Improving Compositional Generalization with Self-Training for Data-to-Text Generation

Sanket Vaibhav Mehta, Jinfeng Rao, Yi Tay, Mihir Kale, Ankur P. Parikh, Emma Strubell

{svmehta, estrubel}@cs.cmu.edu, {jinfeng, yitay, mihirkale, aparikh}@google.com

Summary

- Data-to-text generation focuses on generating fluent natural language responses from structured meaning representations (MRs). Such representations are compositional and it is expensive to collect responses for all possible combinations of atomic meaning schemata, thereby necessitating few-shot generalization to novel MRs.
- In this work, we systematically study the problem of compositional generalization of the state-of-the-art T5 models in few-shot data-to-text tasks. We propose a simple template engine along with a generic BLEURT based self-training approach for improving the model's generalization capabilities.
- On the commonly used Weather and SGD benchmarks, our approach improves tree accuracy by 46%+ and reduces the slot error rate by 73%+ over the strong T5 baselines in few-shot settings.

Semantic Representation

- Tree-structured MR
  - Discourse relations - DS_JUSTIFY,
  - Dialog acts - DG_INFORM,
  - Arguments - LOCATION
- Linearize tree-structured input and target response

Query: Is it jacket weather?

```
[DG_NO
]
[DG_INFORM
   [CONDITION light rain ]
   [HUMIDITY extremely humid ]
   [DATE_TIME [COLLOQUIAL today ] ]
   [LOCATION [CITY Palo Alto ]  ]
]
[DS_JUSTIFY
   [DG_RECOMMEND
      [ATTIRE_NOT jacket ]
      [LOCATION [CITY Palo Alto ] ]
      [DATE_TIME [COLLOQUIAL today ] ]
   ]
   [DG_INFORM
      [CONDITION_NOT cold ]
      [LOCATION [CITY Palo Alto ] ]
      [DATE_TIME [COLLOQUIAL today ] ]
   ]
]
```

Naive Structured Input

```
[DG_NO No
] ,
[DS_JUSTIFY
   [DG_RECOMMEND leave the
      [ATTIRE_NOT jacket ] at home
   ] because
   [DG_INFORM it isn't
      [CONDITION_NOT cold ]
      [DATE_TIME [COLLOQUIAL today ] ]
      [LOCATION in [CITY Palo Alto ] ]
   ] .
]
[DG_INFORM It'll be
   [HUMIDITY extremely humid ] with
   [CONDITION light rain ]
] .
```

Response: No, leave the jacket at home because it isn't cold today in Palo Alto. It'll be extremely humid with light rain.

Structured Target Response

- T5 model - pre-train and fine-tuning discrepancy
- Template engine - recursively traverses the tree-structured MR in a top-down manner to generate structure-aware text representation (template guided input representation)

Query: Is it jacket weather?

```
[DG_NO No
]
[DG_INFORM there will be
   [CONDITION light rain ]
   [HUMIDITY extremely humid ]
   [DATE_TIME at [COLLOQUIAL today ] ]
   [LOCATION in [CITY Palo Alto ]  ]
]
[DS_JUSTIFY
   [DG_RECOMMEND
      [ATTIRE_NOT jacket ] is not recommended
      [LOCATION in [CITY Palo Alto ] ]
      [DATE_TIME at [COLLOQUIAL today ] ]
   ] , because
   [DG_INFORM there won't be
      [CONDITION_NOT cold ]
      [LOCATION in [CITY Palo Alto ] ]
      [DATE_TIME at [COLLOQUIAL today ] ]
   ]
]
```

