元論文

G-EVAL: NLG Evaluation using GPT-4 with Better Human Alignment https://arxiv.org/pdf/2303.16634

作者のGitHub https://github.com/nlpyang/geval

G-EVAL: GPT-4を用いたNLG評価のための人間との整合性向上

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, Chenguang Zhu
Microsoft Cognitive Services Research
{yaliu10, iterdan, yicxu, shuowa, ruox, chezhu}@microsoft.com

要約

自然言語生成（NLG）システムによって生成されたテキストの品質を自動的に測定することは難しい。従来の参照ベースの指標、例えばBLEUやROUGEなどは、特に創造性や多様性が求められるタスクにおいて、ヒューマンジャッジメントとの相関が比較的低いことが示されている。最近の研究では、参照が不要なNLG評価のために、大規模言語モデル（LLM）を利用することが提案されており、これは新しいタスクに対しても適用可能であるという利点がある。しかし、これらのLLMベースの評価者は、中規模のニューラル評価者に比べて、依然として人間との対応が低い。本研究では、連鎖思考（CoT）とフォーム入力のパラダイムを用いて、LLMを活用してNLG出力の品質を評価するためのフレームワーク「G-EVAL」を提案する。我々は、テキスト要約と対話生成という2つの生成タスクで実験を行った。その結果、GPT-4をバックボーンモデルとするG-EVALは、要約タスクで人間とのスピアマン相関が0.514に達し、従来のすべての方法を大幅に上回ることが示された。また、LLMベースの評価者の行動に関する分析を行い、LLM生成テキストに対するバイアスの可能性についても強調する。

1. はじめに

自然言語生成システムの品質評価は、大規模言語モデルが高品質で多様なテキストを生成できるようになった現代でも、依然として難しい問題である。従来の自動評価指標、例えばBLEU、ROUGE、METEORなどは、NLG評価で広く使用されているが、新しいタスクに対して人間の評価との相関が低くなる傾向がある。さらに、これらの指標は参照出力が必要であり、新しいタスクに対してそれを収集するにはコストがかかる。

最近の研究では、LLMを参照なしでNLG評価者として直接使用することが提案されている。このアプローチでは、LLMが高品質で流暢なテキストに対して高い生成確率を割り当てるという前提に基づき、候補出力をスコアリングする。しかし、LLMをNLG評価者として使用する妥当性と信頼性は体系的に検討されておらず、これらの評価者が依然として中規模ニューラル評価者に比べて人間との対応が低いことが示されている。したがって、LLMをNLG評価に使用するためのより効果的で信頼性の高いフレームワークが求められている。

本研究では、LLMを使用し、連鎖思考（CoT）とフォーム入力のパラダイムを組み合わせたNLG出力の品質評価のためのフレームワーク「G-EVAL」を提案する。評価タスクの定義と評価基準を含むプロンプトをLLMに提供し、連鎖思考（CoT）を生成させて、評価タスクをステップバイステップで実行する。この評価者の出力は、フォーム形式でフォーマットされる。さらに、出力された評価トークンの確率を使用して、最終的なメトリックを洗練させることができる。我々は、テキスト要約と対話生成という2つのNLGタスクに関する3つのメタ評価ベンチマークで広範な実験を行った。その結果、G-EVALは人間の評価と高い相関を持つことが示され、既存のNLG評価者を大幅に上回った。また、LLMベースの評価者の行動に関する分析を行い、LLM生成テキストに対するバイアスの可能性についても強調する。

