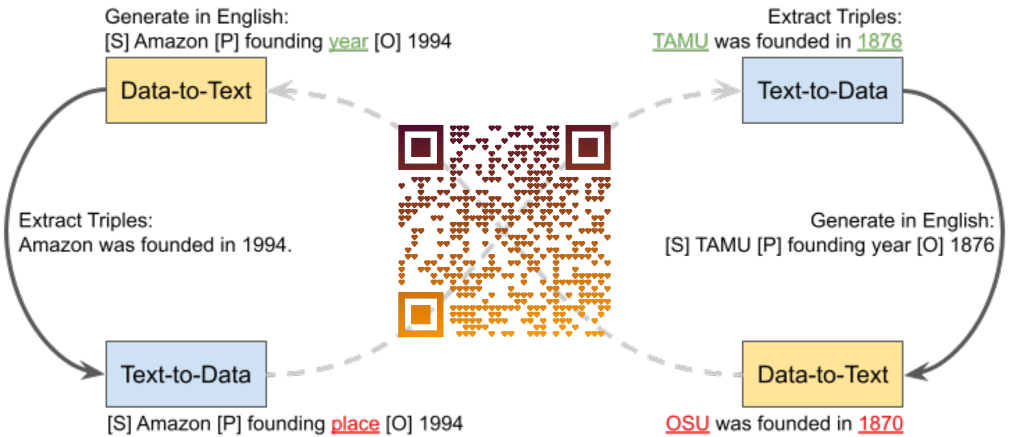
Faithful Low-resource Data-to-Text Generation through Cycle Training

Zhuoer Wang

Nikhita Vedula

Shervin Malmasi



Marcus Collins

Simone Filice

Oleg Rokhlenko

TEXAS A&M UNIVERSITY

amazon

Special thanks to Prof. **James Caverlee** @ Texas A&M University

Introduction

The Data-to-Text Generation Task

[S] The Fellowship of the Ring [P] preceded by [O] The Hobbit

[S] The Hobbit [P] release date [O] 1937-09-21

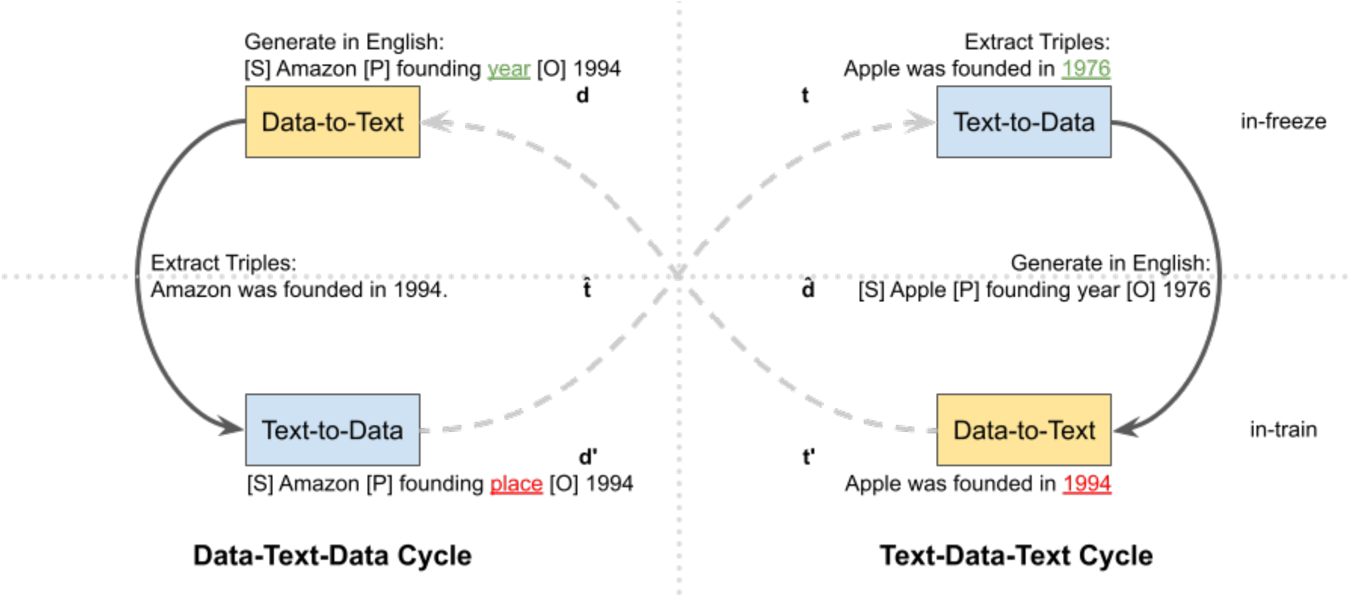
-> The Hobbit, was which published on September 21, 1937, came before The Fellowship of the Ring.

