Paper Review: Entity-level Factual Adaptiveness of Fine-tuning based Abstractive Summarization Models, Song et al.

Review by Jason Gillette

Introduction

This paper seeks to address the problem of factual consistency in summarization tasks for large language models. In a typical summarization tasks the model input is a large string of text of a document containing information and the expected output is a summarization of the input with the expectation of factual consistency between the input and output. Generative models may generate outputs that are consistent with their pre-training data and contrary to the input data, commonly referred to as a hallucination. Reducing the occurrence of counterfactual hallucinations is critical in improving the performance, reliability, and utility of large language models for summarizations tasks. The authors seek to reduce these hallucinations by measuring measuring factual inconsistencies between pre-trained models and common evaluation datasets and by introducing fine-tuning techniques in an effort to reduce counterfactual hallucinations. THeir research question is not expressly stated, but can be paraphrased as follows; how can we improve the robustness of fine-tuned summarization models against knowledge conflicts, ensuring that the generated summaries remain factually accurate when the input contains information that contradicts the model’s pre-trained knowledge?

Contributions

The authors highlight (3) contributions of their research.

1. Introducing the concept of factual adaptiveness in fine-tuning based summarization models using the parametric knowledge of a pre-trained language model. Factual adaptiveness refers to the model's ability stay accurate even when the input includes information that conflicts with pre-trained knowledge.
2. A demonstration that factual consistency on original datasets tends to be orthogonal to factual adaptiveness. Specifically, data filtering largely improves factual adaptiveness while advanced decoding and contrastive learning show minimal differences. In other words, performance in factual consistency, or the ability to produce summaries that accurately reflect the facts in non-conflicting input document, did not have an impact on factual adaptiveness, when using traditional methods. Use of data filtering methods on fine-tuning data reversed this trend and demonstrated improvements in factual adaptiveness.
3. In light of the results, the authors propose a controllable counterfactual data augmentation method that enhances factual adaptiveness while preserving factual consistency on original datasets.

In addition to the (3) contributions noted by the authors, there are (2) key components in this paper. They are the metric of conditional likelihood (MCL) and metric of factual consistency (MFC). Each of these are the primary mechanisms by which the authors are able to demonstrate their methods and justify their contributions. MCL assesses the model's likelihood for the original entity versus a counterfactual entity when generating a summary on a counterfactual input, highlighting the model’s adaptive behavior when

generating entity-level content. MFC evaluates the change in factual accuracy between summaries generated from original and conflicting documents, emphasizing overall summary consistency when knowledge conflicts arise.

Methodology

To set up their experiment, the authors selected (2) well-known models for text summarization tasks, PEGASUS and BART. Each model is fine-tuned on well-known text summarization datasets, XSum and CNN/DailyMail. Each of these datasets are extractive vice generative. This was an intentional choice by the authors, as extractive summaries are more closely aligned with their input documents at the entity-level. With each model pre-trained, the authors introduce counter factual sampling as a means of generating counterfactual data for evaluation. For the original fine-tuning data, the authors preform entity replacement. For instance, to create a counterfactual sample, an entity in the original document (e.g., person or place name) is identified and replaced with a conflicting entity. For instance, "Bangladesh" might be replaced with "Queensland". With fine-tuned models and counterfactual data, the authors use (2) evaluation "scenarios". In the first, they evaluate the model's propensity toward pre-trained knowledge for a certain entity by passing a partial summary and a generic unrelated document to each model. They can then see how the model generates information relative to that entity, despite the lack of input context. In the second scenario, the authors pass the counterfactual samples and use MCL and MFC do gather an empirical measure.

Results and Analysis

The evaluation results of the paper show that while techniques aimed at improving factual consistency, such as data filtering, contrastive learning, and advanced decoding, can improve a model’s ability to generate factually accurate summaries, only data filtering significantly boosts factual adaptiveness. The metrics MCL and MFC demonstrated that data filtering led to the greatest improvement in the model’s ability to handle entity-level conflicts without generating hallucinations based on pre-trained knowledge. Traditional techniques like advanced decoding and contrastive learning showed minimal impact on factual adaptiveness, indicating that general factual consistency improvements do not automatically translate to robustness against conflicting information. The proposed counterfactual data augmentation method (i.e., entity-level replacement) proved effective, showing that models trained with these augmented datasets better handle knowledge conflicts while maintaining overall factual consistency.

Discussion

The authors show a deep understanding of summarization and the challenges associated with it. The contributions of this research shine light on significant challenges and offer a sufficiently novel approach to improving performance. However, there are two criticisms that can be leveraged.

1. Accessibility and Clarity: The paper’s dense descriptions make it challenging to understand, especially for those who are not deeply familiar with abstraction summarization. This may limit the paper’s impact by making it difficult for researchers, practitioners, or others in adjacent fields to fully grasp the contributions, methodology, and significance of the findings. Throughout the paper, sections dedicated to detailed descriptions of methods simply repeated the same terms detailed in the introduction without offering any further explanation or examples. Improved clarity and simplified explanations of concepts like MCL, MFC, and counterfactual data augmentation could make the findings more accessible to a broader community of researchers.

2. Limited Scope of Factual Inconsistency: The authors focus exclusively on entity-level factual inconsistencies and evaluate only a narrow set of entity types like names and locations, which leaves broader types of factual errors unexplored. Factual inconsistency in summarization can also involve relationships between entities, semantic roles, causal chains, and other forms of factual structure, which are critical in complex texts. This limitation suggests that while the paper addresses one specific form of inconsistency, the models may still struggle with other inconsistencies, especially those beyond straightforward entity substitution. This narrow focus might lead to models that are less robust in real-world applications, where factuality issues are often more nuanced. MOreover the evaluation methods proposed and used by the authors may in fact be misleading, as they do not represent a broader scope of counterfactual summarization.

Conclusion

This paper contributes valuable insights into improving factual consistency and adaptiveness in summarization models. While effective in introducing new methods and metrics, the scope is limited to entity-level conflicts, potentially overlooking broader factual inconsistencies. Further research could explore these limitations.

Reference

Entity-level Factual Adaptiveness of Fine-tuning based Abstractive Summarization Models (Song et al., EACL 2024)