G-Eval:

NLG Evaluation using GPT-4 with Better Human Alignmnet

Yang Liu Dan Iter Yichong Xu
Shuohang Wang Ruochen Xu Chenguang Zhu

Microsoft Azure AI
yaliu10@microsoft.com

EMNLP 2023

고경빈

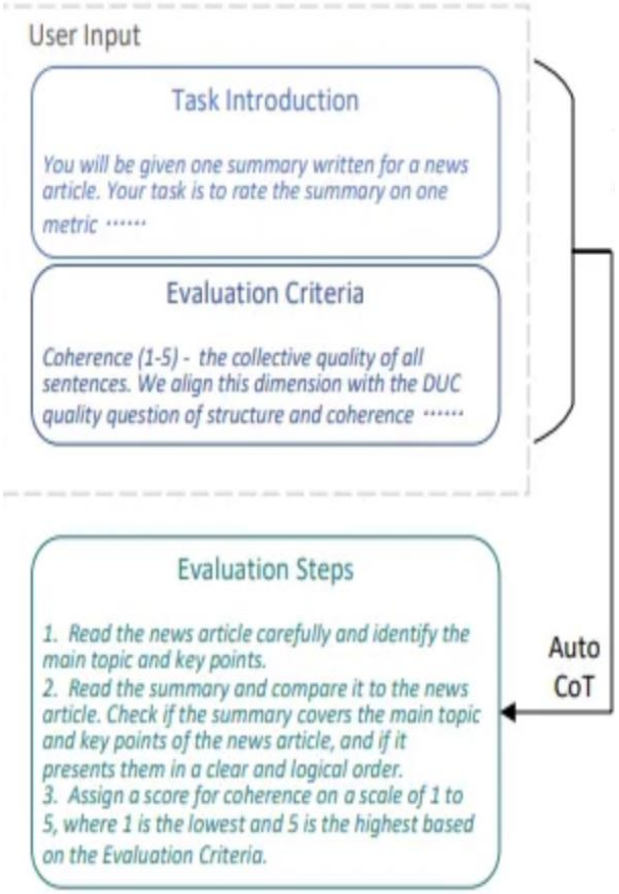
2025.01.10

Background

- Evaluating the quality of NLG systems is hard
- Traditional metrics have low correlation with human judgments
- Validity and reliability of reference-free NLG evaluators have not been investigated and still have low correlation
- A better framework is needed for using LLMs as evaluators

Method

- Prompt for NLG Evaluation
  - Define evaluation task and evaluation criteria
- Auto CoT for NLG Evalation
  - Generate evaluation step by itself
  - Provide more context and guidance
- Scoring Function
  - Perform evaluation task with form-filling paradigm
  - Use probability normalization $score = \sum_{i=1}^{n} p(s_i) \times s_i$



Experiments

- Model
  - GPT-3.5(text-davinci-003) with t=0 → G-EVAL-3.5
  - GPT-4 with n=20, t=1, top_p=1 → G-EVAL-4
- Benchmarks
  - SummEval(Summary): fluency, coherence, consistency, relevance
  - Topical-Chat(Dialogue generation):
    naturalness, coherence, engagingness, groundedness
  - QAGS(Hallucination): consistency

Experiments

- Baselines
  - BERTScore: cosine similarity based on BERT
  - MoverScore: BERTScore + soft alignment + EMD
  - BARTScore: average likelihood of BART
  - FactCC: a summary is consistent with the source document?
  - QAGS: generate questions from summary & check source for answer
  - USR: assess dialogue generation from different perspectives
  - UniEval: evaluate different aspects of text generation as QA tasks
  - GPTScore: formulate the evaluation task as a conditional generation problem

Correlation

- Pearson: linear relationship

$$r = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y}$$

- Spearman: non-linear and monotonic relationship

$$\gamma_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

- Kendall-Tau: non-linear & consistency of rank pairs

$$\tau = \frac{C - D}{C + D}$$

Correlation Coefficient Range	Interpretation
0.00 - 0.19	Very weak or no correlation
0.20 - 0.39	Weak correlation
0.40 - 0.59	Moderate correlation
0.60 - 0.79	Strong correlation
0.80 - 1.00	Very strong correlation

