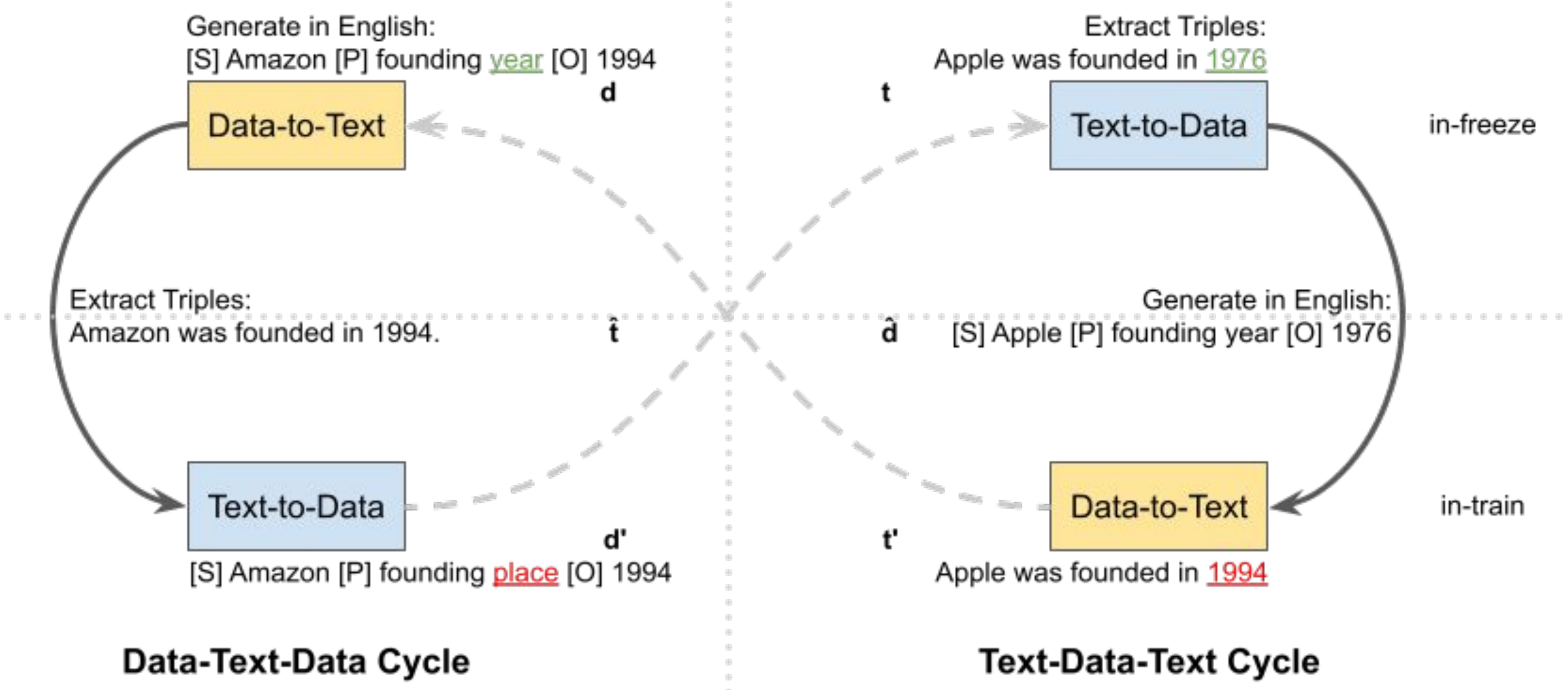
Faithful Low-resource Data-to-Text Generation through Cycle Training

Zhuoer Wang, Marcus Collins, Nikhita Vedula, Simone Filice, Shervin Malmasi, Oleg Rokhlenko

TEXAS A&M UNIVERSITY · amazon

Overview

Our work targets the task of data-to-text generation. Given multiple input triples, a model is expected to generate a fluent and faithful surface realization. Fine-tuning with Large Language Models has achieved strong performance, but it relies on human annotated data that is expensive and time-consuming to obtain. It also may suffer faithfulness issues when the amount of annotated data is limited. To overcome the aforementioned limitations, we adopt the Cycle Training approach.



Cycle training uses two models which are inverses of each other. It consists of the Data-Text-Data cycle that enforces the self-consistency of data and the Text-Data-Text cycle that enforces the self-consistency of text in a reverse manner. As illustrated in the figure above, the upper-level models are frozen to generate the intermediate inputs for the training of the lower-level models that attempt to reconstruct the initial inputs. Through iterative training between the two cycles, cycle training can converge to models with near-supervised performance while ensuring and even improving the faithfulness of the output.

Datasets

Dataset	Domain	Split Size (Train/Dev/Test)	Unique Predicates	Triples/Sample (Median/max)	Vocab size	Tokens/Sample (Median/max)
WebNLG	DBPedia (16 categories)	35,426/4,464/7,305	1,236	3 / 7	20,126	21 / 80
E2E	Restaurants	33,482/1,475/1,475	41	4 / 7	6,158	22 / 73
WTQ	Wikipedia (Open-domain)	3,253/361/155	5,013	2 / 10	11,490	13 / 107
WSQL	Wikipedia (Open-domain)	526/59/38	946	2 / 6	2,353	12 / 34

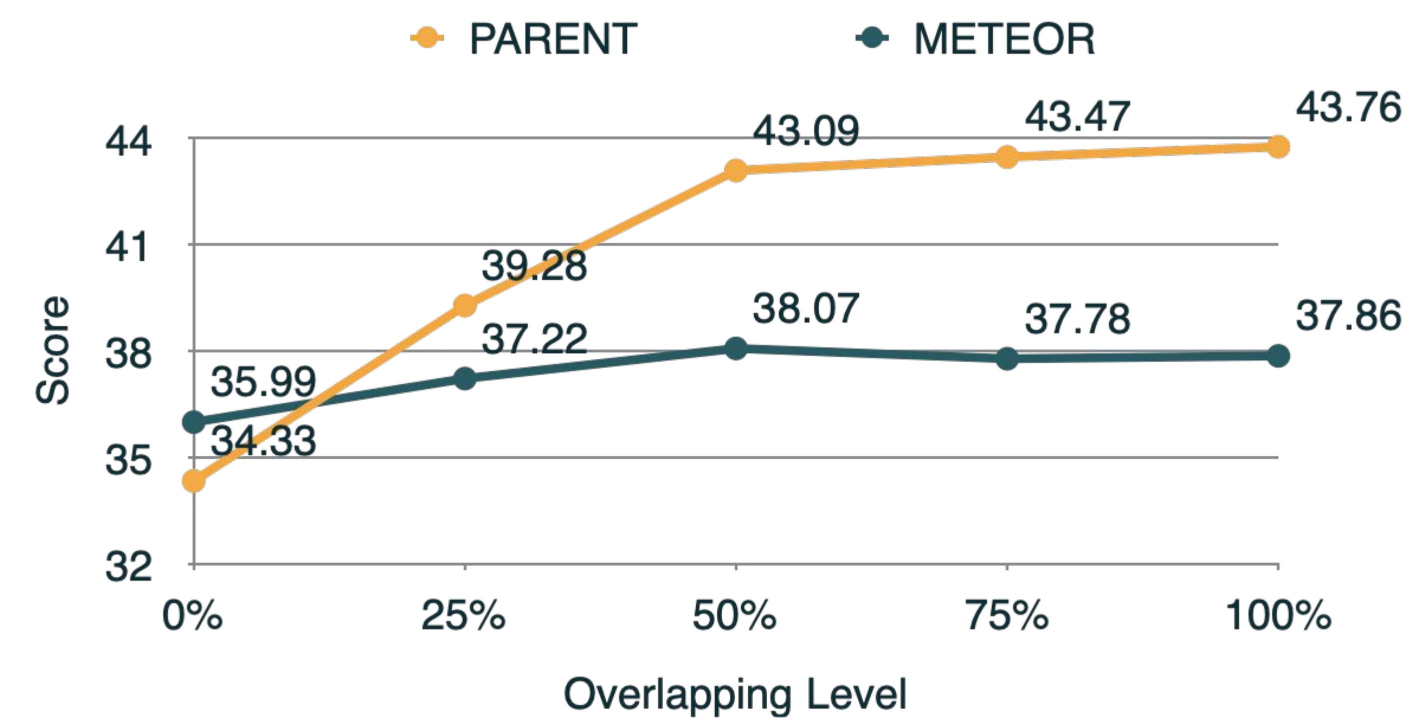
Experiments

- **Fully-supervised fine-tuning** with all labeled samples
- **Low-resource fine-tuning** with 100 labeled samples
- **Additional pretraining** on target domain text and **Low-resource fine-tuning** with 100 labeled samples
- **Unsupervised cycle training** with unpaired samples
- **Low-resource cycle training** with 100 labeled samples for fine-tuning and unpaired samples for cycle training

Method	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	BLEU	BERTScore	PARENT
Tested on WebNLG							
Fully-supervised fine-tuning	59.99	**40.93**	**49.32**	**39.76**	**42.83**	**95.41**	45.67
Low-resource fine-tuning	55.55	36.63	46.21	35.22	33.63	94.60	41.37
+ Additional pretraining	55.28	35.71	45.41	35.26	33.44	94.33	39.47
Unsupervised cycle training	58.65	37.70	46.18	37.98	36.36	94.42	43.24
Low-resource cycle training	**60.21**	40.56	48.71	39.74	41.77	95.18	**46.14**
Tested on WSQL							
Fully-supervised fine-tuning	58.27	32.77	48.40	**37.95**	22.97	**93.18**	24.00
Low-resource fine-tuning	56.37	31.60	49.42	33.57	23.34	92.57	23.68
+ Additional pretraining	56.01	30.92	47.00	35.34	21.18	92.24	22.66
Unsupervised cycle training	42.24	15.17	33.52	29.45	4.03	85.37	14.63
Low-resource cycle training	**58.72**	**33.13**	**51.01**	37.43	**25.60**	93.03	**25.84**

* Additional results on E2E and WTQ available in paper; **Bold**: best of all; Underlined: best of low-resource settings

- **Unsupervised cycle training at different overlapping levels**



* Additional evaluation with other metrics available in paper

Human Evaluation

- **A new quantitative annotation schema that features better objectiveness, consistency, and precision**
  - **Count of Factual Errors** measures the factual correctness of the generated text with respect to the entities (subject and object) and predicates of the input triplets. Factual errors are information in the generations that contradict the information in the input triplets.
  - **Count of Hallucination Errors** measures the relevance of the generated text with respect to the input triplets. Hallucination errors occur when words or phrases in the generation cannot be inferred from the input triplets. Unlike FEs, HEs add information not present in the triplets or reference, but do not directly contradict the triplets.
  - **Count of Information Misses** measures the information coverage of the generated text with respect to the predicates given in the input triplets.
  - **Fluency Preference** measures the quality of the generated text in terms of the grammar, structure, and coherence of the text.

Method	Factual Errors	Hallucination Errors	Information Misses	Fluency Preference
Low-resource fine-tuning	8.05	14.84	21.39	2.00
Low-resource cycle training	**0.49**	**2.57**	**3.36**	1.80
Fully-supervised fine-tuning	2.08	11.48	8.46	**1.73**

* Aggregated results, per dataset results available in paper; FE, HE, and IM are normalized, see details in paper

Main Findings

- **Cycle training, when initialized with a small amount of labeled samples, significantly improves the generation performance over the low-resource fine-tuning method, and it also achieves competitive performance with respect to the fully-supervised method.**
- **Compared to the fully-supervised fine-tuning approach and evident from the PARENT score as well as the human evaluation, low-resource cycle training generated texts have better faithfulness to the input data when applied to multi-domain and open-domain datasets (WebNLG, WTQ, and WSQL).**
- **When the size is the same, the unpaired data corpus and text corpus used for cycle training need to have at least 50% entities (or say, latent information) overlap to achieve performance at an ideal level.**