The Challenge

- **Reliance of human annotated data**
  - Expensive and time-consuming
- **Suffering faithfulness issues when data is limited**
  - Missing Information
  - Factual Errors
  - Hallucination Errors

Approach

- **Cycle Training:** A variable **x** and a bijective mapping function **f** should satisfy **x = g(f(x))**, where **g** is the inverse function of **f**

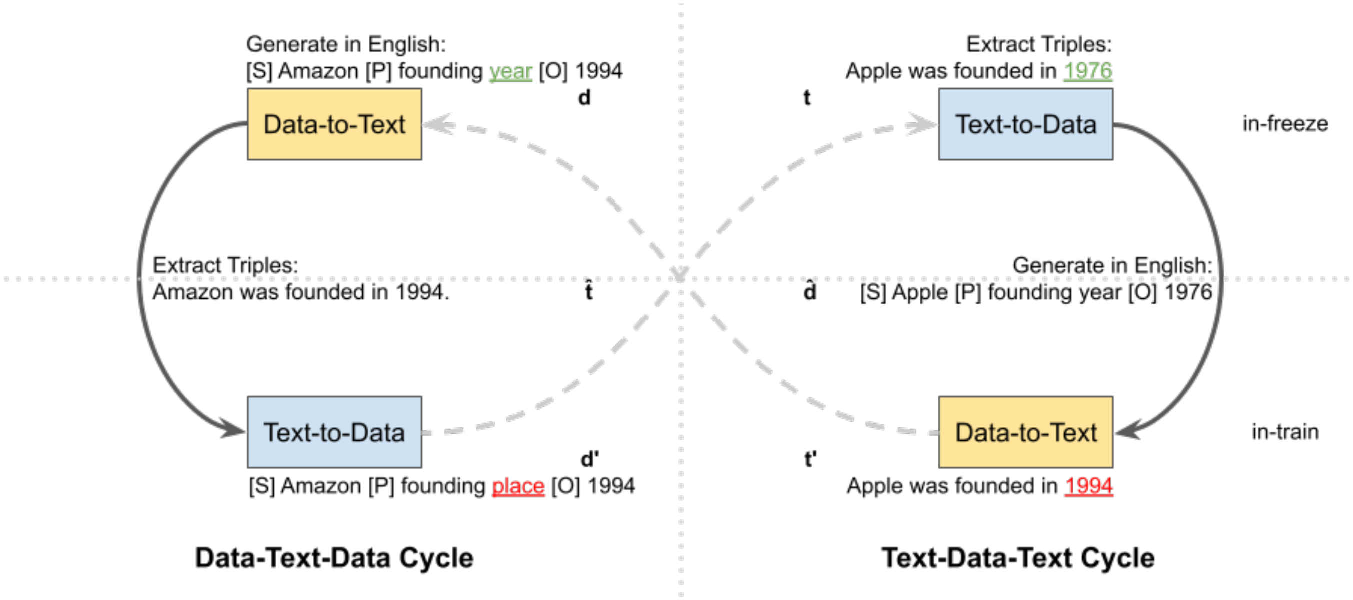


Generate in English:
[S] Amazon [P] founding year [O] 1994
d
Data-to-Text
Extract Triples:
Amazon was founded in 1994.
t̂
Text-to-Data
d'
[S] Amazon [P] founding place [O] 1994
Data-Text-Data Cycle
Extract Triples:
Apple was founded in 1976
t
Text-to-Data
in-freeze
Generate in English:
[S] Apple [P] founding year [O] 1976
d̂
Data-to-Text
in-train
t'
Apple was founded in 1994
Text-Data-Text Cycle

Cycle Training

- **Data-Text-Data Cycle:** enforces the self-consistency of data

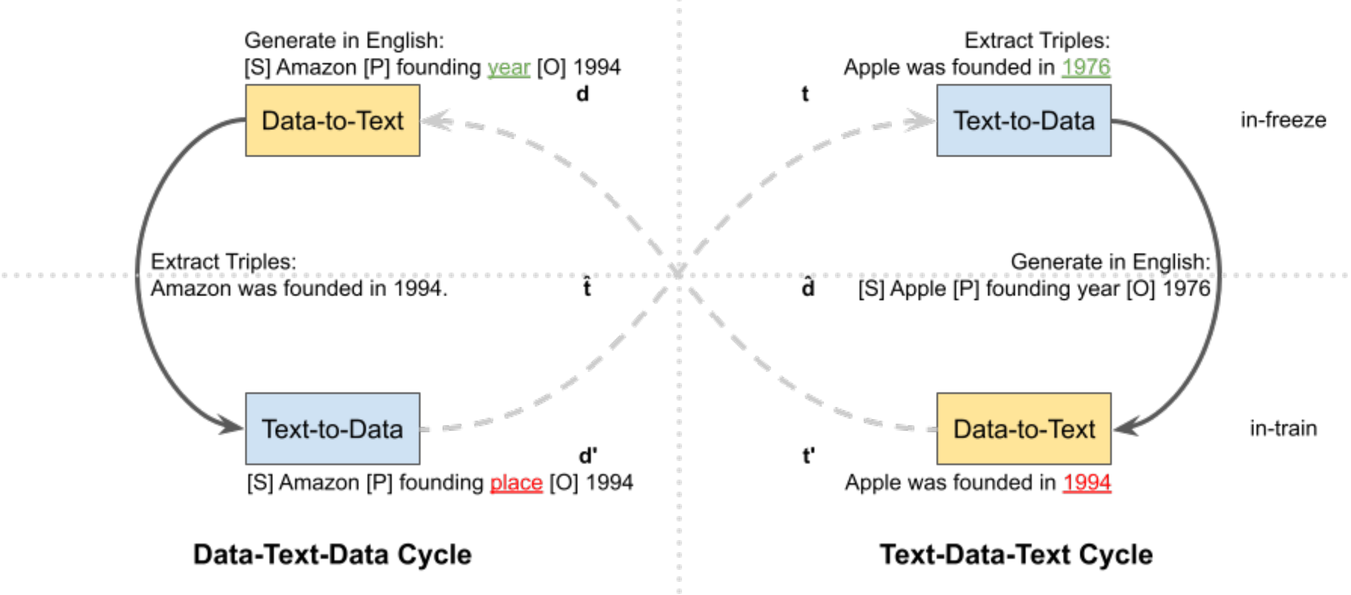
$$\mathcal{L}_{d'} = -\frac{1}{|d|}\sum_{i=0}^{|d|} \log p(d_i \mid d_0, \ldots, d_{i-1}, \hat{t})$$


Generate in English:
[S] Amazon [P] founding year [O] 1994
d
Data-to-Text
Extract Triples:
Amazon was founded in 1994.
t̂
Text-to-Data
d'
[S] Amazon [P] founding place [O] 1994
Data-Text-Data Cycle
Extract Triples:
Apple was founded in 1976
t
Text-to-Data
in-freeze
Generate in English:
[S] Apple [P] founding year [O] 1976
d̂
Data-to-Text
in-train
t'
Apple was founded in 1994
Text-Data-Text Cycle

Cycle Training

- **Text-Data-Text Cycle:** enforces the self-consistency of text

$$\mathscr{L}_{t'} = -\frac{1}{|t|}\sum_{i=0}^{|t|} \log p(t_i \,|\, t_0, \ldots, t_{i-1}, \hat{d})$$