Results for Summarization

Metrics	Coherence		Consistency		Fluency		Relevance		AVG	
	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ
ROUGE-1	0.167	0.126	0.160	0.130	0.115	0.094	0.326	0.252	0.192	0.150
ROUGE-2	0.184	0.139	0.187	0.155	0.159	0.128	0.290	0.219	0.205	0.161
ROUGE-L	0.128	0.099	0.115	0.092	0.105	0.084	0.311	0.237	0.165	0.128
BERTScore	0.284	0.211	0.110	0.090	0.193	0.158	0.312	0.243	0.225	0.175
MOVERSscore	0.159	0.118	0.157	0.127	0.129	0.105	0.318	0.244	0.191	0.148
BARTScore	0.448	0.342	0.382	0.315	0.356	0.292	0.356	0.273	0.385	0.305
UniEval	0.575	0.442	0.446	0.371	0.449	0.371	0.426	0.325	0.474	0.377
GPTScore	0.434	–	0.449	–	0.403	–	0.381	–	0.417	–
G-EVAL-3.5	0.440	0.335	0.386	0.318	0.424	0.347	0.385	0.293	0.401	0.320
- Probs	0.359	*0.313*	0.361	*0.344*	0.339	*0.323*	0.327	*0.288*	0.346	*0.317*
G-EVAL-4	**0.582**	**0.457**	**0.507**	**0.425**	**0.506**	**0.455**	**0.547**	**0.433**	**0.514**	**0.418**
- Probs	0.560	*0.472*	0.501	*0.459*	0.505	*0.473*	0.511	*0.444*	0.502	*0.446*
- CoT	0.564	0.454	0.493	0.413	0.483	0.431	0.538	0.427	0.500	0.407
- Description	0.513	0.424	0.421	0.344	0.447	0.373	0.479	0.388	0.479	0.377

- GPT-based > Reference-free > Reference-based
- G-EVAL-4 > G-EVAL-3.5 → Larger model size is beneficial
- G-EVAL-4 > GPTScore → Form-filling paradigm is effective
- G-EVAL-4-with CoT > G-EVAL-4-without CoT
- G-EVAL-4-with probability > G-EVAL-4 without probability

Results for Dialogue Generation

Metrics	Naturalness		Coherence		Engagingness		Groundedness		AVG	
	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ
ROUGE-L	0.176	0.146	0.193	0.203	0.295	0.300	0.310	0.327	0.243	0.244
BLEU-4	0.180	0.175	0.131	0.235	0.232	0.316	0.213	0.310	0.189	0.259
METEOR	0.212	0.191	0.250	0.302	0.367	0.439	0.333	0.391	0.290	0.331
BERTScore	0.226	0.209	0.214	0.233	0.317	0.335	0.291	0.317	0.262	0.273
USR	0.337	0.325	0.416	0.377	0.456	0.465	0.222	0.447	0.358	0.403
UniEval	0.455	0.330	0.602	0.455	0.573	0.430	0.577	0.453	0.552	0.417
G-EVAL-3.5	0.532	0.539	0.519	0.544	**0.660**	**0.691**	**0.586**	0.567	0.574	0.585
G-EVAL-4	**0.549**	**0.565**	**0.594**	**0.605**	0.627	0.631	0.531	0.551	**0.575**	**0.588**

- G-EVAL > Reference-free > Reference-based
- G-EVAL-4 ~ G-EVAL-3.5 → Easy for G-EVAL

