EMNLP 2023

Critic-Driven Decoding for Mitigating Hallucinations in Data-to-text Generation

Mateusz Lango, Ondřej Dušek

December 6-10, 2023

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

Hallucinations in the Data-to-Text task

This work is about:

- **Data-to-text**: generating textual description for given data
  - Example:

(Alex Plante, birth year, 1989), (Alex Plante, birth place, Manitoba)
$\Rightarrow$
Alex Plante was born in 1989 in Manitoba.

- **Hallucinations**: generated text lacks grounding in the input data
  - Can lead to inaccurate or misleading information
  - Undermines quality and reliability of the output

Motivations

- Current approaches: train a new model
  + require annotation / training procedure change / network architecture change
- Text classifiers: can find incoherence between data & generated text
  (e.g. NLI-based metrics)

Our approach:

☞ Can be used with **any existing LM**, only modifies decoding

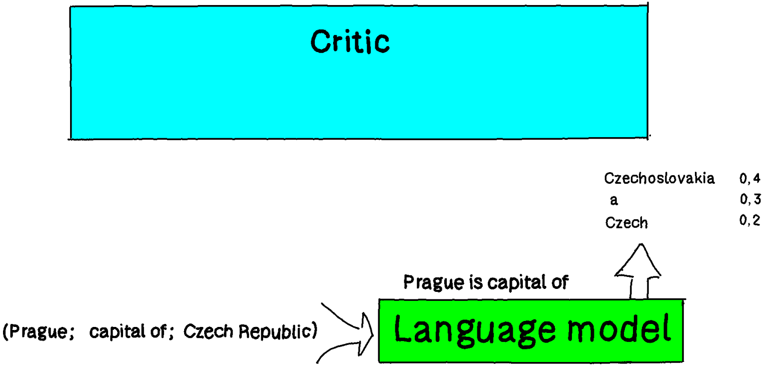
☞ Uses a **text critic classifier** to guide the decoding

☞ Checks **match** between **data & text generated so far**

Critic-driven decoding

- Text critic classifier: compares **input data** vs. **text generated so far**
- LM word probability multiplied by critic's assessment of correctness

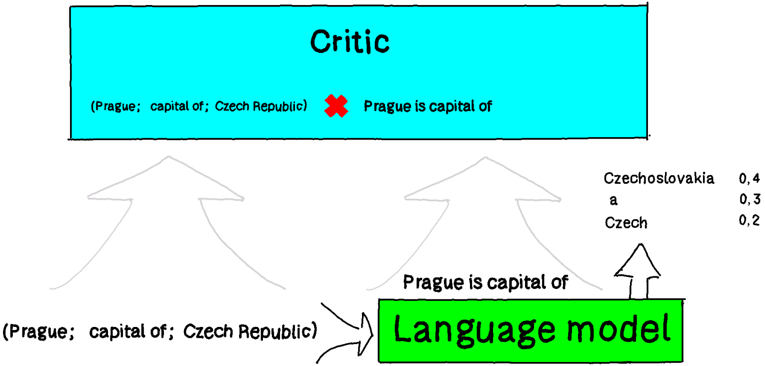
$$P(y_i|y_{\leq i-1}, x, c) \propto P(c|y_{\leq i}, x)P(y_i|y_{\leq i-1}, x)$$



Critic-driven decoding

- Text critic classifier: compares **input data** vs. **text generated so far**
- LM word probability multiplied by critic's assessment of correctness

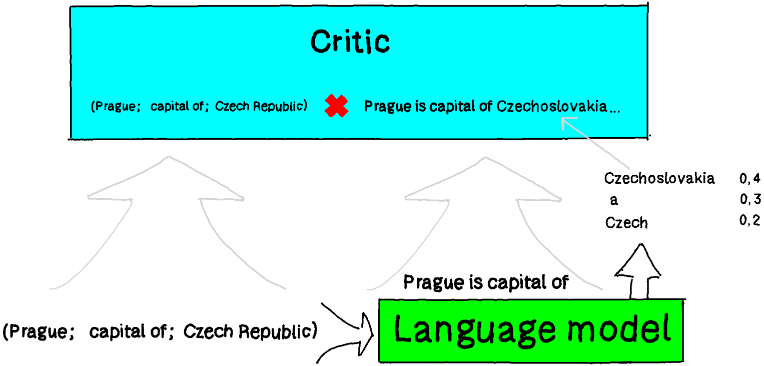
$$P(y_i|y_{\leq i-1}, x, c) \propto P(c|y_{\leq i}, x)P(y_i|y_{\leq i-1}, x)$$