Generate in English:
[S] Amazon [P] founding year [O] 1994
d
Data-to-Text
Extract Triples:
Amazon was founded in 1994.
t̂
Text-to-Data
d'
[S] Amazon [P] founding place [O] 1994
Data-Text-Data Cycle
Extract Triples:
Apple was founded in 1976
t
Text-to-Data
in-freeze
Generate in English:
[S] Apple [P] founding year [O] 1976
d̂
Data-to-Text
in-train
t'
Apple was founded in 1994
Text-Data-Text Cycle

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**
  - All labeled samples
- **Low-resource fine-tuning**
  - 100 labeled samples
- **Low-resource fine-tuning with additional pretraining**
  - 100 labeled samples for fine-tuning and target domain text for pretraining
- **Unsupervised cycle training**
  - Unpaired samples for cycle training
- **Low-resource cycle training**
  - 100 labeled samples for fine-tuning and unpaired samples for cycle training
- **Unsupervised cycle training at different overlapping levels**

Experiments

Dataset	Method	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	BLEU	BERTScore	PARENT
WebNLG	Fully-supervised fine-tuning	**59.99**	**40.93**	**49.32**	**39.76**	**42.83**	**95.41**	**45.67**
	Unsupervised cycle training	58.65	37.70	46.18	37.98	36.36	94.42	43.24
E2E	Fully-supervised fine-tuning	**69.77**	**42.87**	**50.93**	**52.90**	**29.35**	**94.76**	**41.91**
	Unsupervised cycle training	63.43	37.73	45.96	50.49	27.92	93.71	37.97
WTQ	Fully-supervised fine-tuning	**62.25**	**34.59**	**49.41**	**39.17**	**21.18**	**92.88**	**24.18**
	Unsupervised cycle training	61.27	33.45	48.22	39.06	20.46	92.67	23.05
WSQL	Fully-supervised fine-tuning	**58.27**	**32.77**	**48.40**	**37.95**	**22.97**	**93.18**	**24.00**
	Unsupervised cycle training	42.24	15.17	33.52	29.45	4.03	85.37	14.63

Experiments

Dataset	Method	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	BLEU	BERTScore	PARENT
WebNLG	Low-resource fine-tuning	55.55	36.63	**46.21**	35.22	33.63	**94.60**	41.37
	Unsupervised cycle training	**58.65**	**37.70**	46.18	**37.98**	**36.36**	94.42	**43.24**
E2E	Low-resource fine-tuning	**66.62**	**39.68**	**48.59**	48.80	25.31	**94.35**	**39.56**
	Unsupervised cycle training	63.43	37.73	45.96	**50.49**	**27.92**	93.71	37.97
WTQ	Low-resource fine-tuning	55.89	31.60	46.73	31.98	15.34	91.91	**23.36**
	Unsupervised cycle training	**61.27**	**33.45**	**48.22**	**39.06**	**20.46**	**92.67**	23.05
WSQL	Low-resource fine-tuning	**56.37**	**31.60**	**49.42**	**33.57**	**23.34**	**92.57**	**23.68**
	Unsupervised cycle training	42.24	15.17	33.52	29.45	4.03	85.37	14.63

Experiments

Dataset	Method	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	BLEU	BERTScore	PARENT
WebNLG	Low-resource fine-tuning	55.55	36.63	46.21	35.22	33.63	94.60	41.37
	Low-resource cycle training	**60.21**	**40.56**	**48.71**	**39.74**	**41.77**	**95.18**	**46.14**
E2E	Low-resource fine-tuning	66.62	39.68	48.59	48.80	25.31	94.35	39.56
	Low-resource cycle training	**69.53**	**42.48**	**50.51**	**53.02**	**29.22**	**94.74**	**41.39**
WTQ	Low-resource fine-tuning	55.89	31.60	46.73	31.98	15.34	91.91	23.36
	Low-resource cycle training	**61.54**	**34.25**	**49.07**	**39.09**	**20.93**	**92.66**	**24.39**
WSQL	Low-resource fine-tuning	56.37	31.60	49.42	33.57	23.34	92.57	23.68
	Low-resource cycle training	**58.71**	**33.13**	**51.01**	**37.43**	**25.60**	**93.03**	**25.84**

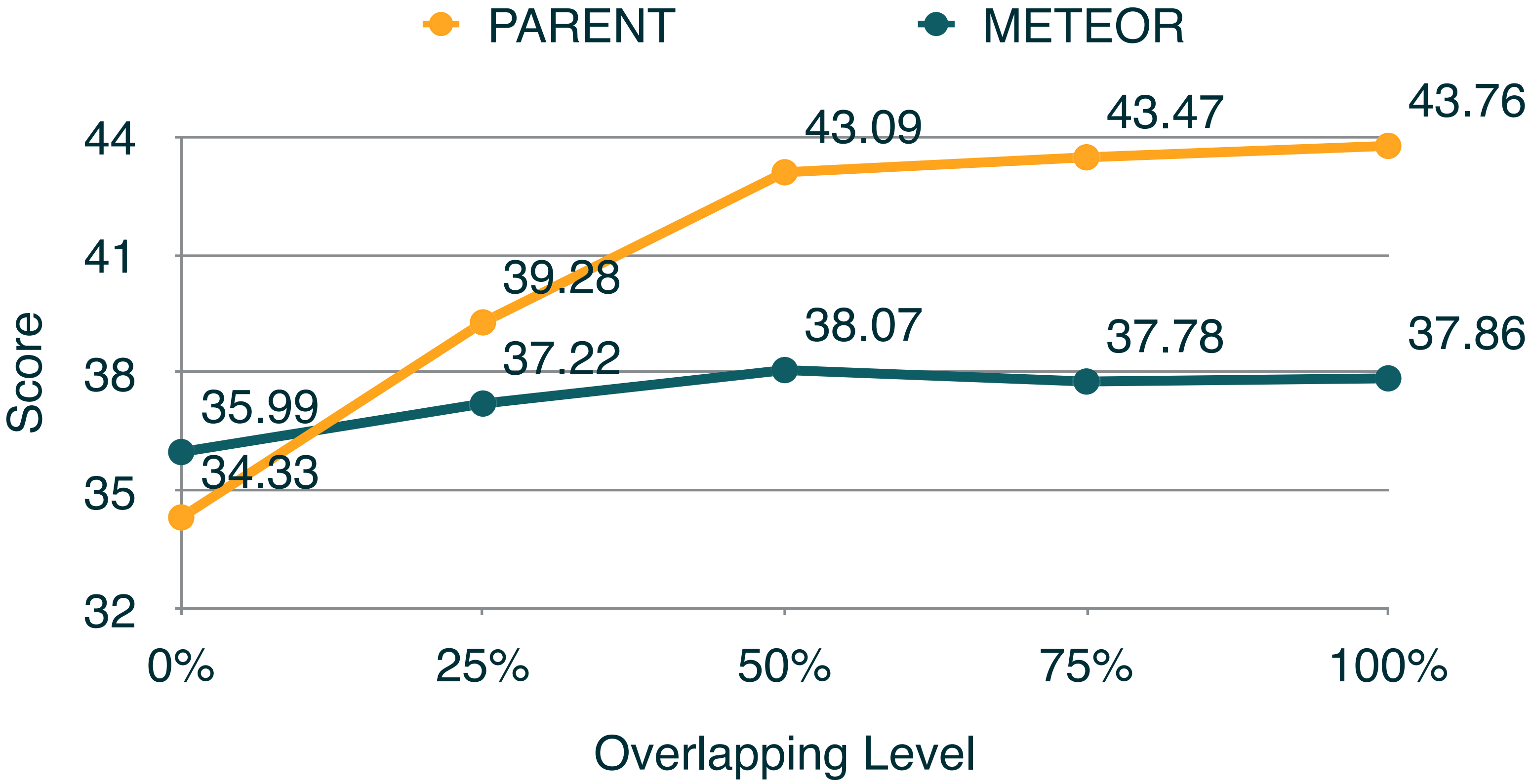
Experiments

Dataset	Method	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	BLEU	BERTScore	PARENT
WebNLG	Fully-supervised fine-tuning	59.99	**40.93**	**49.32**	**39.76**	**42.83**	**95.41**	45.67
	Low-resource cycle training	**60.21**	40.56	48.71	39.74	41.77	95.18	**46.14**
E2E	Fully-supervised fine-tuning	**69.77**	**42.87**	**50.93**	**52.90**	**29.35**	**94.76**	**41.91**
	Low-resource cycle training	69.53	42.48	50.51	53.02	29.22	94.74	41.39
WTQ	Fully-supervised fine-tuning	**62.25**	**34.59**	**49.41**	**39.17**	**21.18**	**92.88**	24.18
	Low-resource cycle training	61.54	34.25	49.07	39.09	20.93	92.66	**24.39**
WSQL	Fully-supervised fine-tuning	58.27	32.77	48.40	**37.95**	22.97	**93.18**	24.00
	Low-resource cycle training	**58.71**	**33.13**	**51.01**	37.43	**25.60**	93.03	**25.84**

Experiments

Dataset	Method	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	BLEU	BERTScore	PARENT
WebNLG	Low-resource FT+additional PT	55.28	35.71	45.41	35.26	33.44	94.33	39.47
	Low-resource cycle training	**60.21**	**40.56**	**48.71**	**39.74**	**41.77**	**95.18**	**46.14**
E2E	Low-resource FT+additional PT	66.88	39.45	48.65	50.11	26.29	94.35	39.65
	Low-resource cycle training	**69.53**	**42.48**	**50.51**	**53.02**	**29.22**	**94.74**	**41.39**
WTQ	Low-resource FT+additional PT	55.57	30.48	44.47	33.73	15.89	91.53	22.88
	Low-resource cycle training	**61.54**	**34.25**	**49.07**	**39.09**	**20.93**	**92.66**	**24.39**
WSQL	Low-resource FT+additional PT	56.01	30.92	47.00	35.34	21.18	92.24	22.66
	Low-resource cycle training	**58.71**	**33.13**	**51.01**	**37.43**	**25.60**	**93.03**	**25.84**

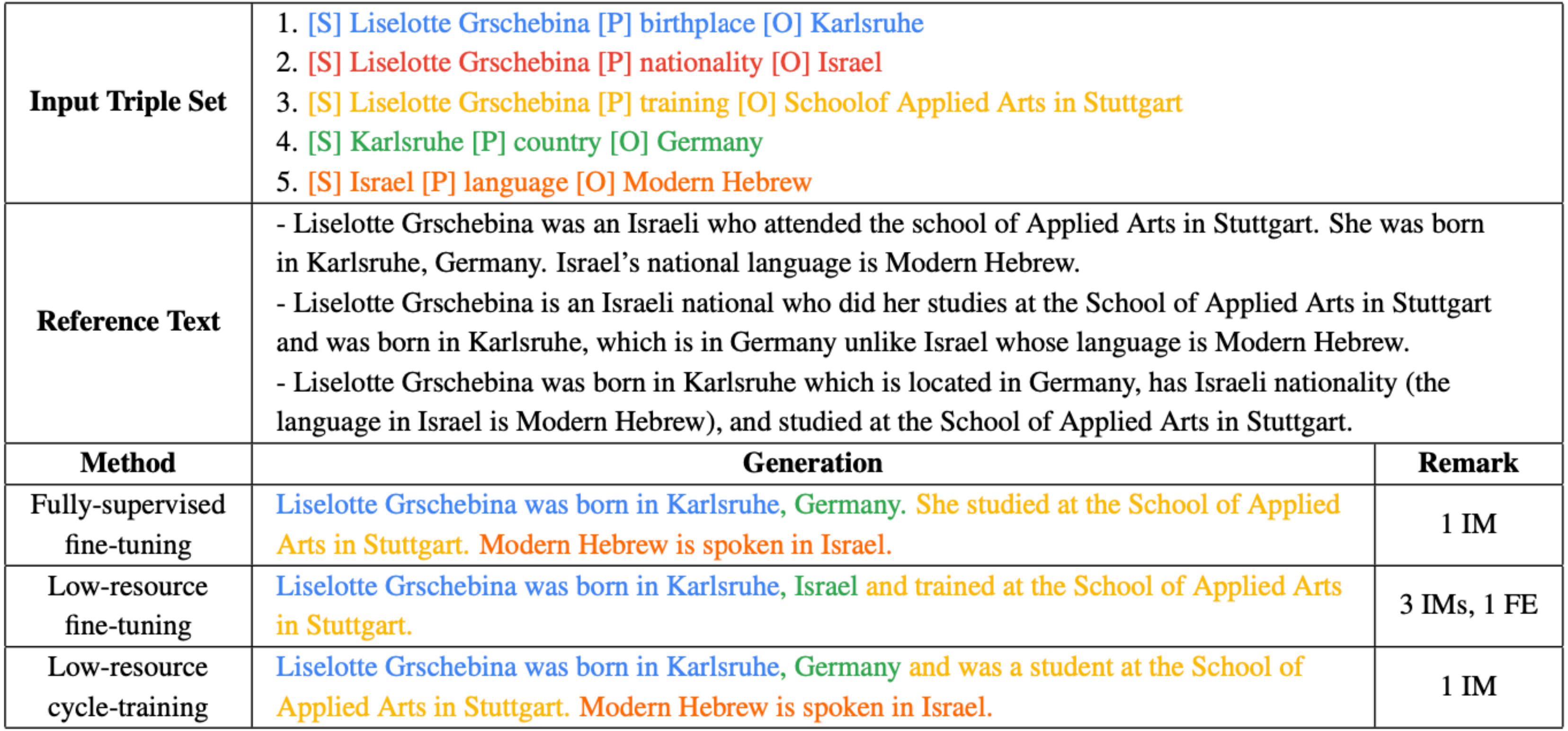
Experiments



Human Evaluation

- **A new annotation schema**
  - Quantitative
  - Better objectiveness, consistency, and precision
- **Three aspects annotated by well trained experts**
  - Count of Factual Errors (FE)
  - Count of Hallucination Errors (HE)
  - Count of Information Misses (IM)
- **One aspect annotated by crowdsource workers**
  - Fluency Preference (FP)

Human Evaluation



Input Triple Set	1. [S] Liselotte Grschebina [P] birthplace [O] Karlsruhe 2. [S] Liselotte Grschebina [P] nationality [O] Israel 3. [S] Liselotte Grschebina [P] training [O] Schoolof Applied Arts in Stuttgart 4. [S] Karlsruhe [P] country [O] Germany 5. [S] Israel [P] language [O] Modern Hebrew	
Reference Text	- Liselotte Grschebina was an Israeli who attended the school of Applied Arts in Stuttgart. She was born in Karlsruhe, Germany. Israel's national language is Modern Hebrew. - Liselotte Grschebina is an Israeli national who did her studies at the School of Applied Arts in Stuttgart and was born in Karlsruhe, which is in Germany unlike Israel whose language is Modern Hebrew. - Liselotte Grschebina was born in Karlsruhe which is located in Germany, has Israeli nationality (the language in Israel is Modern Hebrew), and studied at the School of Applied Arts in Stuttgart.	
Method	**Generation**	**Remark**
Fully-supervised fine-tuning	Liselotte Grschebina was born in Karlsruhe, Germany. She studied at the School of Applied Arts in Stuttgart. Modern Hebrew is spoken in Israel.	1 IM
Low-resource fine-tuning	Liselotte Grschebina was born in Karlsruhe, Israel and trained at the School of Applied Arts in Stuttgart.	3 IMs, 1 FE
Low-resource cycle-training	Liselotte Grschebina was born in Karlsruhe, Germany and was a student at the School of Applied Arts in Stuttgart. Modern Hebrew is spoken in Israel.	1 IM

Human Evaluation

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**

- Showing aggregated results, each dataset's result available in our paper
- Annotation guidelines and interface available in Appendix A of our paper
- Additional generation samples available in Appendix B of our paper

Thank you!

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