Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Saad Obaid ul Islam
Název práce	Tackling Hallucinations in Chart Summarization
Rok odevzdání	2022
Studijní program	Computer Science
Studijní obor	Language Technologies and Computational Linguistics

Autor posudku Rudolf Rosa **Role** oponent

Pracoviště Ústav formální a aplikované lingvistiky

Text posudku:

The Thesis

The thesis focuses on chart to text generation, which is a specific subtask of data to text natural language generation. In this task, we want to generate a text based on a graphical chart (represented by a data table). In this task (as in many data to text tasks), one common issue of current approaches is *hallucination*, i.e. outputing *unfaithful* information which is not based on the inputs. This is a very current and very important issue.

The thesis reviews and analyzes the current situation and proposes two remedies: using a better input format (i.e. linearizing the input data in a better way), and filtering out unfaithful sentences from the training data. Both of the approaches are sensible, reasonably based on existing approaches and of performed analyses of the existing systems. I also like that the approaches are rather simple and elegant. Each of the approaches brings very significant improvements in the produced outputs by significantly reducing hallucinations while typically also improving the overall quality of the outputs. I also highly acknowledge the fact that strong improvements of results were obtained while at the same time also decreasing the computational complexity of the training. Also, it is fascinating that further large improvements were obtained by manually curating only a rather small portion of the dataset. A number of manual analyses and evaluations are performed, which makes the claims rather strong and trustworthy.

The thesis is considerably strong research-wise, definitely above average. Not that many theses manage to significantly improve upon state of the art approaches by targetting a current problem with simple and elegant measures.

The Approach

The approach used in the thesis is definitely its strength, while I do have a few reservations.

As for the improvements towards the input format, I like how simply bringing related pieces of information closer together makes the processing simpler, thus reaching much higher quality while also achieving much faster training times. It is a pity, though, that only one suggested format is used, without any discussion why this one format in particular is the best one and whether other formats might also work well. These experiments seem pretty simple, so I would very much appreciate if multiple variants of formatting the input had been tried out. For example, the Obeid and Hoque format looks good in locality of the information (but seems to be missing the title, which is absurd) while the Kanthara format contains the title (but locality of information is bad); an obvious first experiment thus would be to use the Obeid and Hoque format but also include the title? To me, a most logical format would instead be to encode the information into triplets of (x-label, y-label, value). And there are definitely other simple options that could be easily tried out (or at least discussed in the thesis, with some argumentation of why the one and only format that is tested there had been chosen). Also, the authors argue that the improvement comes from improving the locality of the information, reducing long range dependencies in the

input. However, the effect of including the graph title is probably also very strong and should be measured independently.

Also, as is discussed in future work, it would be nice and very easy to include some statistics into the input (such as range, highest value, lowest value) which the generation system could then pick up and include into the summary. While I understand that this was beyond the goal of the thesis, it seems like an experiment that would have been very easy to do.

As for the second goal of making the chart descriptions in the dataset more faithful, I very much like the approach of automatically filtering the training data, it seems very sensible and has great results.

I do find the experiments with the synthetic Autochart dataset questionable, as I do not fully understand how any experiments on a fully synthetic dataset (and seemingly also of a quite low quality, as shown e.g. in Figure 3.2 on page 21, even though the authors claim to be getting "coherent, fluent" summaries on page 34) are relevant for real life data; I believe that such experiments may merely suggest some things but cannot really prove anything. In any case, I do not see an urgent need to prove Hypothesis 2, as it seems rather logical and uncontroversial to assume it is true (even the text of the thesis does not provide any alternative hypothesis that could explain the observations instead of Hypothesis 2).

As for the follow-up experiments where additional summary sentences were manually written, I very much appreciate that, but the approach does not seem completely logical – first, some sentences are filtered out, which might also incorrectly filter out some faithful sentences; and next, new faithful sentences are manually created and added. The logical approach might rather be to supervise the automated filtering, manually fixing some errors of the filtering, such as preventing the removal of the faithful sentences in the first place? Or maybe this is where the Autochart approach could actually be used to generate some synthetic faithful sentences? Also, the author notes issues due to differences in number formating; obviously, if this is a frequent error, it would be trivial to perform number normalization to get rid of it?

