Review of the paper ”TRUE: Re-evaluating Factual Consistency Evaluation” by Or Honovich et al.

Klavdiia Naumova 342018

May 12, 2023

A paper entitled ”TRUE: Re-evaluating Factual Consistency Evaluation” [1] (hereafter, I am referring to it as TRUE for brevity) presents a survey of existing metrics for factual consistency evaluation of NLP models and an assessment of these metrics on a set of standardized text datasets manually annotated for factual consistency. The authors claim the importance of automatic evaluation metrics for NLP tasks and recommend using Natural Language Inference (NLI) and question generation-and-answering-based metrics for faithfulness evaluation as they performed the best in the study.

1 Introduction

The fast development of NLP models and their entry into our everyday life requires researchers and engineers to look for effective metrics to evaluate generated outputs. Factual consistency is one of the important aspects to be assessed. The authors of TRUE give examples of poor factual consistency of different NLP models and emphasize that despite the existence of various factual consistency evaluation metrics, there is no clear standardized protocol to evaluate the quality of these metrics making it difficult to compare them across different tasks and datasets. Thus TRUE aims to develop such a protocol and find the best metrics for the faithfulness evaluation of NLP models.

Since evaluating text generations is rather challenging and the question ”What is the best metric for this task?” often appears in NLP research, I find the goal of the TRUE paper particularly important. Let me present more details of this study.

2 Methods

The authors chose 12 metrics in total from three groups: Model-based, NLI-based, and QG-QA-based (Question Generation and Question Answering). As a baseline, N-gram-based metrics and token-level F1 scores were used since they are known to have poor correlation with factual consistency.

To evaluate the metrics, authors chose a set of datasets used in such NLP tasks as summarization, dialogue generation, fact verification, and paraphrase detection manually annotated for factual consistency. This choice is adequate since it covers the most common tasks where factual agreement is essential. Here are some important elements of the TRUE approach:

- the authors consider the faithfulness of a text w.r.t its ground truth rather then real-world knowledge;
- all annotations were converted to binary labels which indicate if the whole generated text is factually consistent with its ground truth or not;

- the authors report the Receiver Operating Characteristic Area Under the Curve (ROC AUC) to evaluate the metrics.

Indeed, the first two aspects clearly define what text should be considered faithful, and the ROC AUC as an evaluation metric provides interpretability and eases comparisons.

To verify that the difference between the best and second best methods on each dataset is statistically significant, the authors performed bootstrap resampling and hypothesis testing with two *p*-values.

3 Results and analysis

The main finding of the paper is that NLI and QG-QA-based metrics, namely Q^2, ANLI (Adversarial NLI), and $\mathrm{SC}_{\mathrm{ZS}}$ (averaged maximum NLI scores for the pairs of document/summary sentences in summary consistency task) perform the best on chosen datasets. The authors also found that averaging these three metrics allows them to achieve even better scores.

From additional analysis, the researchers found that the performance of these three metrics decreases with the increase in the input length. The authors note that it was not expected for $\mathrm{SC}_{\mathrm{ZS}}$ as this metric was supposed to handle longer texts. Also, it was surprising to see lower but still considerably good ROC AUC values for Q^2 and ANLI metrics on longer inputs. However, the authors do not provide any ideas of possible reasons for this behavior nor claim that it was a result of using standardized datasets.

The paper shows experimental proof that model-based metrics perform better for larger models. Also, the authors manually analyzed the errors and found that failures of the three best metrics correspond to difficult examples with wrong labels or text with personal statements or subtle inconsistencies. Finally, the paper presents the analysis of cases in which at least one of the best three metrics failed, while the ensemble succeeded.

4 Conclusion

When new metrics and datasets appear fast, it is important to have a standardized procedure to evaluate the quality of those metrics to ensure their appropriate usage. The TRUE paper presents such a protocol and proves its efficiency for factual consistency evaluation. The authors performed all the necessary experiments and analyses and clearly described the results. Moreover, the code is publicly available [1]. I think that the main strength of the paper is the choice and thorough examination of a set of metrics and datasets. There are no pitfalls I could note however I would suggest adding more domain-specific datasets, for example, medical reports or papers to find out the best metric for biomedical NLP tasks.

References

[1] Or Honovich et al. “TRUE: Re-evaluating Factual Consistency Evaluation”. In: *NAACL* (2022). arXiv: 2204.04991v3.

[1] https://github.com/google-research/true