Critic-driven decoding

- Text critic classifier: compares **input data** vs. **text generated so far**
- LM word probability multiplied by critic's assessment of correctness

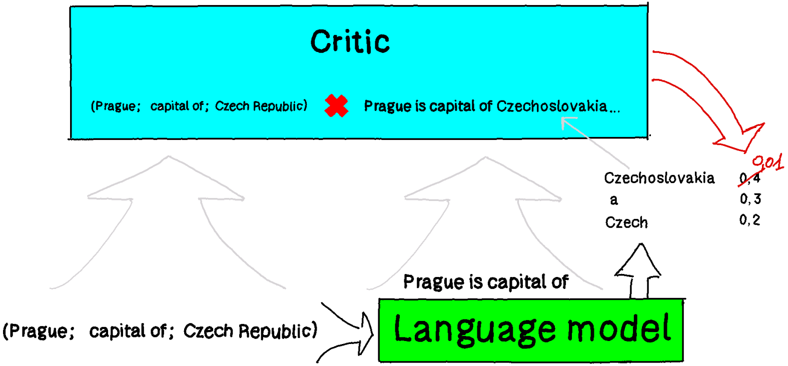
$$P(y_i|y_{\leq i-1}, x, c) \propto P(c|y_{\leq i}, x)P(y_i|y_{\leq i-1}, x)$$



Critic-driven decoding

- Text critic classifier: compares **input data** vs. **text generated so far**
- LM word probability multiplied by critic's assessment of correctness

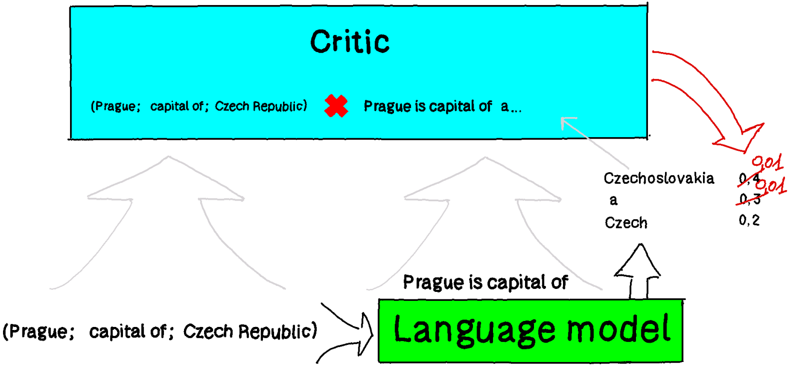
$$P(y_i|y_{\leq i-1}, x, c) \propto P(c|y_{\leq i}, x) P(y_i|y_{\leq i-1}, x)$$



Critic-driven decoding

- Text critic classifier: compares **input data** vs. **text generated so far**
- LM word probability multiplied by critic's assessment of correctness

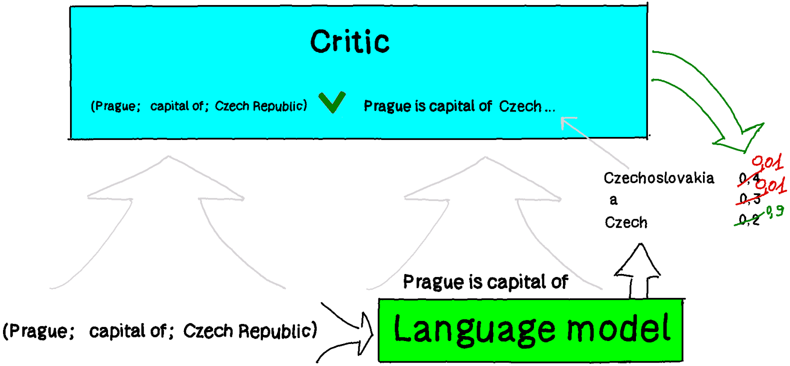
$$P(y_i|y_{\leq i-1}, x, c) \propto P(c|y_{\leq i}, x)P(y_i|y_{\leq i-1}, x)$$



Critic-driven decoding

- Text critic classifier: compares **input data** vs. **text generated so far**
- LM word probability multiplied by critic's assessment of correctness

$$P(y_i|y_{\leq i-1}, x, c) \propto P(c|y_{\leq i}, x) P(y_i|y_{\leq i-1}, x)$$



Generating critic training data

- **Positive training examples** = all prefixes from LM training data

The A-Rosa Luna is powered by a MTU Friedrichshafen engine and is 125.8 metres in length.

