Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Bc. František Trebuňa		
Název práce	Persona-Aware Chatbot Response Generation		
Rok odevzdání	2024		
Studijní program	Informatika	**Studijní obor**	Matematická lingvistika

Autor posudku	Rudolf Rosa	**Role**	oponent
Pracoviště	Ústav formální a aplikované lingvistiky		

Text posudku:

Thesis summary

The thesis tries to improve upon existing approaches to persona-aware chatbot response generation. The thesis first reviews the theory and existing approaches in much detail. It then chooses some of the most promising existing approaches to chatbot modelling and to language modelling in general, and tries to apply and combine these on the task using the standard ConvAI2 dataset.

The thesis reports on a number of experiments, which are automatically evaluated, eventually identifying the most promising setup; the setup technically improves upon previous state of the art, since when the LMEDR approach is complemented with the LambdaRank approach, there is an improvement over the original approach. A nice property of this approach is that it is successful even when the reranker is trained using data from a different system than to which it is applied at inference time, showing nice generalization.

The human evaluation, however, shows that the automated evaluation is not fully trustworthy, and also that the used dataset is not of a sufficiently high quality. In the final experiment, the best developed setups are evaluated against vanilla untuned GPT-3.5, which only underwent a small amount of prompt engineering. Despite the automated evaluation showing the proposed setup to outperform the GPT-3.5 setup, the human evaluation reveals that it is the other way round, and the GPT-3.5 outputs are actually preferable. Moreover, the human evaluation also shows that GPT-3.5-generated outputs are even preferable to the gold standard annotation, i.e. the dialogues contained in the dataset (created by crowdsourcing).

Validity of results and findings

Sadly, in my eyes, this partially invalidates all of the performed research, since the final experiment shows that the quality of the dataset is too low; apparently, a higher quality dataset could be simply generated with GPT-3.5 instead of using the texts crowdsourced from people. Thefore, I believe it is unclear whether the partial findings in the thesis are actually valid: we mostly only have improvements in F1 score measured on the dataset as the sole reason to believe that the methods and approaches used do actually bring improvements in the task. However, since the quality of the dataset is low, being closer to the dataset does not necessarily mean being better. Unfortunately, human evaluation was done only on the final setups, combining all the most promising partial steps; what we would need is also a human evaluation of the individual components of the final setup to actually prove whether they do in fact improve the performance of the system on the task.

Besides, the work discusses a wide range of evaluation metrics, but in the end mostly relies on a single metric, specifically F1 score comparing the generated outputs to the gold texts in the dataset. There is no real discussion how adequate this metric is, how much it correlates with human judgements, and therefore to what extent improvements in F1 score can be interpreted as real improvements in the task itself. The work does acknowledge that automated evaluation is imperfect and only human evaluation is to be trusted, but this is not reflected in the actual methodology of the work, since in most cases, improvements in F1 are without any doubt directly interpreted as improvements in the task.

My reservation towards the validity is thus two-fold: we are estimating the quality of the outputs with a metric which is probably imperfect (and we do not even know how good or bad it is), and the metric is computed on a dataset which we know quite well that is imperfect (and some known problems of the dataset are even dicussed already in section 2.5); the final experiment with human evaluation shows quite starkly how misleading the automated evaluation can be. Therefore, any observed improvements in the value of the metric should be taken only as indications of a possible improvement in the task, and before we make any conclusions about the actual effect of any component of the final method, we should validate these indications with manual evaluation. However, this is not done, the improvements in the automated metric are interpreted directly as improvements in the task, and reported as such in the conclusion of the thesis, without any discussion of the validity of these conclusions. (It may be that even with such an imperfect metric, there is good reason to believe that the indicated improvements are nevertheless real improvements; however, such discussion is not present in the thesis.)

Thus, in my view, the only trustworthy findings of the thesis are the ones validated by human evaluation, which are: (a) prompting an off-the-shelf LLM with a small amount of prompt engineering can easily surpass sophisticated setups trained on the ConvAI2 dataset, and (b) the quality of the ConvAI2 dataset is so low that a better dataset could probably be generated automatically using an off-the-shelf LLM.

The other claimed findings are based only on the automated evaluation which the thesis itself shows not to be very trustworthy. (And even then, many of the reported improvements are quite small, while statistical significance of the score differences is not measured.)