こちらの内容を日本語に翻訳し、マークダウン形式で提供します。

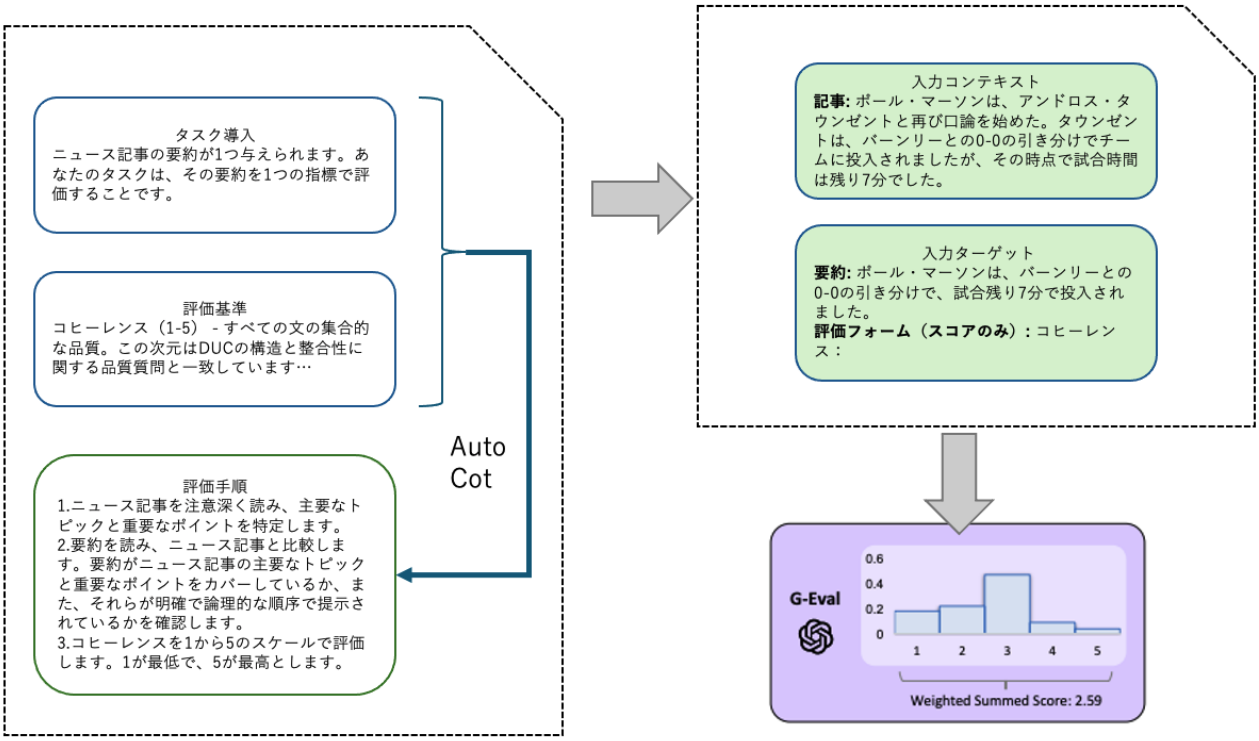

Figure 1 .1 この図は、G-EVALの全体的なフレームワークを示しています。最初にタスク導入と評価基準をLLMに入力し、詳細な評価手順を自動的に生成させます。その後、生成されたCoTを使用して、NLG出力をフォーム入力のパラダイムで評価します。最後に、出力スコアの確率加重総和を最終スコアとして使用します。

まとめると、この論文の主な貢献は以下の通りです：

1. **LLMベースのメトリック**は、特に対話応答生成のようなオープンエンドでクリエイティブなNLGタスクにおいて、人間の品質評価との相関に関して、参照ベースおよび参照不要のベースラインメトリックを一般的に上回ります。

2. **LLMベースのメトリック**は、指示やプロンプトに敏感であり、連鎖思考（CoT）を導入することで、より多くの文脈とガイダンスを提供し、LLMベースの評価者のパフォーマンスを向上させることができます。

3. **LLMベースのメトリック**は、個別のスコアをそれぞれのトークン確率に基づいて再重み付けすることで、より細かい連続スコアを提供することができます。

4. **LLMベースのメトリック**には、LLM生成テキストを人間が書いたテキストよりも好むという潜在的な問題があり、LLMベースのメトリックが自身の改善のための報酬信号として使用される場合、LLMの自己強化につながる可能性があります。

2. 方法

G-EVALは、3つの主要なコンポーネントを持つプロンプトベースの評価者である。1) 評価タスクの定義と評価基準を含むプロンプト、2) LLMによって生成された中間指示の連鎖思考（CoT）、3) LLMを呼び出してトークンの確率に基づいてスコアを計算するスコアリング機能。

NLG評価のためのプロンプト

プロンプトは、評価タスクと評価基準を定義する自然言語の指示である。例えば、テキスト要約の評価では、次のようなプロンプトを使用することができる：

> あなたには、ニュース記事の要約が1つ与えられます。あなたのタスクは、その要約を1つの指標で評価することです。これらの指示を注意深く読み、理解してください。レビュー中はこの文書を開いたままにし、必要に応じて参照してください。