Template Guided Structured Input

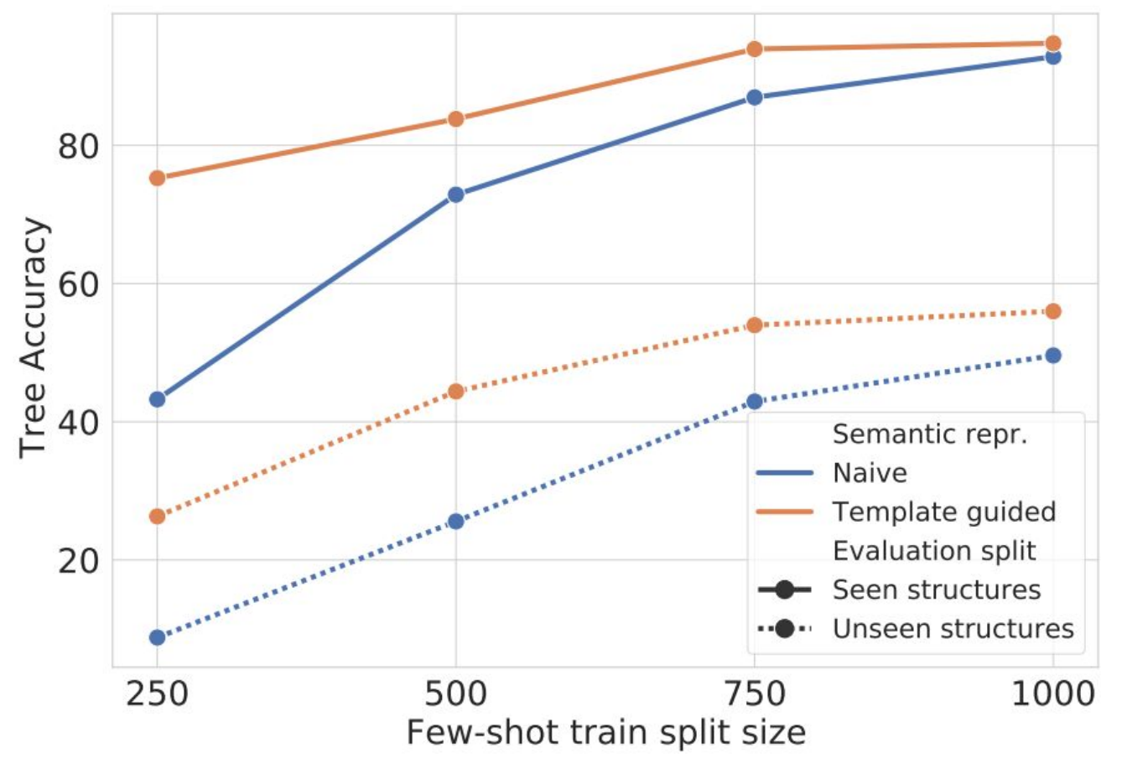
- Example templates

ID	Template Name	Template Body
1	DG_NO	[DG_NO No]
2	DS_JUSTIFY	[DS_JUSTIFY DG_RECOMMEND, because DG_INFORM]
3	DG_INFORM	*IsSet*($condition) ? DG_INFORM_CONDITION : DG_INFORM_CONDITION_NOT
4	DG_INFORM_CONDITION	[DG_INFORM there will be [CONDITION $condition] *Optional*([HUMIDITY $humidity]) DATETIME_AND_LOCATION]
5	DG_INFORM_CONDITION_NOT	[DG_INFORM there won't be [CONDITION $condition] DATETIME_AND_LOCATION]
6	DATETIME_AND_LOCATION	*Optional*(at [DATE_TIME $date_time]) *Optional*(in [LOCATION $location])
7	DG_RECOMMEND	[DG_Recommend [ATTIRE_NOT $attire] is not recommended DATETIME_AND_LOCATION]

Case Study

(Q1) Do current state-of-the-art generation models compositionally generalize?

- Current state-of-the-art generation models (T5-small), see a significant drop in performance on unseen tree-structures
- Naive: 47%-80%, across different few-shot train splits



(Q2) What is an effective semantic representation for tackling compositional generalization?

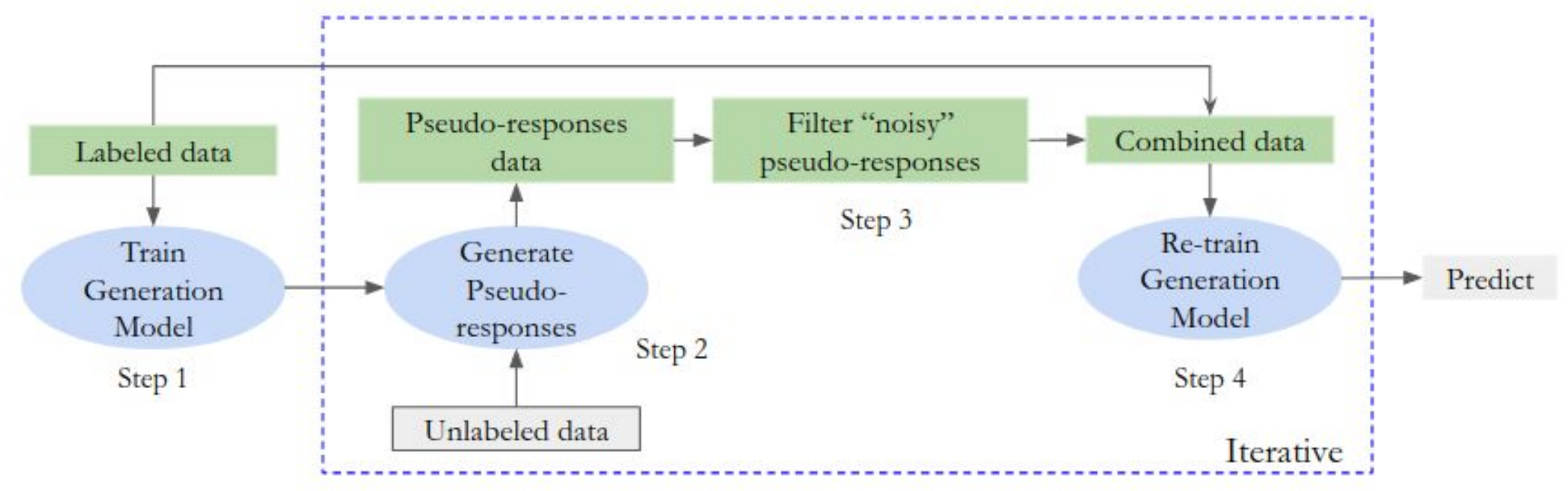
- Template guided: 41%-65%, across different few-shot train splits

(Q3) Does scaling model size (and training data) trivially solve the problem?

- Increasing model size does not close the generalization gap
- T5-small performs similarly or better than its larger counterparts

Model Size	Val. Seen	Val. Unseen
T5-small (77M)	99.54	64.02
T5-base (120M)	99.63	55.80
T5-large (800M)	99.36	58.45

Self-Training using BLEURT



- Self-training is susceptible to "noisy" pseudo-response
- Generation models are prone to hallucinate additional content not supported by the input
- **Solution:** we repurpose BLEURT as a quality estimator to filter "noisy" pseudo-responses during self-training

- Fine-tuning BLEURT

Soruce (text-to-text input): there will be light freezing fog with a temperature high of 74 low of 61 at next friday

Positive candidate (target response): next friday will have a high of 74 , a low of 61 , and a light freezing fog

Negative candidates:

[retrieving similar examples] next friday will be cloudy with a high of 74 , a low of 61 , and thunderstorms and rain

[pairing with reference] there will be light freezing fog with a temperature high of 74 low of 61 at next friday

[swapping words] next friday ~~will~~ of have a high ~~of~~ will 74 , a low of 61 , and a light freezing fog

[repeating phrases] next friday will have a high of 74 , a low of 61 of 61 , and a light freezing fog

[dropping phrases] next friday will have a high of 74 , ~~a low of 61 ,~~ and a light freezing fog

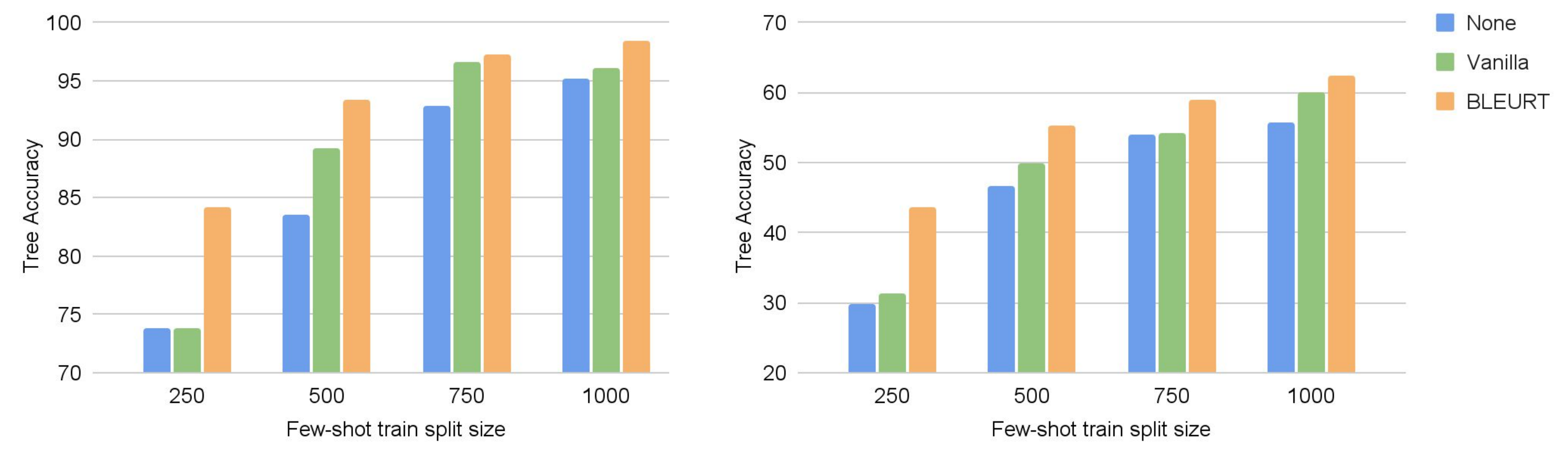
[flipping digits] next friday will have a high of 78 , a low of 61 , and a light freezing fog

Experiments

- Data
  - FewShotWeather: 1shot-250, 1shot-500, 1shot-750, 1shot-1000
  - FewShotSGD: 5-shot, 10-shot, 20-shot, 40-shot
- Metrics - Tree Accuracy, BLEU (↑ is better), Slot Error Rate (↓ is better)
- Model - Seq2Seq: T5.1.1.1 small, BLEURT-20-D12
- Inference - Beam width 4
- Compositional generalization (FewShotWeather - Tree structures)