Other parts of the thesis

Consequently, I would also have reservations towards the suggested future work. The work clearly showed that a major bottleneck of the task is the low-quality dataset; therefore, I am convinced that necessarily, the first crucial step in any future work would be to get or create a higher quality dataset (possibly even with the help of state-of-the-art LLMs and prompt engineering).

Besides that, the thesis focuses on a real and current issue, uses sensible modern models and methods and a standard dataset, and combines the components in novel ways to arrive at new findings. The thesis by itself does not propose any new methods or significant modifications of existing methods (it therefore appropriately lacks the usual section dedicated to methods, and rather jumps from Related Work directly to Experiments), which is a pity, but the range of methods used in the work is sufficient in breadth for a master thesis, and the assignment did not in fact call for new methods to be suggested and developed.

Fulfillment of assignment

What the assignment did call for was a wider evaluation, explicitly evaluating to what extent are the generated dialogues in line with the pre-specified persona descriptions. The thesis does discuss a wide range of metrics, but then resorts to using mostly F1 (and, to some extent, hits at 1); however, none of these metrics are specific to persona-aware chatbot evaluation, and persona-specific evaluations are extremely rare in the thesis (I have only found one instance, in figure 5.2b).

In fact, the persona-awereness is not really specifically targetted by the thesis. The thesis does use a dataset containing persona specifications and persona-specific dialogues, but the actual approach used in the thesis mostly does not contain anything specific for the personas; the persona specifications are simply treated as part of the prefix of the language modelling task and not really handled in any special way, neither in training, nor in inference or evaluation.

The author does at one place (section 5.1.6) notice that the models tend to repeat text from the persona specification, and that the models also tend to confuse the persona specification with the dialogue history, which he simply lists as problems that were noticed, but does not try to address that in any way (such as experimenting with the input format – which could probably help, since the format lacks a clear unique separator between the persona specification and the dialogue).

The text of the thesis

The thesis is written in very good English, the concepts are well explained and easy to understand, the typography and graphics are very good, the broad review of theory and related works is impressive (probably even unnecessarily extensive).

I have found only one confusing inaccuracy in formulations, where the thesis seems to implicitly assume that RLHF is done based on human feedback, while DPO is done without human feedback. I agree it is probably usually so, but it need not be so in principle, and in any case, this should be explicitly stated in the text. The text says in Section 1.5.6 that RLHF is performed based on human feedback, which is true, and then says in Section 1.5.7 that DPO is an alternative to RLHF that differs in substituting the reinforcement learning with a specially crafted loss, which is also true. This logically implies that DPO also uses human feedback (which can be true) as not using human feedback is not mentioned as a difference between DPO and RLHF. However, when DPO is used in Section 5.5, it suddenly goes without saying that no human feedback is actually used, and it is somewhat hard to decode from the text what is used instead of the human feedback, which confuses the reader and contradicts the previous explanation.

Summary

To sum up, the thesis presents an adequate amount of work for a master thesis and shows a good orientation of the author in the problem and methods.

However, most of the claimed findings are not trustworthy, since the final experiments expose the evaluation approach as not trustworthy, probably mainly due to the low-quality dataset. This by itself could be a valid outcome of research; it does happen rather often that after a considerable amount of tedious work, the researcher arrives at a finding that renders all the previous work useless, be it due to an overlooked error or simply due to finding something that was simply not known before (such as the fact that the standard dataset is in fact of a rather low quality). However, a researcher must then be honest and truthful, and must interpret all the results in the light of this finding.

The thesis also fulfills the assignment only partially, especially by mostly following a generic approach applicable to any chatbot modelling (or even to any language modelling in general), without explicitly focusing on the persona-awareness (which was the goal of the assignment).

I believe the quality of the thesis is still sufficient to be defended, but I do not think that it deserves the highest grade.

Questions for the defence

At the defence, I would obviously like the author to defend the approach used and the validity of the findings, especially explaining how the approach is adequate for the task (specifically with the persona-awareness in mind) and whether the findings based on the automated metrics are trustworthy. I would also like the author to comment on why he eventually used F1 score as the primary (and often only) metric.

Besides that, I encourage the author to react to any other parts of my review, especially if he strongly disagrees with my opinion. As chatbots are not really my primary domain of expertize, it is quite possible that I misunderstood or misinterpreted something, probably due to my low experience and knowledge of this field and its common practices. However, I admit that my review is rather extensive, and therefore I do not require the author to explicitly react to each and every point I make in the review.

Práci doporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

V Praze dne 16. 8. 2024

Podpis: