BERTSCORE: EVALUATING TEXT GENERATION WITH BERT

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, Yoav Artzi

Strengths Of Paper

1. Semantic Understanding & Contextual Matching
    a. Uses BERT's contextual embeddings to capture deep semantic relationships.
    b. Goes beyond surface-level word matching used in traditional metrics like BLEU/ROUGE.
    c. Can recognize synonyms, paraphrases, and semantically similar expression.
2. Robust Architecture & Design:
    a. Flexible token alignment through greedy matching
    b. Three complementary metrics (Precision, Recall, F1) for comprehensive evaluation
    c. Handles length variations and word order differences naturally
    d. Built on well-established transformer architecture with proven effectiveness
3. Strong Empirical Results & Practical Benefits:
    a. Better correlation with human judgments across multiple tasks
    b. Generalizes well across different languages and domains

Weakness Of Paper

1. Computational and Resource Intensity
    a. High GPU memory requirements.
    b. Significant processing time compared to traditional metrics.
    c. Impractical for real-time applications.
2. Technical Limitations and Reliability Issues:
    a. BERT's token limit restricts long text evaluation
    b. May miss critical factual errors
    c. Inconsistent handling of numerical values
3. Methodological Gaps:
    a. Limited evaluation across diverse text generation tasks
    b. Not model independent, uses BERT as a backend only

Improvements Of Paper

1. Domain adaptation capabilities to be able to use across domains like legal, medical domain. In addition to be able to adapt to code switch languages like Spanglish, Hinglish.
2. Multi-level or hierarchical evaluation for very long documents, for contextual similarity between documents. Be able to distinguish between two long documents if and when they are talking about Apple as a company or Apple as a farming.
3. Should be compatible with all the models.