Master thesis

University of Tartu

April 1, 2023

1 Meeting notes 7

1. We could train baselines with their [Neeman et al., 2022] codebase, and also point out that we didn't get the same results as in the paper even though we used their codebase on their data. Maybe due to random seed we didn't get same resutls.

2. Yova's colleague from university of Edinburgh is one of the authors of [Welbl et al., 2020]. So in case we have any questions we could ask him.

3. Wait until Ella replies and ask her about different results that we are getting with their codebase and the data. What might be wrong? Maybe they also could provide information about their variance?

4. Our work is not about going after state-of-the-art results but rather to do comprehensive study with our methods.

   Note: However, it is a little bit inconsistent with what Yova said last time: "We need to get better results for counterfactuals data".

5. Describe in my write-up all the methods we are employing. So that Yova could be on the same page with me.

6. Train with my new updated codebase(related to matching training settings with [Neeman et al., 2022]) with factual dataset.

7. Try data-augmentation training(old name was adversarial-training).

8. Kairit suggested to do bootstrapping test. Since [Neeman et al., 2022] didn't provide variance for their results, and we are getting different results. This might raise questions. Thus we could run several training's, and take an average but it would take too much time, considering that

one run take one day. Thus we might try to ask for variance from [Neeman et al., 2022], or do bootstrapping test.

9. Results from hyperparameters search for prompt-tuning isn't promising. I need to try another library for that.

Todos:

1. Write article about Prompt-Tuning, and Data-augmentation training.
2. Train on factual dataset [Neeman et al., 2022] codebase.
3. Instructive in-context-learning for T5-FLAN-XXL(no weights updates)
4. Train baselines on all the datasets(f,f+cf,f+a,f+cf+a).
5. Try data-augmentation training
6. Finish all the hyperparameters search
7. Ping Ella to ask for the variance, and other follow ups.
8. Try PEFT library for prompt-tuning, since my implementation doesn't give promissing results.
9. Perform bootstrapping test
10. Remove non-best checkpoints to free up space
11. Prepare slides for my next presentation
12. Read HPC documentation, to understand allocation of two machines.

Desirable outcome:

1. Finish write-up for Prompt-tuning, and data-augmentation training.
2. Try our new library PEFT for prefix and prompt tuning.
3. Train on factual dataset [Neeman et al., 2022] codebase.
4. Finish hyperparameter search
5. Remove non-best checkpoints
6. Try out data-augmentation training
7. Perform bootstrap test to get variance
8. Train baseline with factual dataset
9. Improved presentation of my experiments

References

[Neeman et al., 2022] Neeman, E., Aharoni, R., Honovich, O., Choshen, L., Szpektor, I., and Abend, O. (2022). Disentqa: Disentangling parametric and contextual knowledge with counterfactual question answering.

[Welbl et al., 2020] Welbl, J., Minervini, P., Bartolo, M., Stenetorp, P., and Riedel, S. (2020). Undersensitivity in neural reading comprehension.