プロンプトには、評価基準のカスタマイズも含まれ、異なるNLGタスクに応じて異なる基準が設定される。例えば、テキスト要約のコヒーレンス（整合性）の評価では、次の内容をプロンプトに追加する：

> コヒーレンス（1–5） – すべての文の集合的な品質。この次元はDUCの構造と整合性に関する品質質問と一致しており、「要約はよく構成され、整理されているべきである。要約は単に関連情報の山ではなく、トピックに関する情報の一貫した本体へと文から文へと築き上げられるべきである。」とされる。

自動生成された連鎖思考（CoT）によるNLG評価

連鎖思考（CoT）は、テキスト生成プロセス中にLLMによって生成された一連の中間表現である。評価タスクでは、一部の基準が単純な定義を超える詳細な評価指示を必要とし、各タスクごとにそのような評価手順を手動で設計するのは時間がかかる。我々は、LLMがそのような評価手順を自動的に生成できることを発見した。CoTは、LLMが生成されたテキストを評価するためのより多くの文脈とガイダンスを提供し、評価プロセスと結果を説明するのにも役立つ。例えば、テキスト要約のコヒーレンスを評価する場合、プロンプトに「評価手順：」という行を追加し、LLMに次のようなCoTを自動生成させる：

1. ニュース記事を注意深く読み、主要なトピックと重要なポイントを特定する。
2. 要約を読み、ニュース記事と比較する。要約がニュース記事の主要なトピックと重要なポイントをカバーしているか、また、それらが明確で論理的な順序で提示されているかを確認する。
3. コヒーレンスを1から5のスケールで評価する。1が最低で、5が最高とする。

> 訳者注意：
> 筆者のGitHubを見てもこの作成プロンプトは存在しない。
> おそらく、以下のような文章を与えて評価手順を作成させるという意図のように見える
>
> ```
> You will be given one summary written for a news article.
>
> Your task is to rate the summary on one metric.
>
> Please make sure you read and understand these instructions carefully.
> Please keep this document open while reviewing, and refer to it as needed.
>
> Evaluation Criteria:
>
> Coherence (1–5) – the collective quality of all sentences. We align this
> dimension with the DUC quality question of structure and coherence whereby
> "the summary should be well–structured and well–organized. The summary
> should not just be a heap of related information, but should build from
> sentence to a coherent body of information about a topic."
> ```

```
Evaluation Steps:


ここで評価基準が作成されるのでこれを後続のプロンプトで使う。
これをAuto-COTと言っていいのかが謎。
```

スコアリング機能

スコアリング機能は、設計されたプロンプト、自動生成されたCoT、入力文脈、および評価対象のテキストを使用してLLMを呼び出し、評価タスクを実行する。GPTScore（Fu et al., 2023）がターゲットテキストの生成確率を評価指標として使用するのとは異なり、G-EVALはフォーム入力パラダイムで直接評価タスクを実行する。例えば、テキスト要約のコヒーレンスを評価するために、プロンプト、CoT、ニュース記事、要約を連結し、定義された基準に基づいて各評価側面のスコアを1から5までででLLMに出力させる。

しかし、我々はこの直接的なスコアリング機能に2つの問題があることを発見した：

1. 一部の評価タスクでは、1つの数字がスコア分布を支配することが多く、例えば1から5のスケールで3が多くなる。このため、スコアの分散が低くなり、ヒューマンジャッジメントとの相関が低くなる可能性がある。
2. LLMは通常、プロンプトで明示的に小数値を要求しても、整数スコアしか出力しない。これにより、生成されたテキスト間の微妙な違いを捉えない評価スコアが多数発生する。

これらの問題に対処するために、LLMからの出力トークンの確率を使用してスコアを正規化し、それらの重み付き総和を最終結果として提案します。具体的には、プロンプトで定義されたスコアのセット $S = {s_1, s_2, \ldots, s_n}$ が与えられた場合、各スコア $ p(s_i) $ の確率はLLMによって計算され、最終的なスコアは次のようになります：

$$ \text{score} = \sum_{i=1}^{n} p(s_i) \times s_i \tag{1} $$

この方法により、生成されたテキストの品質と多様性をよりよく反映する、より細かい連続スコアを得ることができます。

```
訳者注：
ChatGPTのAPIでn=20を与えて20個結果を取得する。
https://github.com/nlpyang/geval/blob/main/gpt4_eval.py#L41


スコアリングについては以下のロジックで単純な平均をとっているように見える。
https://github.com/nlpyang/geval/blob/main/meta_eval_summeval.py#L57


これ、chatgpt呼び出す際にlogprobsを指定して、その確率の重みづけをするつもりだったんだろうけど実際のコードはされていないように見える。
```

