Entity-Level Factual Consistency of Abstractive Text Summarization

Anirudh Kansal
Dept. of Electrical Engineering
IIT Bombay
Mumbai, India
20d070013@iitb.ac.in

Mohit
Dept. of Electrical Engineering
IIT Bombay
Mumbai, India
20d070052@iitb.ac.in

***Abstract*—Factual consistency has been a concern with abstractive summarization; The models may tend to produce plausible sounding yet factually incorrect summaries of documents. The problem of generation of names in the summaries that do not exist in the source document, or 'entity hallucination' has been addressed by simply filtering the training data. Further, a technique has been used to promote entity-level precision, which involves the generation of salient named-entities before the complete summary.**

***Index Terms*—Abstractive summarization, entity hallucination, JAENS**

I. Introduction

There are many limitations facing neural text summarization the most serious of which is their tendency to generate summaries that are not factually consistent with the input document; a factually consistent summary only contains statements that can be derived from the source document.[2]

The factual inconsistency shows up at either *entity level* or *relation level*. At the entity level, a model generated summary may contain named entities that never appeared in the source document. It has been referred to as the *entity hallucination problem*.[2]

[1] shows examples of an article, whose summary generated was - *"A six-year old boy was injured in a car accident"*, but the article did not mention the boy's age. This is an example of factually inconsistent summary.

[2] shows an example of entity hallucination where the model generated the summary - *"People in Italy and the Netherlands are more likely to consume fewer cups of coffee than those in the UK, a study suggests."*. Meanwhile, the article never mentioned "UK" in the content."UK" was a result of model hallucination.

The relation level inconsistencies occur when the entities indeed exist in the source document but the relations between them are not in the source document. This type of inconsistency is much harder to identify.[2]

II. Methods Used

Entity based data filtering: To find every named entity in a dataset, Spacy NER was run on the ground truth summary, or gold summary. The sentence containing the entity was removed from the ground truth summary if any of the entities cannot be found in the source document. The document-summary pair was eliminated from the dataset if the ground truth summary is one sentence long and had to be discarded. In this method, we can be sure that the ground truth summary in our filtered dataset is free of entity hallucinations.

This is based on the assumption that the problem of entity hallucination originates from training on the kind of data the filtering process removes.

JAENS: The JAENS (Join sAlient ENtity and Summary generation) approach has been discussed in [2]. This is another technique used to enhance the quality of summarization. The paper attempted at marking the things which were *"summary worthy"*. A named entity in the source document that is worthy of being included in the generated summary is one that is prominent enough to be included in the ground truth summary. Encoder representation was used for the same, and this was accomplished by integrating a classification head into the BART encoder. Initially, the matching named-entities were located in the ground truth summary for then the classification of "summary worthy" entities to take place.

Then the labels - 'Beginning','Inside' or 'Outside', were assigned to each token of the source document to denote if the token was beginning, inside or outside of a summary-worthy named-entity, respectively. During training,the classification loss for each token was added at the encoder to the original sequence-to-sequence loss.

JAENS encourages the model to jointly learn to identify the summary-worthy named-entities while learning to generate summaries. Since the decoder generates the salient named-entities first, the summaries that JAENS generate can further attend to these salient named-entities through decoder self-attention.[2]

III. Results

The techniques discussed in section 2 were implemented. The provided code (link given in section 5), has three segments.

The first file is an implementation of pretrained BART[1] on the CNN dailymail dataset. It compares the results of the model's summaries before fine tuning on the dataset and after