The Text

The text is the weaker part of the thesis, and I have numerous reservations here. Some parts are written very well, but in other parts the writing could be better, as it is often too brief, does not give enough details, does not sufficiently explain or disucss some aspects, etc.

Interestingly, it is never explained or dicussed whether hallucinations are always bad. It is simply posited that a hallucination is something not based on the input, and equals being unfaithful or non-factual. However, I strongly believe that a hallucination may be unfaithful to the input but still truthful and thus factual? The text of the thesis even contains such examples, such as a generated description of a GDP chart which also gives a (correct) definition of GDP (Table 5.4 on page 32), where the thesis author simply states that the definition is not based on the input and we can not check its factuality based on the input, but does not even touch on it being actually correct. Also, I believe that such unfaithful but truthful hallucinations can actually improve the quality of the output. However, in the thesis, no distinction between good and bad hallucinations is made, neither in theory nor in methodology, experiments and, very importantly, evaluations (so when hallucinations are reported, we do not know which of them are truthful and which are not). It is understandable that it is hard to evaluate truthfulness and it is thus probably safer to simply get rid of all hallucinations, but this should be discussed. The thesis simply assumes all hallucinations are bad without even discussing this. Of course, we might simply define our task to be generating text without hallucinations, but without discussing this further, this makes the task somewhat artficial, disconnected from the real world tasks.

Also the definitions of hallucinations are not completely clear. E.g. if the input data contain "New York" and the generated text also talks about "USA", is that hallucinated if the string "USA" is not on the input but only inferred by the model from the string "New York"? Such information on the output is clearly true but technically this is adding info that is not explicitly present on input.

Even the definition of the task of chart summarization is not clearly given (page 8). As chart to

text generation is treated here within the data to text paradigm, it should be clearly defined how chart to text differs from general data to text. For example, it is not clear whether the underlying data contain only the input data tables, or also metadata about the chart (such as chart type, chart size, scales, logscales, various labels, coloring scheme, etc.). This becomes partially clear with the examples in later chapters of the thesis, but even then this is not completely obvious. The thesis seems to silently assume that given a data table, the chart can be generated automatically, which is true, but the actual charts in some of the datasets are manually curated, and this manual curation may add information or interpretations not captured by the data table. The thesis does mention that chart to text can also be treated as an image to text task, which then does have access even to such information, but the thesis does not discuss these implications.

The motivation for the task is also not given. Is it meant to generate textual descriptions accompanying the charts (which is how some of the datasets are collected), or is it rather to generate texts replacing the charts, e.g. for situations where visual presentation is unsuitable (such as audio-based presentation)? This has implications for goals and evaluation of the outputs, such as evaluating the coverage if the summary.

Related work for hallucination mitigation (i.e. the actual topic of the thesis) is extremely scarce (16 lines of text in section 3.3.1 on page 22-23). Since this is the main focus of the thesis, this section should have been much more detailed. The section refers to 6 works in total. Two of these target hallucinations by ensuring the datasets are fact grounded; one does so by manual filtering of the data, for the other this is not stated in the thesis. The work of Nie et al. (2019) seems highly relevant, as they use NLU in preprocessing "to improve the equivalence between input data and target", which sounds quite close to the approach used in the thesis; however, the thesis does not elaborate further on this work. Two other works are mentioned as using planning to mitigate hallucinations, but no explanation is given as to what approaches exactly are used and how they contribute to increasing the faithfulness of the outputs. Finally, Rebuffel et al. (2022) is mentioned as using a "multi-branch decoder that leverages world (sic) level alignment labels between the input data and target text to learn relevant parts", but from such a brief description I have no idea what this means. For all of the approaches mentioned here, it is never discussed how they are relevant for the thesis, how the approach in the thesis compares to them, and whether they could be also used for the problem focused in the thesis (or why not).

