| Eval on Daily Dialogue(dd) datasets | BLEU2 | BLEU4 |
| --- | --- | --- |
| T5-Base (unfinetuned)\* | 4.56 | 1.14 |
| T0-3B (unfinetuned) | 5.27 | 1.31 |
| Prompt T5-Base(220M) | 1.35 | 0.31 |
| Prompt T0-3B | 5.57 | 1.49 |
| T5-Base + LLMR | 0.076 | 4.1e-154 |
| T5-Base + CE | 6.19 | 1.69 |
| T5-Base + CE + LLMR (ours) | 6.46 (+0.27) | 1.72 (+0.03) |
| T5-Base + CE + LLMR (ours)(with constant scheduler) | 6.61 (+0.42) | 1.67 (-0.02) |

**\*T5-Base+CE = T5-base ce initialized with t0-3b prompt generated pseudo labels**

\*We include the performance of T5-Base (unfinetuned), as it is the starting point of our distillation. This may not be a practical model but T5-Base cannot accomplish dialogue generation without fine-tuning

1.Smoothing the language model prediction?(absorbed in lr?)

2.Nucleus sampling(to truncate)

**October 27, 2022 by Dongheng Li**

**Experimental purpose**: Initialize T5\_base with ost using CE

**Setup:T5\_base\_CE, dataset=ost, same setting as RebTeg**

Ongoing process: 71600/100000

| Eval on OST | BLEU2 | BLEU4 |
| --- | --- | --- |
| model-004k(best so far) | 3.73 | 1.21 |

.

**October 27, 2022 by Dongheng Li**

**Experimental purpose**: disabled the drop out using args.deterministic

**Setup:LLMR\_2, dataset=dd, induced reward KD from t0p**

if self.args.deterministic:

self.model.eval()

Where model.eval() disabled the dropout function

model-001k

| Eval on Daily Dialogue(dd) datasets | BLEU2 | BLEU4 |
| --- | --- | --- |
| T5-Base + LLMR(determ) | 6.36 | 1.69 |

**Conclusion:**

Disabling the dropout function does not help

**Observation:**

The performance steadily decrease in first 10 ckpt.However, the best performance was at 21th ckpt achieved by former stable experiments, more observations required.

**October 25, 2022 by Dongheng Li**

**Experimental purpose**: mask out the extra\_ids using args.mask\_extra

if args.mask\_extra:

logits[..., 32000:] = -torch.inf (where logits[..., 32000:] means after 32000 vocabulary, all extra\_ids are assigned to -inf.

**Setup:T5-Base+RL, dataset=dd, induced reward KD from t0p**

**48744748**

| Eval on Daily Dialogue(dd) datasets | BLEU2 | BLEU4 |
| --- | --- | --- |
| T5-Base + LLMR | 4.27e-15 | 6.24-92 |

**Conclusion:**

Generate <extra\_id> is not the problem causing the low performance (short length)

**Observation:**

There is always a generation of an empty token in the beginning of the sentence.

**October 24, 2022 by Dongheng Li**

**Experimental purpose**: disabled the drop out using args.deterministic

**Setup:LLMR\_2, dataset=dd, induced reward KD from t0p**

if self.args.deterministic:

self.model.eval()

Where model.eval() disabled the dropout function

**Scheduled process**

**48751836**

**October 24, 2022 by Dongheng Li**

**Experimental purpose**: mask out the extra\_ids using args.mask\_extra

if args.mask\_extra:

logits[..., 32000:] = -torch.inf (where logits[..., 32000:] means after 32000 vocabulary, all extra\_ids are assigned to -inf.

**Setup:T5-Base+RL, dataset=dd, induced reward KD from t0p**

**Scheduled process**

**48744748**

**October 20, 2022 by Dongheng Li**

**Experimental purpose**: Improve the T5-Base + CE + LLMR (ours)

**Setup:T5-Base+CE+RL, dataset=dd, induced reward KD from t0p**

**--scheduler constant**

**-lr 1e-6**

**model-021k**

| Eval on Daily Dialogue(dd) datasets | BLEU2 | BLEU4 |
| --- | --- | --- |
| T5-Base + CE | 6.19 | 1.69 |
| T5-Base + CE + LLMR (ours) | 6.46 (+0.27) | 1.72 (+0.03) |
| T5-Base + CE + LLMR (ours)(with constant scheduler) | 6.61 (+0.42) | 1.67 (-0.02) |

**October 15, 2022 by Dongheng Li**

**Experimental purpose**: Improve the T5-Base + CE + LLMR (ours)

**Setup:T5-Base+CE, dataset=dd, induced reward KD from t0p**

**--scheduler constant**

**-lr 1e-6**

**Scheduled process**

**48100514**

**October 13, 2022 by Dongheng Li**

**Experimental purpose**: Improve the T5-Base + CE + LLMR (ours)

**Setup:T5-Base+CE, dataset=dd, induced reward KD from t0p**

**--scheduler constant**

**-lr 1e-5**

**Results:**

**model-001k from validation models**

|  | BLEU2 | BLEU4 |
| --- | --- | --- |
| T5-Base + CE + LLMR (ours) | 6.46 | 1.72 |
| LLMR\_10 | 6.64e-157 | 9.98e-233 |
| T5-dd + LLMR | 0.028 | 0.0048 |

**Conclusion:**

1. Constant scheduler may require a smaller lr?

**Todo:**

1. **Try lr 1e-6**

**October 9, 2022 by Dongheng Li**

**Experimental purpose**: baseline T5-dd + LLMR

**Setup: T5-dd, dataset=dd, induced reward KD from t0p**

**Results:**

**model-001k from validation models**

|  | BLEU2 | BLEU4 |
| --- | --- | --- |
| Prompt T0-3B  [test] | 5.56 | 1.49 |
| T5-dd + LLMR | 0.028 | 0.0048 |

**Oct 4:**

**Observation:** Student is better teacher

* Consistent with previous learning from search results [read it, NeurIPS20, ACL’22]
* May find more specific explanation.
  + Hypothesis: sometimes prompt’s output doesn’t make sense at all, like continue generating the instruction rather than provide the answer

<https://proceedings.neurips.cc/paper/2020/file/7a677bb4477ae2dd371add568dd19e23-Paper.pdf>

<https://aclanthology.org/P19-1338.pdf>

<https://aclanthology.org/2021.findings-emnlp.307.pdf>

<https://aclanthology.org/2022.acl-long.545.pdf>

**October 2, 2022 by Dongheng Li**

**Experimental purpose**: baseline T5-Base + CE + LLMR (ours)

ENTROPY tuning

**Setup: T5+pseudo label, dataset= dd\_t0p, induced reward KD**

**default:5000**

T5-Base + CE + LLMR\_5= --update-per-sync 1000

T5-Base + CE + LLMR\_6 = --update-per-sync 10000

**Results:**

**model-001k from validation models**

|  | BLEU2 | BLEU4 |
| --- | --- | --- |
| Prompt T0-3B  [test] | 5.56 | 1.49 |
| T5-Base + CE  [test] | 6.19 | 1.69 |
| T5-Base + CE + LLMR\_1  [test]  (crashed in training) | 6.45 | 1.72 |
| T5-Base + CE + LLMR\_2  [test]  (crashed in training) | 6.46 | 1.72 |
| T5-Base + CE + LLMR\_3  [test]  (crashed in training) | 6.27 | 1.68 |
| T5-Base + CE + LLMR\_4  [test]  (stable in training) | 6.26 | 1.68 |
| T5-Base + CE + LLMR\_4.25  [test]  (crashed in training) | 6.46 | 1.72 |
| T5-Base + CE + LLMR\_4.5  [test]  (crashed in training) | 6.45 | 1.72 |
| T5-Base + CE + LLMR\_5  [test]  (stable in training) | 6.28 | 1.68 |
| T5-Base + CE + LLMR\_6  [test]  (stable in training) | 6.27 | 1.67 |

**Conclusion:**

1. Behavior policy updates seems not so influential from this experiment.
2. From all experiments, when training is stable the performance is lower, while the best performance seemed to be achieved at 0.5 exploration rate in the first 1000 updates and then crash.Even for those stable experiments, the performance is decreasing…

**Todo:**

1. How to solve the problem 1.stability vs performance 2.Decresing of the performance
2. Try 48h training for at least 8 ckpts in entropy0.1 topk5 --update-per-sync 1000 since in default --update-per-sync is set to 5000 and 24h is not enough time to get the first ckpt that has updated behavior policy.

**September 31, 2022 by Dongheng Li**

**Experimental purpose**: baseline T5-Base + CE + LLMR (ours)

ENTROPY tuning

**Setup: T5+pseudo label, dataset= dd\_t0p, induced reward KD**

T5-Base + CE + LLMR\_4 = entropy 0.1

T5-Base + CE + LLMR\_4.25 = entropy 0.25

T5-Base + CE + LLMR\_4.5 = entropy 0.5

**Results:**

**model-001k from validation models**

|  | BLEU2 | BLEU4 |
| --- | --- | --- |
| Prompt T0-3B  [test] | 5.56 | 1.49 |
| T5-Base + CE  [test] | 6.19 | 1.69 |
| T5-Base + CE + LLMR\_1  [test]  (crashed in training) | 6.45 | 1.72 |
| T5-Base + CE + LLMR\_2  [test]  (crashed in training) | 6.46 | 1.72 |
| T5-Base + CE + LLMR\_3  [test]  (crashed in training) | 6.27 | 1.68 |
| T5-Base + CE + LLMR\_4  [test]  (stable in training) | 6.26 | 1.68 |
| T5-Base + CE + LLMR\_4.25  [test]  (crashed in training) | 6.46 | 1.72 |
| T5-Base + CE + LLMR\_4.5  [test]  (crashed in training) | 6.45 | 1.72 |

**Conclusion:**

1. Training stables at entropy = 0.1, which means the smaller exploration rate will stable the training, however get a lower performance.
2. Topk will not influence the results much and will choose topk = 5 and entropy = 0.1 for following experiments

**Todo:**

1. To explore more possibilities, we decided to try to change the number of updates for updating behavior policy in RL training.
2. Tune with --update-per-sync

**September 29, 2022 by Dongheng Li**

**Experimental purpose**: baseline T5-Base + CE + LLMR (ours)

TOPK tuning

**Setup: T5+pseudo label, dataset= dd\_t0p, induced reward KD**

T5-Base + CE + LLMR\_1 = topk5

T5-Base + CE + LLMR\_2 = topk2

T5-Base + CE + LLMR\_3 = topk10

**Results:**

**model-001k from validation models**

|  | BLEU2 | BLEU4 |
| --- | --- | --- |
| Prompt T0-3B  [test] | 5.56 | 1.49 |
| T5-Base + CE  [test] | 6.19 | 1.69 |
| T5-Base + CE + LLMR\_1  [test]  (crashed in training) | 6.45 | 1.72 |
| T5-Base + CE + LLMR\_2  [test]  (crashed in training) | 6.46 | 1.72 |
| T5-Base + CE + LLMR\_3  [test]  (crashed in training) | 6.27 | 1.68 |

**Conclusion:**

1. Training unstable, crushes from second save(2000 update),and generate strange long text.
2. Topk will not influence the results much and will choose topk = 5 for following experiments

**Todo:**

1. Arg.entropy controls the exploration and exploitation of the RL training, is it possible that 0.5(default) too high,that it explorate too much
2. Tune with entropy value.

**September 27, 2022 by Dongheng Li**

**Scheduled** baseline t5cet0p + LLMR (ours)

**September 25, 2022 by Dongheng Li**

**Experimental purpose**: CE training T5-base model with pseudo label from t0\_3b + prompt

**Setup: T5+pseudo label, dataset= dd\_t0p**

**Results:**

**model-004k is the ßbest validation models**

|  | BLEU2 | BLEU4 |
| --- | --- | --- |
| Prompt T0-3B | 5.56 | 1.49 |
| T5-Base + CE  [valid] | 6.30 | 1.71 |
| T5-Base + CE  [test] | 6.19 | 1.69 |

**Todo:**

1. Do baseline T5-Base + CE + LLMR (ours)
2. Do baseline Prompt T5-Base(100M)

**September 20, 2022 by Dongheng Li**

**Experimental purpose**: CE training T5-base model with pseudo label from t0\_3b + prompt

**Setup: T5+pseudo label, dataset= dd\_t0p**

* **~~Ongoing process: 100000/100000~~**

Estimate time: ~13h

Training loss: 3.341 to 1.386 (at the end of 35800 updates)

Grad\_norm: 1.881 to 0.01978 (at the end of 35800 updates)

Estimate completion time**:**9/21/2022

**Todo:**

1. Monitor the training loss
2. Valid and test

**September 11, 2022 by Dongheng Li**

**Experimental purpose**: Generate the pseudo ground truth for T5 initiation

**Setup: T0 + prompt, dataset= dd\_ost\_singleturn**

* **~~Ongoing process: 60005/60005~~**

Total sentence to process: 979230

Avg. processing speed: 5it/s ~ 6it/s

Estimate time: ~49h

Estimate completion time**:**9/15/2022

**Todo:**

1. Observed unwanted quotation marks in generation, preprocess needed for performing basic KD in the next step

1. Basic KD withpseudo ground truth
2. T5-base zero shot generation

**September 9, 2022 by Dongheng Li**

**Experimental purpose**: Try if which prompt work the best with T0-3b

**Setup: T0 + prompt, dataset= dd\_singleturn**

**Prompt 10: dialog response of "[SRC]" is:**

Prompt 1: Generate the dialog response of "[SRC]"

Prompt 2: reply to "[SRC]":

Prompt 3: Answer to "[SRC]" is:

Prompt 4: Generate the response of "[SRC]"

Prompt 5: "[SRC]". Said the man

More in [LLMR templates for t0-3b](https://docs.google.com/document/d/1QHAYxx12deKdodjyRAyxyladsjBa3FMQt7ym-i_9YYA/edit)

|  | BLEU2 | BLEU4 |
| --- | --- | --- |
| **BEST (prompt 10)** | 5.56 | 1.49 |
| Prompt 1 | 5.11 | 1.27 |
| Prompt 2 | 4.73 | 1.24 |
| Prompt 3 | 4.51 | 1.16 |
| Prompt 4 | 4.86 | 1.26 |
| Prompt 5 | 3.25 | 1.01 |
| Prompt 6 | 4.66 | 1.23 |
| Prompt 7 | 4.21 | 1.18 |
| Prompt 8 | 4.75 | 1.25 |
| Prompt 9 | 3.77 | 1.12 |
| Prompt 10 | 5.56 | 1.49 |

**Conclusion:**

**1.By observation, prompt generation is more accurate when the template is more instruction-like.**

**2.The special operator([,({}) seemed not to be as influential as imagined(suggested by guoqing)**

**TODO:**

1. Cross-entropy training with generated sequences as pseudo-groundtruth (hard): <https://arxiv.org/abs/1606.07947>
2. Cross-entropy training against predicted probabilities (soft)
3. A list of papers about KD ([KD list](https://docs.google.com/document/d/17s3U2cWtjzbW_uiPb4yhxjiE-MUR63uPls7j0QD6dyw/edit))

**September 1, 2022 by Dongheng Li**

**Experimental purpose**: Try if which prompt work the best with T0-3b

**Setup: T0, dataset=?,**

Prompt 1:

Prompt 2:

|  | BLEU2 | BLEU4 | Dist-1 | Dist-2 |
| --- | --- | --- | --- | --- |
| BEST (prompt 1) | 5.11 | 1.27 |  |  |
| Prompt 1 | 5.11 | 1.27 |  |  |
| Prompt 2 |  |  |  |  |
| … |  |  |  |  |

TODO:

1. Cross-entropy training with generated sequences as pseudo-groundtruth (hard): <https://arxiv.org/abs/1606.07947>
2. Cross-entropy training against predicted probabilities (soft)
3. A list of papers about KD ([KD list](https://docs.google.com/document/d/17s3U2cWtjzbW_uiPb4yhxjiE-MUR63uPls7j0QD6dyw/edit))