T.C.

MARMARA UNIVERSITY

FACULTY OF ENGINEERING

DEPARTMENT OF COMPUTER ENGINEERING

CSE4078 Introduction to NLP

LLM SFT REPORT 2

May 30th, 2024

Group 8 Members

Alper Özdemir

Eren Başpınar

Emirhan Erdoğan

Faruk Akdemir

Şule Koca

Introduction

This report aims to examine the advanced evaluation metrics used for Large Language Models (LLMs). LLMs are designed to perform highly in natural language processing tasks, and various metrics are used to objectively measure their performance. This report will cover advanced evaluation metrics such as perplexity, BLEU, ROUGE, METEOR, BERTScore, MoverScore, and embedding-based similarity metrics.

Advanced Evaluation Metrics

BLEU (Bilingual Evaluation Understudy)

BLEU is a metric used for machine translation and text generation. It calculates the n-gram overlaps between the generated text and reference texts. A higher BLEU score indicates that the model's output is closer to the reference texts. BLEU is particularly suitable for short sentences and includes a penalty term to prevent excessively short sentences.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

ROUGE is a metric used for text summarization and text generation. ROUGE measures n-gram, word sequence, and word pair overlaps. It has different variants like ROUGE-1, ROUGE-2, and ROUGE-L. Higher ROUGE scores indicate that the model's output is similar to the reference texts.

METEOR (Metric for Evaluation of Translation with Explicit Ordering)

METEOR is a metric used for machine translation and considers synonyms and word stems. METEOR uses word alignments, stemming, and synonymy information to produce more flexible and human-aligned results. A higher METEOR score indicates that the model produces semantically accurate translations.

BERTScore

BERTScore evaluates the similarity between two texts using deep learning models like BERT. BERTScore captures the meaning of the texts better because it considers word embeddings and contextual meaning. A higher BERTScore indicates that the model produces semantically accurate and consistent texts.

FrugalScore

FrugalScore is an evaluation metric for natural language generation (NLG) tasks that balances accuracy with computational efficiency. By training a low-cost model to approximate the results of more resource-intensive metrics, FrugalScore retains most of the original performance while significantly reducing computational costs. This makes it ideal for large-scale or real-time applications where traditional metrics would be impractical. In our text summarization project, using FrugalScore allowed us to assess the quality of generated summaries efficiently and reliably.

Embedding-based Similarity Metrics

Embedding-based similarity metrics use models like GloVe, Word2Vec, or BERT to measure the similarity between word or sentence embeddings. These metrics assess the contextual and semantic similarity of texts. Higher embedding-based similarity scores indicate that the model produces semantically consistent texts.

Evaluation Results

Below are the performance results of six different models evaluated using ROUGE, BLEU, BERT, METEOR, QaEval, and FrugalScore metrics:

Model	ROUGE-1 Mean	ROUGE-2 Mean	ROUGE-L Mean	BLEU Mean	BERT Mean	METEOR Mean	QaEval Mean	FrugalScore Mean
Model 1	0.218	0.130	0.170	0.035	0.835	0.230	0.311	0.654
Model 2	0.220	0.133	0.172	0.036	0.840	0.236	0.317	0.659
Model 3	0.211	0.125	0.165	0.034	0.837	0.220	0.307	0.652
Model 4	0.252	0.139	0.196	0.035	0.879	0.212	0.292	0.715
Model 5	0.306	0.161	0.248	0.037	0.872	0.153	0.196	0.708
Model 6	0.259	0.124	0.212	0.021	0.853	0.107	0.142	0.692

Model	ROUGE-1 Std	ROUGE-2 Std	ROUGE-L Std	BLEU Std	BERT Std	METEOR Std	QaEval Std	FrugalScore Std
Model 1	0.132	0.113	0.115	0.059	0.159	0.200	0.386	0.136
Model 2	0.129	0.112	0.112	0.057	0.151	0.200	0.387	0.133
Model 3	0.131	0.111	0.113	0.057	0.140	0.200	0.385	0.148
Model 4	0.132	0.126	0.118	0.073	0.063	0.189	0.379	0.075
Model 5	0.159	0.154	0.152	0.098	0.052	0.168	0.321	0.068
Model 6	0.140	0.122	0.128	0.064	0.057	0.130	0.288	0.061

Conclusion

This report has examined the advanced metrics used to evaluate the performance of LLMs and presented the evaluation results of six different models. These metrics play a crucial role in the development and improvement of LLMs, providing robust tools to objectively assess model performance. Proper use of these metrics ensures that LLMs produce effective and reliable results in real-world applications.