$\Rightarrow$

The
The A-Rosa
The A-Rosa Luna
The A-Rosa Luna is
...

Generating critic training data

Negative training examples: 5 approaches

1. **base** – replacing random words

   "The Cruises", "The A-Rosa the", "The A-Rosa Luna located", ...

2. **base with full sentences** – replacing random sentences
3. **vanilla LM** – replacing words by sampling from (unconditioned) LM

   "The United", "The A-Rosa is", "The A-Rosa Luna powers", ...

4. **fine-tuned LM** – words sampled from data-conditioned LM

   "The A-Rosa Luna is 125.8m", "The A-Rosa Luna is supplied", , ...

5. **fine-tuned LM with full sentences** – full sentences from data-conditioned LM

Results - manual evaluation

- 100 instances from WebNLG test set
- Base LM: fine-tuned BART
- Minor and major hallucinations, omissions, disfluencies, repetitions
- Overall relative ranking of text quality

decoding approach	min. hal.	maj. hal.	omi.	disfl.	rep.	avg. rank
baseline	0.22	0.40	0.25	0.20	0.08	3.61
1. critic (base)	0.21	0.30	**0.20**	0.17	**0.04**	**3.38**
2. critic (base with full sent.)	0.21	**0.29**	0.27	**0.11**	0.08	3.43
3. critic (vanilla LM)	**0.18**	**0.29**	0.23	0.19	0.05	3.54
4. critic (fine-tuned LM)	0.22	0.37	0.26	0.21	0.07	3.53
5. critic (fine-tuned LM with full sent.)	0.20	0.37	0.26	0.18	0.07	3.54

Results - summary

Automatic evaluation

- BLEU, METEOR, BERTScore, NLI, BLEURT on WebNLG and OpenDialKG
- ☞ Stat. significant improvements of hallucination-oriented measures (NLI, BLEURT)
- ☞ Text quality measures not affected (sometimes slight improvements)

Results - summary

Automatic evaluation

- BLEU, METEOR, BERTScore, NLI, BLEURT on WebNLG and OpenDialKG

☞ Stat. significant improvements of hallucination-oriented measures (NLI, BLEURT)

☞ Text quality measures not affected (sometimes slight improvements)

Analysis of **changes introduced by critics**

☞ Many outputs intact (30-70%), most changes only a few words (2-5), length preserved

Results - summary

Automatic evaluation

- BLEU, METEOR, BERTScore, NLI, BLEURT on WebNLG and OpenDialKG

☞ Stat. significant improvements of hallucination-oriented measures (NLI, BLEURT)

☞ Text quality measures not affected (sometimes slight improvements)

Analysis of **changes introduced by critics**

☞ Many outputs intact (30-70%), most changes only a few words (2-5), length preserved

In-domain and out-of-domain evaluation on WebNLG

☞ Improvements on both, but more significant on out-of-domain

Results - summary

Automatic evaluation

- BLEU, METEOR, BERTScore, NLI, BLEURT on WebNLG and OpenDialKG
- Stat. significant improvements of hallucination-oriented measures (NLI, BLEURT)
- Text quality measures not affected (sometimes slight improvements)

Analysis of **changes introduced by critics**

- Many outputs intact (30-70%), most changes only a few words (2-5), length preserved

In-domain and out-of-domain evaluation on WebNLG

- Improvements on both, but more significant on out-of-domain

Additional results with **beam search** decoding

- Consistent improvements, similar to greedy (baseline) decoding

Takeaways

- **Critic-driven decoding** – a general method which can be coupled with different LMs and decoding strategies
- **Critic training data** can be generated with **simple strategies** (like replacing random words)
- The approach **significantly reduces hallucinations** without any changes to base LM

Paper: https://t.ly/HXnl9
Code: https://t.ly/563XE

Mateusz Lango
@LangoMateusz

Ondřej Dušek
@tuetschek

This research was supported by the European Research Council (Grant agreement No. 101039303 NG-NLG). It used resources provided by the LINDAT/CLARIAH-CZ Research Infrastructure (Czech Ministry of Education, Youth, and Sports project No. LM2018101).