3. 実験

Zhong et al.（2022）に従い、3つのベンチマーク、SummEval、Topical-Chat、QAGSで評価を行い、ヒューマンジャッジメントとの相関を測定した。

3.1 実装の詳細

OpenAIのGPTファミリーをLLMとして使用し、GPT-3.5（text-davinci-003）およびGPT-4を使用する。GPT-3.5では、モデルの決定性を高めるためにデコード温度を0に設定した。GPT-4では、トークン確率の出力をサポートしていないため、'n = 20, temperature = 1, top p = 1'を設定し、20回サンプリングしてトークン確率を推定する。G-EVAL-4は、GPT-4をバックボーンモデルとするG-EVALを示し、G-EVAL-3.5はGPT-3.5をバックボーンモデルとするG-EVALを示す。各タスクのプロンプト例は付録に記載している。

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**	**0.455**	**0.378**	**0.547**	**0.433**	**0.514**	**0.418**
- Probs	0.560	*0.472*	0.501	*0.459*	0.438	*0.408*	0.511	*0.444*	0.502	*0.446*
- CoT	0.564	0.454	0.493	0.413	0.403	0.334	0.538	0.427	0.500	0.407

表1: SummEvalベンチマークでの異なるメトリックのサマリーレベルでのSpearman（ρ）およびKendall-Tau（τ）の相関。G-EVALは、確率を用いない場合（斜体）は、他のメトリックとの公正な比較とは見なされるべきではありません。これは、スコアに多くの同点が生じるためです。このため、Kendall-Tauの相関が高くなりますが、真の評価能力を正確に反映しているわけではありません。詳細はセクション4を参照してください。

3.2 ベンチマーク

3つのメタ評価ベンチマークを採用し、G-EVALとヒューマンジャッジメントとの相関を測定する。

- **SummEval** (Fabbri et al., 2021)
  SummEvalは、さまざまな評価方法を比較するためのベンチマークであり、各要約の流暢さ、コヒーレンス、一貫性、関連性についての人間による評価を提供する。これは、CNN/DailyMailデータセットに基づいている。

- **Topical-Chat** (Mehri and Eskenazi, 2020)
  知識を用いた対話応答生成システムの評価者をメタ評価するためのテストベッドである。Zhong et al.（2022）に従い、自然さ、コヒーレンス、魅力的であること、根拠に基づくことの4つの側面で人間の評価を使用する。

- **QAGS** (Wang et al., 2020)
  要約タスクにおける幻覚の評価のためのベンチマークであり、2つの異なる要約データセットに対する一貫性の次元を測定する。

3.3 ベースライン

G-EVALを、最先端の性能を達成したさまざまな評価者と比較する。

- **BERTScore** (Zhang et al., 2019)
  BERTの文脈化された埋め込みを基に、2つのテキスト間の類似性を測定する。

- **MoverScore** (Zhao et al., 2019)
  BERTScoreにソフトアライメントと新しい集計方法を追加し、より堅牢な類似性尺度を得る。

- **BARTScore** (Yuan et al., 2021)
  事前学習されたエンコーダ・デコーダモデルであるBARTの平均的な尤度で評価する統一された評価者。

- **FactCC** および **QAGS** (Kryściński et al., 2020; Wang et al., 2020)
  生成された要約の事実的一貫性を測定する2つの評価者。FactCCは、元の文書に対する要約の一貫性を予測するBERTベースの分類器。QAGSは、要約から質問を生成し、元の文書に回答が見つかるかどうかを確認する質問応答ベースの評価者。

- **USR** (Mehri and Eskenazi, 2020)
  対話応答生成のさまざまな側面を評価する評価者。

- **UniEval** (Zhong et al., 2022)
  テキスト生成のさまざまな側面をQAタスクとして評価できる統一された評価者。

- **GPTScore** (Fu et al., 2023)
  GPT-3のような生成事前学習モデルを使用してテキストを評価する新しいフレームワーク。G-EVALとは異なり、GPTScoreは評価タスクを条件生成問題として定式化する。

3.4 要約の結果

Zhong et al.（2022）に従い、summary-level SpearmanおよびKendall-Tauの相関を使用して異なる要約メトリックを評価した。表1は、参照テキストとの意味的な類似性を比較するメトリックの結果を示しており、これらのメトリックはほとんどの側面で低い性能を示した。2番目の部分は、要約の品質に対する人間の評価から学習するニューラルネットワークを使用するメトリックの結果を示し、これらのメトリックは、類似性ベースのメトリックよりもはるかに高い相関を示している。最後の部分は、GPTベースの評価者に対応しており、G-EVALは、SummEvalベンチマークで従来の評価者を大幅に上回っている。GPT-4を使用したG-EVALは、SpearmanおよびKendall-Tauの相関で従来の評価者を大幅に上回り、GPT-4の大規模なモデルサイズが要約評価に有利であることを示している。

3.5 対話生成の結果

Topical-Chatベンチマークを使用して、対話応答の品質に対する人間の評価と評価者の一致を測定する。各対話のターンごとにPearsonおよびSpearmanの相関を計算した。表2は、類似性ベースのメトリックが、対話の魅力や根拠に基づく側面で人間とよく一致するが、他の側面ではそうではないことを示している。学習ベースの評価者に関しては、G-EVALがUniEvalと同等の結果を達成している。特に、G-EVAL-3.5はG-EVAL-4と同等の結果を達成しており、このベンチマークがG-EVALモデルにとって比較的容易であることを示している。

3.6 幻覚の結果

先進的なNLGモデルは、しばしば入力された文脈に合致しないテキストを生成することがある（Cao et al., 2018）。最近の研究では、強力なLLMでもこの幻覚問題に悩まされることが明らかになっている。これが、要約タスクにおける一貫性の側面を測定するための評価者の設計の動機となっている。我々は、CNN/DailyMailとXSumという2つの異なる要約データセットを含むQAGSメタ評価ベンチマークをテストした。表3は、BARTScoreがより抽出

的なサブセット（QAGS-CNN）で高い相関を示し、より抽象的なサブセット（QAGS-Xsum）では低い相関を示すことを示している。UniEvalは、両方のデータセットで良好な相関を示している。G-EVAL-4は、QAGS全体で最先端の評価者を上回り、QAGS-Xsumで大きな差を示している。一方、G-EVAL-3.5はこのベンチマークでうまく機能せず、一貫性の側面がLLMの能力に敏感であることを示している。

4. 分析

LLMベースの出力をG-EVALが好むのか？LLMを評価者として使用する際の懸念の1つは、LLM自身が生成した出力を人間が書いた高品質のテキストよりも好む可能性があるということである。この問題を調査するために、我々は要約タスクに関する実験を行い、LLM生成要約と人間が書いた要約の評価スコアを比較した。Zhang et al.（2023）によって収集されたデータセットを使用し、フリーランスのライターにニュース記事の高品質の要約を書かせ、次にアノテータに人間の書いた要約とLLM生成要約（GPT-3.5、text-davinci-003）を比較させた。このデータセットは、次の3つのカテゴリに分類できる：

1. 人間の書いた要約がGPT-3.5の要約よりも高い評価を受けた場合
2. 人間の書いた要約がGPT-3.5の要約よりも低い評価を受けた場合
3. 人間の書いた要約とGPT-3.5の要約が同等の良さと評価された場合

我々は、G-EVAL-4を使用して各カテゴリの要約を評価し、平均スコアを比較した。結果は図2に示されており、G-EVAL-4は、人間の評価者が人間の書いた要約を好む場合に、人間の書いた要約に高いスコアを割り当て、人間の評価者がGPT-3.5の要約を好む場合に、GPT-3.5の要約に低いスコアを割り当てる。しかし、G-EVAL-4は、人間の評価者が人間の書いた要約を好む場合でも、GPT-3.5の要約に対して常に高いスコアを与えている。この現象の潜在的な理由として、次の2つが考えられる：

1. 高品質なシステムからのNLG出力は、自然に評価が難しい。元の論文の著者は、人間が書いた要約とLLM生成要約を評価する際のアノテータ間の合意が非常に低いことを発見し、Krippendorffのアルファは0.07であった。

2. G-EVALは、評価と生成の両方でモデルが同じ評価基準の概念を共有している可能性があるため、LLM生成要約に対するバイアスを持つ可能性がある。

我々の研究はこの問題に関する予備的な研究として考慮されるべきであり、LLMベースの評価者の行動を完全に理解し、その固有のバイアスを軽減するためには、さらなる研究が必要である。我々は、LLMベースの評価者が評価スコアを報酬信号として使用して自身を改善する場合に、LLMが自身の評価基準に対して過剰適合し、本来のNLGタスクの評価基準に対して適合しなくなる可能性があることを強調する。

5. 関連研究

Ngramベースのメトリック

Ngramベースのメトリックは、生成されたテキストと参照テキスト間の語彙的な重複を測定することでNLGモデルを評価するスコアである。BLEU（Papineni et al., 2002）は、機械翻訳評価のために最も広く使用されている指標であり、修正されたn-gram精度の幾何平均とブレヴィティペナルティを計算する。ROUGE（Lin, 2004）は、要約評価のための再帰指向の指標であり、生成された要約と参照要約セット間のn-gramの重複を測定する。最近の研究では、NLGの60%以上の論文がROUGEまたはBLEUのみに依存してシステムを評価していることが示されている（Kasai et al., 2021）。しかし、これらの指標はコンテンツの質を測定することができず、NLGシステムの信頼性を正確に反映していない。

埋め込みベースのメトリック

埋め込みベースのメトリックは、生成されたテキストと参照テキスト間の意味的な類似性を、単語や文の埋め込みに基づいて測定するスコアである。WMD（Kusner et al., 2015）は、単語の埋め込みに基づいて2つのテキスト間の距離を測定するメトリックである。BERTScore（Zhang et al., 2019）は、BERTの文脈化された埋め込みを基に2つのテキスト間の類似性を測定する。MoverScore（Zhao et al., 2019）は、BERTScoreにソフトアライメントと新しい集計方法を追加し、より堅牢な類似性尺度を得る。

タスク特化の評価者

タスク特化の評価者は、特定のタスク要件に基づいて生成されたテキストの品質を測定するスコアである。例えば、要約タスクでは生成された要約の一貫性を評価し（Kryściński et al., 2020; Wang et al., 2020; Cao et al., 2020; Durmus et al., 2020）、対話応答生成タスクでは生成された応答のコヒーレンスを評価する（Dziri et al., 2019; Ye et al., 2021）。しかし、これらの指標は他のNLGタスクに一般化することができず、生成されたテキストの全体的な品質を測定することができない。

統一された評価者

最近、入力および出力内容を変えることで、複数の次元からテキストの品質を評価する評価者が開発されている（Yuan et al., 2021）。UniEval（Zhong et al., 2022）は、QAタスクとしてテキスト生成のさまざまな側面を評価できる統一された評価者であり、質問の形式を変更することで異なる評価タスクに対応できる。

LLMベースの評価者

Fu et al.（2023）は、GPTScoreという新しいフレームワークを提案し、GPT-3のような生成事前学習モデルを使用してテキストを評価することを提案している。これは、生成事前学習モデルが指示と文脈に従って生成された高品質なテキストに高い確率を割り当てるという前提に基づいている。Wang et al.（2023）は、ChatGPTをNLG評価者として使用することに関する予備的な調査を行っている。KocmiとFedermann（2023）は、機械翻訳タスクの評価にGPTモデルを使用することを提案している。

6. 結論

本研究では、連鎖思考（CoT）を用いて生成されたテキストの品質を評価するためのフレームワーク「G-EVAL」を提案する。我々は、テキスト要約と対話生成という2つのNLGタスクに関して広範な実験を行い、G-EVALが最先端の評価者を上回り、高い人間対応性を達成できることを示した。また、LLMベースの評価者の行動に関する予備的な分析を行い、LLM生成テキストに対するバイアスの可能性についても強調した。我々の研究が、NLG評価にLLMを使用するためのさらなる研究を促し、また、評価者としてLLMを使用する際の潜在的なリスクと課題についての意識を高めることを期待する。

参考文献

- Banerjee

, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, 65–72.

