PRECOG TASK

READING TASK

SWARANG JOSHI
2022114010

Introduction

- Existing metrics for evaluating text generation rely on surface-form similarities like n-gram overlap (BLEU) or have other limitations

- Paper proposes BERTSCORE, a new evaluation metric that computes semantic similarity using contextualized word embeddings from pretrained BERT models

- BERTSCORE addresses pitfalls of existing metrics by better handling paraphrases and capturing long-range dependencies

Working

- Use pretrained BERT to get contextual embeddings for tokens in reference and candidate sentences
- Compute pairwise cos similarity between token embeddings and use greedy to maximize similarity score.
- Compute precision and recall thru matching
- Take F1 of precision and recall as final BERTSCORE
- NOTE -Can optionally weight token matches by inverse document frequency (idf)

Results

- Evaluated on machine translation and image captioning using 363 systems
- BERTSCORE correlates better with human judgments than existing metrics like BLEU, METEOR, etc.
- Shows stronger system-level and segment-level correlations on many language pairs
- Outperforms task-specific metrics like SPICE on image captioning
- More robust than other metrics on adversarial paraphrase detection task

Thank you!