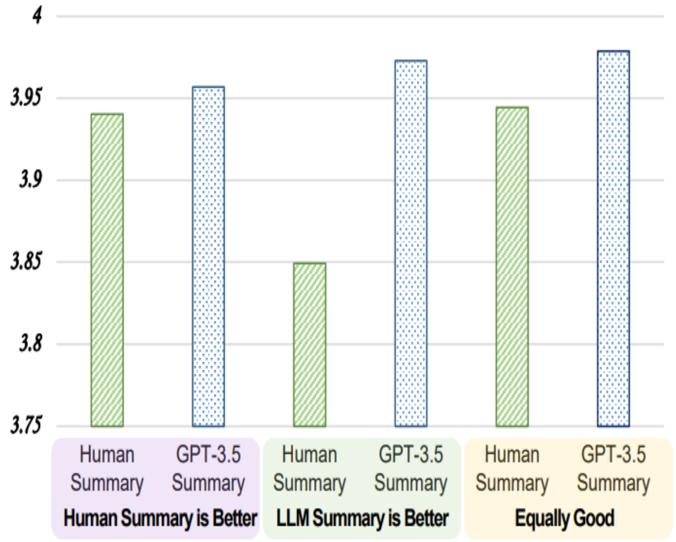
Results for Hallucination

Metrics	QAGS-CNN			QAGS-XSUM			Average		
	r	ρ	τ	r	ρ	τ	r	ρ	τ
ROUGE-2	0.459	0.418	0.333	0.097	0.083	0.068	0.278	0.250	0.200
ROUGE-L	0.357	0.324	0.254	0.024	-0.011	-0.009	0.190	0.156	0.122
BERTScore	0.576	0.505	0.399	0.024	0.008	0.006	0.300	0.256	0.202
MoverScore	0.414	0.347	0.271	0.054	0.044	0.036	0.234	0.195	0.153
FactCC	0.416	0.484	0.376	0.297	0.259	0.212	0.356	0.371	0.294
QAGS	0.545	-	-	0.175	-	-	0.375	-	-
BARTScore	**0.735**	0.680	0.557	0.184	0.159	0.130	0.459	0.420	0.343
CTC	0.619	0.564	0.450	0.309	0.295	0.242	0.464	0.430	0.346
UniEval	0.682	0.662	0.532	0.461	0.488	0.399	0.571	0.575	0.465
G-EVAL-3.5	0.477	0.516	0.410	0.211	0.406	0.343	0.344	0.461	0.377
G-EVAL-4	0.631	**0.685**	**0.591**	**0.558**	**0.537**	**0.472**	**0.599**	**0.611**	**0.525**

- G-EVAL-4 > UniEval > BARTScore
- G-EVAL-3.5 failed to perfrom well on this benchmark
    - Consistency is sensivity to the LLM's capacity

Will G-EVAL perfer LLM-based outputs?

- Dataset
  - Humman-written summaries > GPT-3.5 sumaries
  - Humman-written summaries < GPT-3.5 sumaries
  - Humman-written summaries ~ GPT-3.5 sumaries



- G-EVAL-4 assigns scores to human summaries reasonably
- G-EVAL-4 always gives higher scores to GPT-3.5's summaries
  - High-quality NLG outputs are hard to evaluate
  - G-EVAL may have a bias towards the LLM-generated texts

Limitations

- Bias towards the LLM-generated texts
  - Using evaluation scores as reward signals can cause LLMs to self-reinfo rce and overfit their evaluation criteria
- Limited by the availability and accessibility of LLMs
- LLM updates can cause inconsistent evaluations
- In a free-form paradigm, criteria need to be flexible and adaptive
- G-EVAL could be wrong

From Open Review

- Instability of GPT-4

	Coh. (ρ)	Coh. (τ)	Con. (ρ)	Con. (τ)	Flu. (ρ)	Flu. (τ)	Rel. (ρ)	Rel .(τ)
G-EVAL-4	0.582	0.457	0.507	0.425	0.455	0.378	0.547	0.433
G-EVAL-4(0613)	0.593	0.462	0.508	0.425	0.465	0.382	0.545	0.430

- Lack of baseline comparison

	Coh. (ρ)	Coh. (τ)	Con. (ρ)	Con. (τ)	Flu. (ρ)	Flu. (τ)	Rel. (ρ)	Rel .(τ)
G-EVAL-4	0.582	0.457	0.507	0.425	0.455	0.378	0.547	0.433
GPT-4(simple prompt)	0.513	0.424	0.421	0.344	0.447	0.373	0.479	0.388

- Human Evaluation from previous research
    - SummEval: 1~5, Topical-Chat: 1~3, Fact-based Benchmark: 0~1

Conclusion

- G-EVAL outperform other baseline metrics in terms of correlation with human judgments
- CoT can improve the performance of G-EVAL
- G-EVAL can give a detailed score with probability normalization
- G-EVAL has a potential issue of bias toward LLM-generated texts

My Review

- This is an important first study on the validity and reliability of using LLMs for NLG evaluation
- Using three correlation metrics enhances evaluation objectivity
- The paper addresses both the strengths and issues of G-EVAL
- No correlation threshold values were provided

Open Question

- Is it valid to use G-EVAL as an evaluator?

  If not, how can G-EVAL be improved?