- Cao, M., Dong, Y., Wu, J., & Cheung, J. C. K. (2020). Factual error correction for abstractive summarization models. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 6251–6258.

- Clark, E., Celikyilmaz, A., & Smith, N. A. (2019). Sentence mover's similarity: Automatic evaluation for multi-sentence texts. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2748–2760.
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 4171–4186.
- Durmus, E., He, H., & Diab, M. (2020). FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 5055–5070.
- Dziri, N., Kamalloo, E., Mathewson, K., & Zaiane, O. R. (2019). Evaluating coherence in dialogue systems using entailment. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 3806–3812.
- Fabbri, A. R., Kryściński, W., McCann, B., Xiong, C., Socher, R., & Radev, D. (2021). SummEval: Re-evaluating summarization evaluation. Transactions of the Association for Computational Linguistics, 9, 391–409.
- Fu, J., Ng, S. K., Jiang, Z., & Liu, P. (2023). GPTScore: Evaluate as you desire. arXiv preprint arXiv:2302.04166.
- Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., & Blunsom, P. (2015). Teaching machines to read and comprehend. Advances in Neural Information Processing Systems, 28.
- Kasai, J., Sakaguchi, K., Le Bras, R., Dunagan, L., Morrison, J., Fabbri, A. R., Choi, Y., & Smith, N. A. (2021). Bidimensional leaderboards: Generate and evaluate language hand in hand. arXiv preprint arXiv:2112.04139.
- Kocmi, T., & Federmann, C. (2023). Large language models are state-of-the-art evaluators of translation quality. arXiv preprint arXiv:2302.14520.
- Kryściński, W., McCann, B., Xiong, C., & Socher, R. (2020). Evaluating the factual consistency of abstractive text summarization. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 9332–9346.
- Kusner, M., Sun, Y., Kolkin, N., & Weinberger, K. (2015). From word embeddings to document distances. International Conference on Machine Learning, 957–966.
- Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 7871–7880.
- Lin, C. Y. (2004). ROUGE: A package for automatic evaluation of summaries. Text Summarization Branches Out, 74–81.
- Mehri, S., & Eskenazi, M. (2020). USR: An unsupervised and reference-free evaluation metric for dialog generation. arXiv preprint arXiv:2005.00456.
- Narayan, S., Cohen, S. B., & Lapata, M. (2018). Don't give me the details, just the summary! Topic-aware convolutional neural networks for extreme summarization. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 1797–1807.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730–27744.

- Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: A method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 311–318.
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21, 1–67.
- Reiter, E., & Belz, A. (2009). An investigation into the validity of some metrics for automatically evaluating natural language generation systems. Computational Linguistics, 35(4), 529–558.
- Stent, A., Marge, M., & Singhai, M. (2005). Evaluating evaluation methods for generation in the presence of variation. Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing, 341–351.
- Wang, A., Cho, K., & Lewis, M. (2020). Asking and answering questions to evaluate the factual consistency of summaries. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 5008–5020.
- Wang, J., Liang, Y., Meng, F., Shi, H., Li, Z., Xu, J., Qu, J., & Zhou, J. (2023). Is ChatGPT a good NLG evaluator? A preliminary study. arXiv preprint arXiv:2303.04048.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q., & Zhou, D. (2022). Chain of thought prompting elicits reasoning in large language models. arXiv preprint arXiv:2201.11903.
- Ye, Z., Lu, L., Huang, L., Lin, L., & Liang, X. (2021). Towards quantifiable dialogue coherence evaluation. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2718–2729.
- Yuan, W., Neubig, G., & Liu, P. (2021). BARTScore: Evaluating generated text as text generation. Advances in Neural Information Processing Systems, 34.
- Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). BERTScore: Evaluating text generation with BERT. arXiv preprint arXiv:1904.09675.
- Zhang, T., Ladhak, F., Durmus, E., Liang, P., McKeown, K., & Hashimoto, T. B. (2023). Benchmarking large language models for news summarization.
- Zhao, W., Peyrard, M., Liu, F., Gao, Y., Meyer, C. M., & Eger, S. (2019). MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 563–578.
- Zhong, M., Liu, Y., Yin, D., Mao, Y., Jiao, Y., Liu, P., Zhu, C., Ji, H., & Han, J. (2022). Towards a unified multi-dimensional evaluator for text generation. arXiv preprint arXiv:2210.07197.