Evaluating Question Answering Evaluation

Anthony Chen
University of California, Irvine
anthony.chen@uci.edu

Gabriel Stanovsky
Allen Institute for Artificial Intelligence
gabis@allenai.org

Sameer Singh
University of California, Irvine
sameer@uci.edu

Matt Gardner
Allen Institute for Artificial Intelligence
mattg@allenai.org

Abstract
Current metrics for evaluating question answering (QA) datasets are based on n-gram matching, which have a number of known shortcomings. In this work, we examine the quality of current metrics by how well they correlate with human judgements across three diverse QA datasets. Our work indicates that current metrics do reasonably well in evaluating current datasets, but as QA datasets require more abstract generative answering, metrics that go beyond n-gram matching will be required.

1 Metrics
We assess the quality of five commonly used n-gram based metrics: BLEU-1, BLEU-4, METEOR, ROUGE-L, and F1. In addition to these metrics, we explore whether metrics that rely on distributional similarity can outperform their n-gram counterparts. To this end, we study sentence mover’s similarity (SMS), commonly used in summarization, and BERTScore, which was recently proposed for evaluating translation.

2 Datasets
We describe the datasets used in this study and provide examples in Table 1.

NarrativeQA is a generative QA dataset on books and movie scripts (Kocisky et al., 2017). While generative, the answers often have high overlap with words in the context. The official evaluation metrics of NarrativeQA are BLEU-1, BLEU-4, METEOR, and ROUGE-L.

SemEval-2018 Task 11 is a multiple-choice dataset which focuses on commonsense reasoning (Ostermann et al., 2018). We convert this into a generative QA dataset by using the correct answer as a generative target. We hypothesize that this results in a more difficult generative dataset to evaluate as a number of the answers in the SemEval dataset have no overlap with the question or context.

ROPES is a span-based QA dataset with questions that focus on cause-and-effect relationships (Lin et al., 2019). The official evaluation metric is F1. A unique characteristic of ROPES is that incorrect and correct answers often have some n-gram overlap. We hypothesize F1 will struggle to accurately assign scores because of this.

3 Models
We use the outputs of trained models as the data to be annotated. For NarrativeQA and SemEval, we train a multi-hop pointer generator model (Bauer et al., 2018). For ROPES, we finetune a BERT model (Devlin et al., 2019).

Southern California Machine Learning Symposium 2020, San Diego, California.
4 Results

After training, we extract 500, 500, and 300 model predictions from the validation sets of NarrativeQA, ROPES, and SemEval respectively. We filter out cases where the prediction matches the gold answer. Annotators are then asked to rate (from 1 to 5) how closely a prediction captures the same information as a gold answer. We compute the Spearman and Kendall correlation of the annotations to the scores assigned by metrics. Results are presented in Table 2.

4.1 Discussion

ROPES proves to be a challenging dataset for F1 to evaluate. This highlights the fact that while F1 is a reasonable metric for many span-based QA datasets, the types of questions and answers can influence how well it works in practice.

We discover with SemEval that existing metrics do considerably worse compared to NarrativeQA. This aligns with our hypothesis that more free-form generative QA datasets leads to a degradation in n-gram based metrics’ performance.

Finally, BERTScore and SMS fall behind the best metric for each dataset. This points to the fact that metrics that perform well for evaluating summarization and translation do not necessarily indicate success in evaluating question answering.

5 Conclusion

We present a study of existing metrics by comparing their correlation to human accuracy judgements on three QA datasets. We find that while existing metrics do reasonable on existing datasets, as generative QA datasets become more abstractive in nature, better metrics that go beyond n-gram matching will be required.
References