The structure of the text is sometimes suboptimal, e.g. in chapter 3 where descriptions of datasets and NLG approaches are intermixed (even though here the situations is complicated as the papers introducing the datasets also introduce some of the NLG approaches).

Sections 2.3.4 or 2.4 are given mostly in bullets without any sane introduction as of why this information is presented here, which looks more like an outline of the section that was left unfinished. The parapgraph on BLEU even has a missing citation ("It was proposed by ? to...") and a typo in equation 2.15 (missing closing parenthesis).

Problems of some of the metrics are listed in section 2.4, but not for other metrics; this then makes the evaluations later in the thesis harder to interpret, as some of the automated numbers might be explained by problems of some of the metrics, but the author does not seem to be aware of the shortcomings and thus does not provide any sensible interpretation of the observed results. Most importantly, Table 6.4 on page 40 contains some results of automated metrics which are not in agreement (some report an improvement, some not, BLT drops drastically) but the table is not interpreted or discussed in the text at all (the text only says "Table 6.4 shows the results" but does not comment on the results).

In the data filtering experiments, it is noted that the generated summaries become shorter (while containing less hallucinations), and this is treated as a negative thing. I believe the shortness is not the problem, the problem is rather that the summaries are less informative? But informativeness is not among the evaluated qualities.

Table 4.1 on page 27 is difficult to understand as the descriptions do not sufficiently explain what we see in the table:

- The "templateTitle" strings seem to link to the title of the graph ("U.S. Millenials : most popular social network 2016 , by age group"), but this is never explained explicitly

- If so, the texts and indices do not seem to completely match the input; i.e. "United States" is replaced here by "templateTitle[0]" but the string "United States" does not appear anywhere in the title; "Millenials" is replaced by "templateTitleSubject[0]" which is also not clear where this comes from.
- "2016" is not replaced, even though it is contained in the title
- "August" disappears without being represented by a template string
- Why is "between 18 and 24 years old" replaced by "between 18 and templateValue[1][1] years old"? Both "18" and "24" are contained in the input similarly, so how come one is replaced and the other is not?
- Also this part contributes to the discussion towards faithfulness to the input; is the "years old" part a hallucination, as the input talks about "age groups" but does not explicitly say that e.g. the range "18-24" is in years (while age could potentially also be given in months)?
- The title of the graph does not seem to be at all contained in the actual input, which either is an error in the thesis, or it seems like an apparent error in the dataset; I do not see how should the NLG system be able to refer to specific parts of the title (such as templateTitle[5]) without knowing the actual title, not even its length? From the thesis it seems that the title is indeed not included, then it is not that surprising that including the title dramatically improves the results?

There are occasional language errors (e.g. singular-plural), typos, inaccuracies and other minor issues, e.g.:

- Definition of hallucinations on page 6: "factually correct" instead of "factually incorrect"?
- On page 17: "The reason we use this metric is interesting"; what should this mean?
- Table 3.1 on page 19: no units given – is it charts, sentences or tokens?
- Table 4.3 on page 28: chart data are shown here without showing the actual charts.
- Page 37 talks about the Autochart experiments, stating "the training data contains additional information", while in fact what was added into the dataset was automatically generated noise; I do not agree to referring to noise as information.
- Page 40 says 46 out of 50 summaries did not contain hallucinations, but we cannot compare this number to anything as in section 5.1.2 the number of hallucinations was not reported.
- Also on page 40, it seems contradictory that "NLI+T5 produces single sentence summaries for all the input data" while at the same time "65% of the summaries are of length two" (assuming this is the length in sentences).
- etc.

And a final suggestion: since the thesis focuses on chart to text NLG, and the text of the thesis does in fact contain some charts (and also some data tables that could be made into charts), it would have been very nice and interesting to also include their automatically generated descriptions.

The Grade

Even though I have a number of reservations, especially towards the text of the thesis, I find this thesis to be sufficient in all aspects, and even slightly above average in the research aspects.

I suggest to grade the thesis with the grade 2 or 3.

Práci doporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

V Praze dne 19. 1. 2023

Podpis: