Master Thesis Review

Faculty of Mathematics and Physics, Charles University

Thesis Author	Saad Obaid ul Islam
Thesis Title	Tackling Hallucinations in Chart Summarization
Submission Year	2022
Study Program	Computer Science
Branch of Study	Language Technologies and Computational Linguistics
Review Author	Ondřej Dušek **Role** Supervisor
Department	Institute of Formal and Applied Linguistics

Review Text:

Thesis Topic Current state-of-the-art methods in natural language generation (NLG) are based on pretrained neural language models (PLMs) finetuned to the particular target task, such as data-to-text generation, dialogue response generation, or chart summarization, the latter of which is the topic of Saad Obaid ul Islam's thesis. While PLM-based NLG systems produce very fluent and convincing outputs, their major problem are *hallucinations*, i.e. system outputs that are not grounded in the input – the PLM "hallucinates" text that fits the particular situation and may or may not be true, but has no support in the input data/table/chart. Removing hallucinations is a major hurdle to PLMs' real-world usability, and there is a lot of active research aiming towards that goal. Saad's thesis is thus very topical and aims at state-of-the-art results in chart summarization.

The approach taken by Saad uses the current PLMs and finetunes them, same as previous approaches, but makes two main changes to the overall setup specifically to combat hallucination: (1) changing the way the PLMs used (with a two-step setup where the second step corrects the output of the first one) and, most importantly, (2) improving the quality of the input data the PLM is finetuned on. The data improvements are again twofold: (a) optimizing the linearization of the chart, which is an important step but overlooked in current literature, and (b) removing noise from the training summaries. The approaches taken are well positioned to reduce hallucination and provide better grounding to the generating PLM.

Contents Summary The thesis has three main sets of experiments, all aiming at reducing hallucination and improving chart summary quality:

1. *Linearization improvements:* Saad analyzed prior state-of-the-art works on chart summarization and identified problems with the input chart linearizations. The input linearizations were missing some important information (chart or axes titles), and included long-distance dependencies (corresponding x-y values were not adjacent). Therefore, Saad proposed an improved input chart linearization that includes all important information from the data and puts related chart values together.

2. *Data cleaning:* Saad analyzed the underlying datasets and found noise in the training summaries. He further demonstrated that this noise is one of the causes of hallucination in a simple experiment: He used a noise-free template-based chart summarization dataset and injected noise into it by generating random sentences from a vanilla PLM. While a PLM finetuned on the noise-free data did not produce hallucinations, the same PLM finetuned on noised data did.[1] This lead to an additional proposal for automatically cleaning the training data using a natural language inference (NLI) model. NLI had previously been used with success to check for hallucinations in data-to-text NLG system outputs. Here, the NLI model was used to check if each sentence from a training chart summary is entailed by the corresponding chart data. If the entailment probability was lower than a threshold, the sentence was removed. Finally, Saad simply finetuned his PLM on the data cleaned in this way.

3. *Two-step decoding:* Saad also devised a two-step decoding approach to further improve results on the cleaned data. This also worked around a problem with output summaries being too short, which emerged as a result of data cleaning. Using a small portion of the data which he hand-cleaned and expanded to include more

[1] Note that the noise-free template-based dataset is perfect for an experiment like this, but is not applicable in a real-world scenario as finetuning on it leads to the PLM reproducing the templates and producing repetitive outputs.

information based on the chart data, he further finetuned the PLM for a second, "correcting" step. The chart summary is thus first generated by a first PLM and then further corrected (and potentially expanded) by a second PLM.

All models are carefully evaluated using automatic metrics and manual error analysis. The improved linearization is compared against previous approaches to linearization on two datasets, with one selected for manual analysis. The data cleaning and two-step decoding experiments then compare to the improved linearization model as a baseline. The experiments with two-step decoding are properly ablated, checking for contributions from the two-step setup and from training on hand-cleaned data. In addition to performing manual error analysis, Saad ran a human evaluation experiment via crowdsourcing on the Prolific platform for the data cleaning and two-step decoding models, comparing them to the improved linearization baseline.

The experimental results are mostly positive: The improved linearization yields both major gains in automatic metric scores and a reduction of in number of hallucinations as per manual analysis. The data cleaning does not increase automatic metric scores, but shows a significantly reduced number of hallucinations. Finally, the two-step decoding leads to better automatic metric scores and better informativeness according to human evaluation (corresponding to expanded summaries), but does not significantly reduce hallucinations.

Text Structuring The text of the thesis is structured into 7 numbered chapters: Chapter 1 consists of a short introduction and a summary of the thesis contributions, Chapter 2 summarizes the important theoretical background (NLG architectures and evaluation metrics). Chapter 3 introduces related work, i.e. the PLM-based architectures this thesis builds upon and extends. Saad's original contributions are included in the next three chapters: Chapter 4 contains an analysis of the data (both training datasets and baseline system outputs) and formulates hypotheses about linearization and data cleaning. Chapter 5 then acts on these hypotheses, introducing improved linearization and experimentally demonstrating that training data noise causes hallucinations. Chapter 6 recounts the experiments with data cleaning via NLI and two-step decoding, including the human evaluation setup and results. Finally, Chapter 7 sums up the main take-aways and adds some remarks on potential related work.

Work Progress The work on this thesis started in early 2022, in close collaboration with co-supervisors Vera Demberg and Iza Škrjanec from Saarland University. We held frequent meetings and collectively discussed further steps based on Saad's results and data analyses. In this time, Saad demonstrated his ability to build successful deep-learning-based models and properly design and run experiments with them. The text of the thesis emerged in late 2022 after the experiments were finished, and again involved a lot of detailed feedback and multiple rounds of reading from both myself and the co-supervisors from Saarland.

Overall evaluation I believe that the resulting thesis fulfills all the requirements and standards of Charles University. The approaches taken are novel and significantly reduce the number of hallucinations produced in chart summarization. This is also why we prepared a paper based on this thesis, which is now under review at the prestigious Association for Computational Linguistics conference (ACL'23).

I am very satisfied with the end result. Any significant problems in the thesis from my point of view were already discussed between Saad, myself and the co-supervisors, and subsequently solved by Saad before the submission. I do not have any specific questions for the defense.

I recommend that the thesis be defended.

I do not nominate the thesis for a special award.

Prague, 24 January 2023

Signature: