Ahmed Iqbal
CS 4395.001
4/8/2023

ACL Paper Summary

The title of the paper is “Hallucinated but Factual! Inspecting the Factuality of Hallucinations in Abstractive Summarization.” It was written by Meng Cao, Yue Dong, and Jackie Chi Kit Cheung. The paper is affiliated with the School of Computer Science at McGill University in Montreal, QC, Canada and MILA, Montreal, QC, Canada.

The problem that these researchers have presented is that abstractive summarization systems, which aim to generate summaries of the source text by singling out the main points, often produce hallucinations, which are inferences from the source text. These inferences may or may not be factual, and they represent true facts in the real world that didn’t come from the source texts. Previous studies start with the assumption that all hallucinations are undesirable and thus try to eliminate them. The researchers of this paper are trying to find a way to separate the factual hallucinations from the unfactual ones, which will increase the quality of these abstractive summary systems.

The prior work in this field concentrates around two poles: model hallucination, and summary factuality. For model hallucination, previous studies have found that all models generate hallucinations. There have also been recent studies to investigate how to reduce model hallucination, such a loss truncation algorithm that filters out noisy training samples and verification systems to recognize nonfactual hallucinations. Another method suggests that a conditional learning model puts more probability on non hallucinations. This current study will diverge from this prior work in that it will also account for the factuality of the hallucinations. For summary factuality, priors works suggested various ways to check for this. One involves training a model on a wrong dataset to then check for errors. Another involves checking whether the semantic relationships in the summary are present in the original source text.

The unique contributions of the authors begin with their stance that hallucinations are not always undesirable. The method that these authors propose is to check the factuality of a hallucination from an abstractive summarization system by comparing the summary to other factually correct summaries that cover the same topic. This “factuality detection module” uses human written summaries to ensure that the summaries that the abstractive summaries are compared against are factually correct. They also find a way around the noise, content from outside source text in this case, by using the predictions from the classifier that they detail in their paper as the factuality reward signals to guide the training of the model. SInce there was no dataset available to the authors of this paper for their purposes, they created their own, which is also a contribution.

The authors evaluate their method in multiple ways. One of them is to compare their hallucination labels and factuality labels with those before the other baseline approaches. Another is to train a summarization model using their factuality rewards on the XSum test set. Then they calculate the percentage of n-grams that are present in the summary and not in the source text, as these novel n-grams demonstrate how much a model is extracting beyond the source text.

Meng Cao has 175 citations, Yue Dong has 170, and Jackie Chi Kit Cheung has 2199. Jackie Chi Kit Cheung has the most number of citations.

I think the work of these authors is important because it provides a novel way to evaluate whether a hallucination is factual or not. Hallucinations are a part and parcel of abstractive summary systems just because of the nature of how these systems work. They are going to make inferences, just like humans often do. What the authors contribute is a way to sift through these inferences and determine which are factual or not, which has implications as to the efficacy of an abstractive summarization system. After all, if a user wants to select a system for their summarization task, they would avoid ones that are prone to including non factual information.