Developing a Metric for Measuring Hallucinations in Summary Text

15-300, Fall 2021

Abby Shrack

https://abbyshrack.github.io/summarization/

November 12, 2021

1 Project Description

I will be working with Professor Lori Levin in the Language Technologies Institute, with advising from Pengfei Liu and Maria Ryskina, postdoctoral researcher and PhD student in the Language Technologies Institute, to develop a metric for measuring hallucinated text in outputs of summarization models.

Abstractive summarization generates a short summary of a given source document by paraphrasing the source text rather than only selecting subsections. This typically relies on the NLP field of text generation - generating relevant text from a general vocabulary. However, text generation is imperfect, and the outputs of abstractive text summarization frequently suffer from so-called “hallucinations”, or information that is not present in the source document. This incorrect information often renders such summaries useless in practice, so it is imperative to reduce the erroneous test. One important part of this is being able to recognize incorrectly generated text and measure the accuracy of summaries generated by different models. To address this issue, we aim to develop a metric that can automatically evaluate the correctness of summarization outputs given the source document.

We propose several potential directions for this metric. The first of these deals with question answering. Questions can be generated from the summary by masking spans of the text and generating questions based on those masks. These questions are then passed with the source document into a question answering system, and if the answer differs from the masked text, the summary is marked as inconsistent [1]. We will consider modifications to this method using data augmentation, since current question answering systems can only generate a small set of questions from a given summary. The second method we consider modifies a recent development in using the probability that a text would be generated by a certain model as an evaluation of its correctness. This has been done previously using the pre-trained model BART [9], but we consider using additional models such as T5 and PEGASUS, as well as fine-tuning parameters to work more specifically for correctness of generated summaries. Finally, we will also consider alternative methods including the use of natural language inference and entailment, embedding spaces, and rule-based classifiers.

If we are successful, this metric will provide a simpler, more efficient method of measuring correctness of summarization outputs than having human annotators assess whether the summary is correct. This will allow for better assessment of summarization models, as improvements in text generation and neural networks continue to improve the quality of summarization systems. We plan to evaluate the efficacy of our metric by comparing it to existing human annotations on summary outputs, as well as through other meta-evaluation metrics.

The primary focus of this research is developing different metrics and testing how closely they relate to human annotations on summary model outputs. A considerable number of human annotations exist for the CNN/DailyMail corpus [2], but we hope to consider additional datasets as well.

The major challenges in this project will be fine-tuning models to improve metrics without overfitting to the specific data, training new models as appropriate, and developing new metrics unrelated to previous metrics. It will also be challenging to determine the efficacy of our metric on datasets where human annotations are not already available.

2 Project Goals

2.1 75% Project Goal

- Propose a metric for evaluating the factuality of a generated summary using different strategies
- Evaluate this metric against the SummEval human annotations for the CNN/DailyMail dataset

2.2 100% Project Goal

- Test multiple metrics for evaluating the factuality of a generated summary using different strategies.
- Evaluate metrics against the SummEval human annotations for the CNN/DailyMail dataset
- Propose one or more metrics that outperform existing metrics.

2.3 125% Project Goal

- Test multiple metrics for evaluating the factuality of a generated summary using different strategies.

- Evaluate metrics against the SummEval human annotations for the CNN/DailyMail dataset
- Propose one or more metrics that outperform existing metrics
- Evaluate metrics against human annotations for other summarization datasets

3 Project Milestones

3.1 First Technical Milestone:

I intend to complete my literature review and related work section, allowing me to get a more complete understanding of the state of the art in summarization and summarization metrics, as well as other relevant metrics for the broader field of text generation.

3.2 First Biweekly Milestone: February 1st

I intend to reimplement one or two of the current state-of-the-art metrics in evaluating the factuality of generated summaries and/or generated text. By doing so, I will gain insight into what my own metrics should look like and a jumping off point for where to start.

3.3 Second Biweekly Milestone: February 15th

I intend to have determined which datasets I will be able to use to test my metric. This may involve launching human annotations of my own, or simply identifying existing human annotations for factuality on commonly used datasets.

3.4 Third Biweekly Milestone: March 1st

I hope to have chosen a metric to implement or modify, and to have the fundamental code developed for this algorithm. By this point, I will also have begun training for any pre-trained models if applicable.

3.5 Fourth Biweekly Milestone: March 15th

I expect to have implemented more than one potential metric at this point, and to have begun fine-tuning my models to improve accuracy of classification.

3.6 Fifth Biweekly Milestone: March 29th

I hope to have evaluated all metrics against the datasets previously chosen, and to have found one model for a metric that outperforms other existing metrics against human annotations or meta-evaluation systems.

3.7 Sixth Biweekly Milestone: April 12th

By this point, I intend to have developed a qualitative understanding of the performance of my metrics relative to existing metrics. This involves understanding why the metrics outperform or underperform existing metrics and future directions for improvement.

3.8 Seventh Biweekly Milestone: April 26th

I expect to have my research finished by this point, with a paper mostly or fully written and of publishable quality. This also involves finishing all goals of the 100% project goal as above.

4 Literature Search

Text summarization is a well-established field, with widely varying models, datasets, and evaluation metrics. Summarization has recently moved more towards abstractive summarization rather than extractive summarization, as this allows for more flexibility in the generated summaries. However, since abstractive summarization largely requires text generation models, this introduces the problem of hallucinations, or information not present in the source document. These have been shown to be widespread [8] and are understandably problematic for summarization in practice. In particular, as models become more abstractive, they tend to become less faithful to the source text [6].

Existing metrics for detecting hallucinations or measuring factuality include question answering [1], training a classifier on rule-based perturbations of the source text[5][7], using natural language inference to determine whether the summary sentence is entailed by the source document [3], text generation as a metric[9], or creating a fact database of entity relation triples to check the summary against [4].

5 Resources Needed

We expect to primarily use Python and the deep learning framework Pytorch. We may also use the Transformers library developed by the Hugging Face team and/or the bert-extractive-summarizer library. We plan to use datasets available through SummEval [2], as well as potentially other common summarization datasets. We may also need access to external GPU resources if possible, depending on the length of training time needed for the models.

References

[1] 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*.

[2] Fabbri, A.R., Kryściński, W., McCann, B., Xiong, C., Socher, R. and Radev, D., 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics, 9*, pp.391-409.

[3] 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).

[4] Goodrich, B., Rao, V., Liu, P.J. and Saleh, M., 2019, July. Assessing the factual accuracy of generated text. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 166-175).

[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] Ladhak, F., Durmus, E., He, H., Cardie, C. and McKeown, K., 2021. Faithful or Extractive? On Mitigating the Faithfulness-Abstractiveness Trade-off in Abstractive Summarization. *arXiv preprint arXiv:2108.13684*.

[7] 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*.

[8] Maynez, J., Narayan, S., Bohnet, B. and McDonald, R., 2020. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*.

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