Literature Review: BLEURT: Learning Robust Metrics for Text Generation (Sellam et al., 2020)

Céline Kalbermatten | 328519 | celine.kalbermatten@epfl.ch
chatbots-r-us

1 Summary

In natural language processing (NLP), evaluation metrics are used to evaluate the quality of text generated by models. The goal is to replace human judgement by automatic evaluation metrics that closely align with it. Traditional metrics such as Bilingual Evaluation Understudy (BLEU) and Recall-Oriented Understudy for Gisting Evaluation (ROUGE) are mostly based on comparing n-grams between the generated and the reference text. However, these metrics often struggle capturing the nuances of human judgment.

The paper presents an evaluation metric designed for text generation task, which is learned based on the Bidirectional Encoder Representations from Transformers (BERT) and which better approximates human judgement than the most popular metrics as BLEU and ROUGE as well as newer metrics like BERTScore. The metric is called BLEURT and uses both supervised and unsupervised data in order to predict human judgement of text quality. It performs better than the traditional metrics that often fail in nuanced evaluations requiring an understanding beyond lexical similarities.

BLEURT was developed involving a pre-training on synthetic data generated from Wikipedia. This helps the model to generalize across different types and quality of textual data. After the pre-training, a fine-tuning on human-labelled data from tasks such as the WMT Metrics Shared Task and the WebNLG dataset was performed. This fine-tuning step ensured that BLEURT aligns closely with human judgement.

The main innovation of BLEURT is its ability to train with limited labelled data by using a large volume of unlabelled data. By learning from the patterns present in both types of data, the model can generalize better and becomes more robust against data scarcity and biases.

The paper describes several experiments demonstrating that BLEURT performs better than other existing metrics across different datasets and languages. This indicates BLEURT's effectiveness in real-world applications.

In conclusion, the paper represents a significant step forward in the direction of better alignment between automated metrics and human perception of generated text quality. By integrating the power of BERT with innovative training techniques, BLEURT offers a more reliable and nuanced assessment of text generation quality than traditional metrics. BLEURT does not only improve alignment with human judgment but also increases the possibility to fine-tune and assess NLP models more effectively.

2 Strengths

The paper presents a unique application of both pre-training on synthetic data and fine-tuning on human-labelled data. Together, the two steps not only enhance BLEURT's adaptability and accuracy but also align with the author's claims of robustness across different textual inputs.

Another notable strength of the paper is its empirical rigor. The author's claim that BLEURT aligns more closely with human judgement than existing metrics is supported by a validation across multiple benchmarks as the WMT Metrics Shared Task and the WebNLG dataset.

Furthermore, the paper provides all its source code and models to the public. This does not only allow verification and extension of the author's work but also guarantees transparency.

In sum, the strengths of the paper lie in its methodological innovation in the field of text evaluation metrics, its rigorous empirical validation and its transparency and accessibility. The mentioned elements support the author's claims and increase the impact of the paper on the field of natural language generation evaluation.

3 Weaknesses

The paper highlights the model's robustness and superiority over existing metrics like BLEU and BERTScore through its innovative use of synthetic pre-training data. This approach is beneficial for overcoming the scarcity of labelled data. However, it may not fully capture the complexities and nuances of real-world linguistic variations.

The generalization of BLEURT is primarily supported by controlled experimental setups, which might not adequately reflect its performance in practical, diverse linguistic environments. Since the BLEURT metric depends on the BERT architecture, it also inherits its limitations such as potential biases and problems when dealing with domain-specific language. The dependency of BLEURT on BERT might decrease its effectiveness for tasks outside the well-represented domains in BERT's training data.

Additionally, the assertions of superiority could be supported by more comparative analysis across a wider range of language tasks and more challenging datasets that more closely resemble practical applications.

Moreover, the problem statement of the paper focuses on aligning automated metrics with human judgment. This could be expanded to address more detailed aspects of text quality, such as cultural nuances or the complexities of informal language, which are crucial in translation and text generation but often overlooked in automated metrics. Such expansions could help in providing a more comprehensive evaluation of BLEURT's effectiveness and highlight potential areas where it may not perform as expected. This nuanced approach would give a more balanced view of BLEURT's capabilities and limitations.

To conclude, it can be said that while BLEURT represents a significant advancement, its application might still face challenges in more complex or nuanced linguistic tasks.

References

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. Bleurt: Learning robust metrics for text generation.