Seen structures

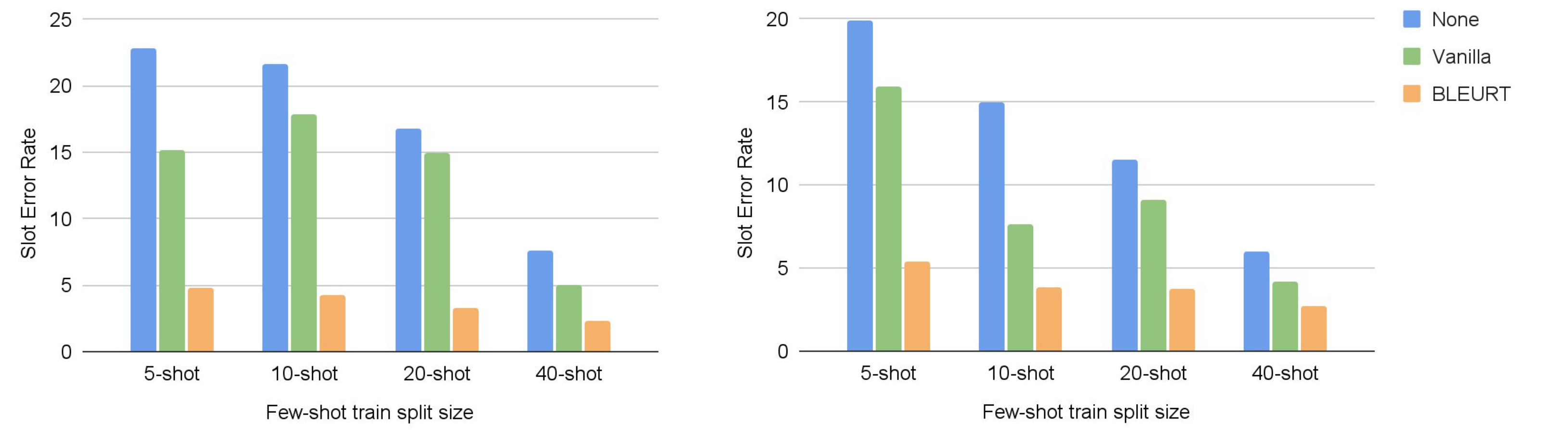
Unseen structures



- Few-shot generalization (FewShotSGD - Flat structures)

Seen schema

Unseen schema



- Performance w.r.t self-training iterations & quality of BLEURT model

Model	Self-training iteration	No. of training examples	FewShotWeather			
			Seen structures		Unseen structures	
			BLEU ↑	Tree Acc. ↑	BLEU ↑	Tree Acc. ↑
Baseline	-	250	69.16	73.68	50.40	29.83
Vanilla	1	+ 14, 742	69.25	73.77	51.87	31.37
	2	+ 4, 170	68.72	73.06	51.92	31.11
BLEURT-250	1	+ 14, 742	69.64	83.85	52.10	41.03
	2	+ 4, 170	69.59	84.12	52.34	43.68
BLEURT-1000	1	+ 14, 021	70.95	84.83	52.13	45.47
	2	+ 4, 772	**70.47**	**85.64**	**53.08**	**47.44**

- Model performance improves across the self-training iterations (2-3 iterations might be sufficient)
- Self-training is sensitive to the quality of the BLEURT model (BLEURT-X denotes BLEURT model fine-tuned using 1-shot-X train split)

- Qualitative analysis (human evaluation study)

Fields	BLEURT	Gram	Nat	Info	Acc	Input or output response
User query	-	-	-	-	-	On the 12th of this month would be great.
Template	-	-	-	-	-	Would you like to fly with American Airlines? The onward flight takes off at 4 am. It has a layover. The returning flight takes off at 12:45 pm. The ticket costs $552
Reference	-	-	-	-	-	How about a connecting American Airlines flight taking off at 4 am and costing $552? The return time is at 12:45 pm.
Predictions						
Baseline	-0.004	2.50	2.17	0.83	0.0	Okay. I've found an American Airlines flight departing at 4 am and returning at 12:45 pm. I inform you that the flight has a return flight at 12:45 pm. The ticket is $1052.
Self-training	0.996	3.00	2.83	0.67	1.0	I've found an American Airlines flight departing at 4 am and returning at 12:45 pm. This will cost you $552.
Full	0.998	2.00	2.00	0.50	1.0	There is an American Airlines flight that leaves at 4 am and has a layover and a return flight at 12:45 pm for $552.

References

Balakrishnan, Anusha, Jinfeng Rao, Kartikeya Upasani, Michael White, and Rajen Subba. *"Constrained Decoding for Neural NLG from Compositional Representations in Task-Oriented Dialogue."* ACL, 2019.

Kale, Mihir, and Abhinav Rastogi. *"Template Guided Text Generation for Task-Oriented Dialogue."* EMNLP, 2020

Sellam, Thibault, Dipanjan Das, and Ankur Parikh. *"BLEURT: Learning Robust Metrics for Text Generation."* ACL, 2020.

Code and data: github.com/google-research/google-research/tree/master/compgen_d2t