[1]https://huggingface.co/facebook/bart-base

doing so. Some summary pairs before and after fine tuning are as follows.
Before (1):
(CNN) – A Florida judge sentenced Rachel Wade, the 20-year-old Wade
After (1):
NEW: Rachel Wade's lawyer says the sentence is "very fair"
Before (2):
HONG KONG, China (CNN) – From the runway to the store rack
After (2):
Vivienne Tam is one of the world's leading fashion designers.
The second file applies both the entity based filter and JAENS approach to produce output, and we observe the following summary before and after pairs. Due to the training sets being small in size, we do not find reduction in number of document-summary pairs due to filters, although it significantly reduced the size of ground truth summaries because the **training time reduced from 2 hours to 1 hour**.One summary pair before and after fine tuning with filtering and JAENS are as follows.
Before (1):
(CNN) – British mercenary Simon Mann has been jailed for 34 years for his part
After (1):
Simon Mann Equatorial Guniea Simon Mann was arrested after a plane carrying him and about 60
(We observe the keywords being paced at the start of the genrated summary. Due to limitation on cpacity, unfortunately, we cannot observe the potential complete results. But going by the idea of self attention as mention in section 2, this would most likely lead to better summary generation).

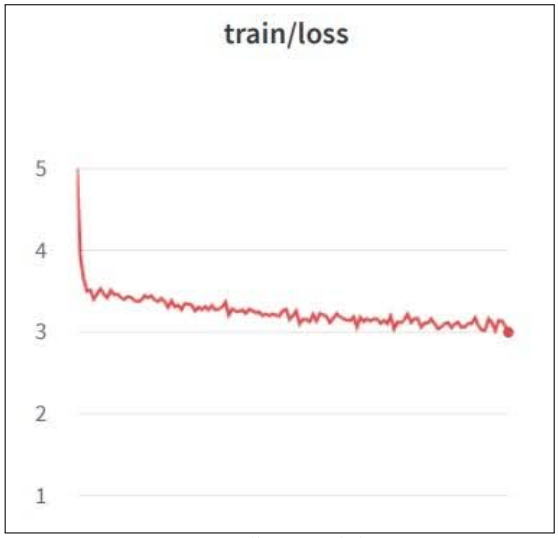

```
{'eval_loss': 5.800720691680908,
 'eval_rouge1': 15.9206,
 'eval_rouge2': 5.2129,
 'eval_rougeL': 12.6713,
 'eval_rougeLsum': 14.5374,
 'eval_gen_len': 20.0,
 'eval_runtime': 340.1088,
 'eval_samples_per_second': 8.444,
 'eval_steps_per_second': 2.111}
```


Normal finetuning of BART

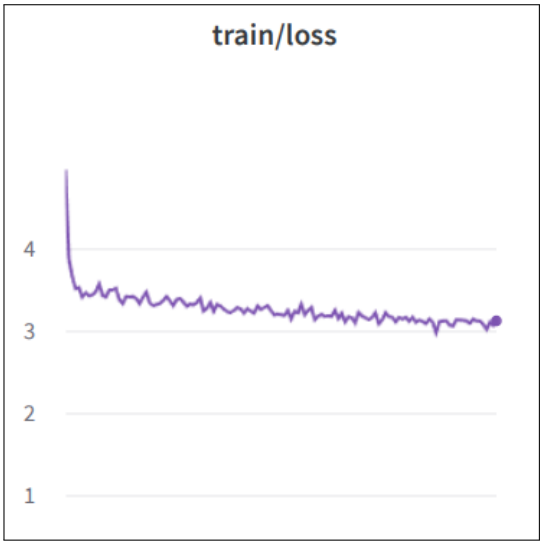

```
{'eval_loss': 3.6375041007995605,
 'eval_rouge1': 21.7051,
 'eval_rouge2': 8.5602,
 'eval_rougeL': 17.3585,
 'eval_rougeLsum': 20.0343,
 'eval_gen_len': 20.0,
 'eval_runtime': 262.5708,
 'eval_samples_per_second': 10.938,
 'eval_steps_per_second': 2.735,
 'epoch': 1.0}
```


