Developing a Metric for Measuring Hallucinations in Summary Text

Milestone 1

15-300, Fall 2021

Abby Shrack

With advising from Lori Levin, Language Technologies Institute

https://abbyshrack.github.io/Summarization/

November 12, 2021

1 What you have accomplished so far

I have mostly completed my background research into the existing set of summarization models, metrics, and meta-evaluation techniques. I am in the process of identifying which models and metrics I want to recreate for my testing and improvement purposes. In particular, I would like to select one or two models to recreate and one or two metrics to recreate. These will most likely be state-of-the-art models (e.g. general language model pretraining [1]) and metrics (e.g. FEQA[2] or BARTScore [3]). My recreation of metrics will necessarily be more important as I work towards creating a more effective metric of my own. Therefore, though I hope to select only one or two metrics as a baseline, I will necessarily recreate more as my research progresses. These may include NLI entailment models [4] and rule-based perturbations [5][6], as described in my proposal. In my selection, I may prioritize models and metrics that have existing code available online alongside prioritizing effectiveness.

2 Major Changes

I have made no major changes to my project plan.

3 Meeting your milestone

I have not yet met the milestone exactly as described in my project proposal. However, as I expect to gain more information as I progress with my research, I have decided to write the related work section closer to the end of 07-400. This will allow me to incorporate changes that I make as I go, rather than having to rewrite the related work if things change and as I gain more information and specificity in my direction.

4 Surprises

I have not run into any major surprises with the content of my project. However, I have been busier than expected with other end-of-semester tasks, and as such have been unable to get a head start on choosing models and metrics to recreate. I hope to be able to start this over winter break instead.

5 Revisions to your 07-400 milestones

The only revision I have made is that for this milestone I have decided not to fully write my related work section, as explained in the previous section. I still intend to be able to meet the rest of my milestones, as the extra work needed to write the related work section should be relatively minimal and accomplishable either as I go or nearer to the end stages of my project.

6 Resources needed

At this stage I do not expect to need any additional resources. However, I have not yet identified a source for external GPU resources if pretraining of models or metrics needs to be done, as mentioned in my project proposal.

References

[1] Du, Z., Qian, Y., Liu, X., Ding, M., Qiu, J., Yang, Z. and Tang, J., 2021. All NLP Tasks Are Generation Tasks: A General Pretraining Framework. *arXiv preprint arXiv:2103.10360*.

[2] Durmus, E., He, H. and Diab, M., 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. *arXiv preprint arXiv:2005.03754*.

[3] Yuan, W., Neubig, G. and Liu, P., 2021. BARTScore: Evaluating Generated Text as Text Generation. *arXiv preprint arXiv:2106.11520*.

[4] Falke, T., Ribeiro, L.F., Utama, P.A., Dagan, I. and Gurevych, I., 2019, July. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 2214-2220).

[5] Kryściński, W., McCann, B., Xiong, C. and Socher, R., 2019. Evaluating the factual consistency of abstractive text summarization. *arXiv preprint arXiv:1910.12840*.

[6] Liu, T., Zhang, Y., Brockett, C., Mao, Y., Sui, Z., Chen, W. and Dolan, B., 2021. A Token-level Reference-free Hallucination Detection Benchmark for Free-form Text Generation. *arXiv preprint arXiv:2104.08704*.