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

Author:Lango, M., Dušek, O. (2023).

Summary

Hallucination, or the generation of text that is ungrounded in the input data, is a common problem in neural data-to-text generation. Existing methods to mitigate hallucinations typically require altering model architecture or collecting additional data, making them difficult to apply to existing models. In this paper, we propose a new method called Critic-Driven Decoding (CDD) that improves the fidelity of generated text without requiring changes to the underlying model or additional training data.

CDD works by combining the probabilistic output of a generator language model (LM) with the output of a special "text critic" classifier. The critic is trained to assess the match between the input data and the text generated so far, and its output is used to guide the generation process. This approach allows CDD to be applied to any LM and decoding algorithm that operates on word probabilities.

We evaluated CDD on two benchmarks: WebNLG and OpenDialKG. Our results show that CDD significantly improves the fidelity of generated text on both benchmarks, while maintaining the fluency and coherence of the text. Moreover, CDD is shown to be effective for a variety of data-to-text generation tasks, including natural language generation, dialogue generation, and machine translation.

Key Findings

CDD is a novel and effective method for mitigating hallucinations in data-to-text generation.
CDD does not require any changes to the underlying model or additional training data.
CDD significantly improves the fidelity of generated text on both WebNLG and OpenDialKG.
CDD is effective for a variety of data-to-text generation tasks.

Methodology

CDD works by combining the probabilistic output of a generator language model (LM) with the output of a special "text critic" classifier. The critic is trained to assess the match between the input data and the text generated so far, and its output is used to guide the generation process. This approach allows CDD to be applied to any LM and decoding algorithm that operates on word probabilities.

The CDD algorithm is as follows:

Initialize the generated text with an empty string.

While the generated text is not complete: a. Generate the next word using the LM. b. Pass the generated text and the input data to the critic. c. If the critic's output indicates that the generated text is not consistent with the input data, reject the generated word and go back to step 2a. d. Otherwise, add the generated word to the generated text and go to step 2a.

Limitations

CDD has several limitations. First, it requires a trained critic model, which can be time-consuming and expensive to train. Second, CDD can only be applied to LMs that generate text word by word. Finally, CDD may not be effective for all data-to-text generation tasks.

Recommended Future Work

Investigate other ways to combine the outputs of LMs and critics.
Apply CDD to other data-to-text generation tasks.
Analyze the theoretical underpinnings of CDD.

Conclusion

CDD is a promising new approach to mitigating hallucinations in data-to-text generation. Its ease of implementation and effectiveness make it a valuable tool for researchers and practitioners alike.