in the paper.

Conclusions

- Decoding algorithms that lead to the highest accuracy in the task and the lowest hallucination rate, at the same time generate highly repetitive text and use a restricted vocabulary.
- Hyper-parameter configuration plays a crucial role in stochastic strategies, and reveals an interesting trade-off between lexical variety, hallucination rate, and task accuracy.
- Taking into account the model's intermediate predictions about the referent, as *Confirm-it* does, represents a promising direction also for stochastic strategies, aiming at preserving their lexical richness while reducing hallucinations.