Training of entity filter + JAENS on BART

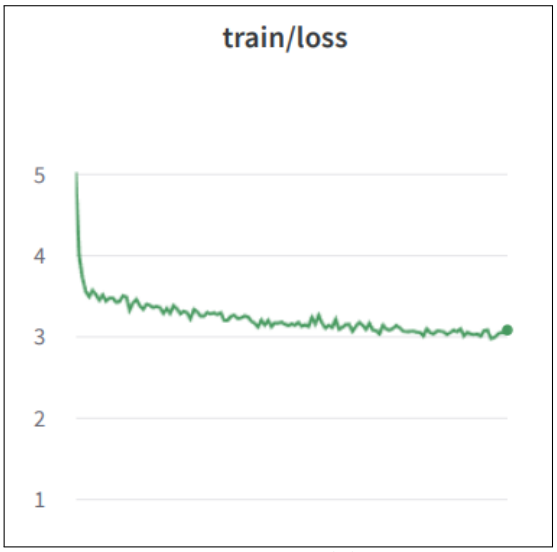
For the last file of code, which is a demo of the model training done, and compares the baseline model with - (i) entity-filtered model, (ii) JEANS model, and (iii) entity-filtered + JAENS model. Following are the training loss graphs with the number of documents of the different models.



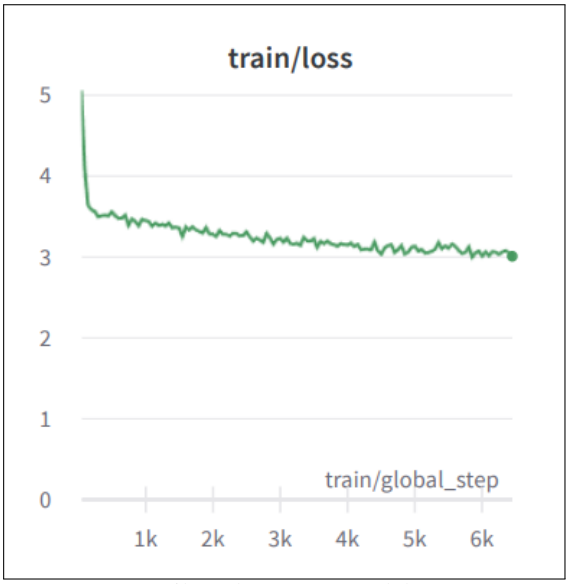
Baseline model



Filtered model



JAENS model



Filtered+JAENS model

IV. Conclusions and Discussion

There were some problems faced, related to the high training time of the data. Along with the training time, the decoder max length was limited, which could have very well affected the output. In fact, we observe the same when we observe the produced summary versus the ground truth summary.

JAENS and entity filter seems to be an effective way to deal with factual inconsistency in the summary generation. The papers [2] and [3] discuss new metrics for evaluating the quality of a summary. [3] uses a technique of Question Answering to achieve the purpose. Use of this metric could accelerate the summary generation to a better result.

[1] discusses QUALS (QUestion Answering with Language model score for Summarization) method which is claimed to be an improvement from QAGS (Wang et al., 2020) protocol.

The project has not exploited all the ideas of the papers, and further implementing them could prove to be very advantageous for the models of summary generation

References

[1] Feng Nan, Cicero Nogueira dos Santos, Henghui Zhu, Patrick Ng, Kathleen McKeown, Ramesh Nallapati, Dejiao Zhang, Zhiguo Wang, Andrew O. Arnold, and Bing Xiang. 2021. Improving Factual Consistency of Abstractive Summarization via Question Answering. *In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6881–6894, Online. Association for Computational Linguistics.

[2] Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cicero Nogueira dos Santos, Henghui Zhu, Dejiao Zhang, Kathleen McKeown, and Bing Xiang. 2021. Entity-level Factual Consistency of Abstractive Text Summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2727–2733, Online. Association for Computational Linguistics.

[3] Daniel Deutsch, Tania Bedrax-Weiss, and Dan Roth. 2021. Towards Question-Answering as an Automatic Metric for Evaluating the Content Quality of a Summary. *Transactions of the Association for Computational Linguistics*, 9:774–789.

V. Key Links

- Drive link - https://drive.google.com/drive/folders/1yUbYTNKQ6WQmA2YV5yp_MhBBsS2JSRJI (Consists of video recording also)