

機械翻訳の品質を迅速かつ包括的に評価する新手法 CATER (Comprehensive AI-assisted Translation Error Rating) の提案

飯田蔵土/Kurando IIDA, 株式会社ErudAite CEO

イントロダクション

CATER (Comprehensive AI-assisted Translation Error Rating) は、機械翻訳の品質評価における課題を解決するために、私、飯田蔵土がErudAite翻訳エンジンの実務開発過程で考案した評価フレームワークである。翻訳評価に長年使用されてきたBLEUやTERには、以下のような実務的な課題が存在していた:

1. 参照訳の準備が高コスト

BLEUやTERでは必ず参照訳が必要であり、低リソース言語や多様な翻訳タスクにおいて、評価準備が実現不可能な場合が多い。

2. 人間翻訳の評価が困難

従来の指標では、参照訳とのN-gram一致率を基にスコアが算出されるため、多様性に富んだ人間翻訳を評価することが難しい。

3. 文脈や意味を反映しにくい

単純な一致や編集量に基づく評価では、文脈の一貫性や意味の適切性を十分に反映できない。

CATERはこれらの課題を解決するため、**TERの基本概念 (Edit率の計算) をLLMを活用する形で拡張したものである**。これにより、参照訳不要で、実務的な翻訳改善に直結する評価フレームワークを実現した。

CATERの特徴

1. 実務に基づいた設計/エラー特定の容易さ

CATERは、ErudAiteの開発過程での翻訳評価と改善プロセスから生まれた実用的な評価手法である。そのため、翻訳者や翻訳エンジン開発者が、**エラー箇所を正確に把握**し、効率的に改善することを支援する。

2. LLMを活用した評価

原文と訳文をLLMに与えることで、翻訳エラーを5つのカテゴリ (文法、意味、文脈、スタイル、情報完全性) に分類し、具体的に特定する。**参照訳が不要**であるため、BLEUやTERと比較して評価準備のコストが低い。

3. 適用範囲の広さ

人間の手による参考訳への「近さ」によって翻訳を評価する場合、人間の翻訳自体を評価することは困難であった。**CATERは人間翻訳、機械翻訳を無差別に扱う**ことが可能である。

CATERプロンプト例と入出力例

評価方法とスコア算出

CATERは以下の計算式で翻訳の品質を評価する。

編集率 (Edit Ratio, ER%)

$ER\% = \frac{\text{修正が必要な単語数}}{\text{総単語数}} \times 100$

スコア算出式

$\text{スコア} = (1 - ER\% \times \text{重み}) \times 100$

各カテゴリのスコアを加重平均し、最終的な総合スコアを算出する。この手法により、エラーの量的評価と質的重要度の両方を反映可能である。

CATERの実装は、LLMを活用したシンプルなプロンプト設計が特徴である。以下はCATERで使用される評価プロンプトの一例 (抜粋) である。

System Instructions:
You are an advanced translation evaluation assistant. Your task is to evaluate the provided translation against the source text based on the following categories:
1. Linguistic Accuracy: Grammar, spelling, fluency.
2. Semantic Accuracy: Retention of original meaning.
3. Contextual Fit: Logical flow and coherence.
4. Stylistic Appropriateness: Tone and style suitability.
5. Information Completeness: Missing or unnecessary information.

Instructions:
For each category, identify errors, calculate the Edit Ratio (ER%), and provide a category score. Summarize the results as follows:
- Category Scores: [List each category and its score]
- Overall Score: [Provide the weighted average score]
- Error Locations and Reasons: [Identify specific errors with reasoning]

CATERのフレームワーク

評価の5つの軸

1. 言語的正確性 (Linguistic Accuracy)

文法、表記、流暢さを評価。

2. 意味の正確性 (Semantic Accuracy)

翻訳が原文の意味を正確に反映しているかを評価。

3. 文脈的適合性 (Contextual Fit)

翻訳の文脈的な一貫性や論理性を評価。

4. スタイルの適切性 (Stylistic Appropriateness)

読者や目的に応じた文体やトーンの適切性を評価。

5. 情報の完全性 (Information Completeness, 重み5)

情報の欠落や不要な追加 (ハルシネーション) を評価。

入力1/原文(川端康成『雪国』より)

国境の長いトンネルを抜けると雪国であった。夜の底が白くなった。信号所に汽車が止まった。
向側の座席から娘が立って来て、島村の前のガラス窓を落した。雪の冷気が流れこんだ。娘は窓いっぱいになり出して、遠くへ叫ぶように、
「駅長さん、駅長さん。」
明りをさげてゆっくり雪を踏んで来た男は、襟巻で鼻の上まで包み、耳に帽子の毛皮を垂れていた。

入力2/訳文(ErudAite β版 ver4.1による翻訳)

The train emerged from the long tunnel into the snow country. The depths of night had faded to white. The train came to a stop at a signal station.
A young woman rose from her seat across the aisle and lowered the window in front of Shimamura. A gust of cold, snowy air rushed in. Leaning out as far as she could, she called into the distance:
"Stationmaster! Oh, Stationmaster!"
A man approached slowly through the snow, carrying a lantern. His scarf was wrapped up to his nose, and the fur trim of his cap hung low over his ears.

出力/CATER (GPT4o, Temperature=0) ※抜粋

1. Category Scores
Linguistic Accuracy: 85
Edit Ratio (ER%): 15%
Errors Identified:
"The train emerged from the long tunnel into the snow country.": 「雪国であった」という原文の文末表現が「into the snow country」と訳されており、ニュアンスがやや異なる。原文の断定的な表現を反映する必要がある。
"The depths of night had faded to white.": 「夜の底が白くなった」という原文の詩的な表現が「faded to white」と訳されており、やや平坦な印象を与える。原文の詩的なニュアンスを強調するべき。
Semantic Accuracy: 80
Edit Ratio (ER%): 20%
Errors Identified:
"A young woman rose from her seat across the aisle and lowered the window in front of Shimamura.": 「向側の座席から娘が立って来て」という原文の「娘」が「a young woman」と訳されているが、原文の「娘」には特定の若さや親密さのニュアンスが含まれているため、より適切な訳語が必要。
"Leaning out as far as she could, she called into the distance.": 「遠くへ叫ぶように」という原文のニュアンスが「called into the distance」としてやや弱まっている。

CATERと従来指標の比較

項目	CATER	COMET	BLEU/TER
参照訳の必要性	不要	推奨されるが必須ではない	必須
文脈・意味の評価	独立カテゴリで評価可能	埋め込みベースで意味的類似性を評価	限定的
情報完全性の評価	欠落情報や不要な挿入を評価可能	間接的に評価	評価対象外
エラーの具体性	修正箇所やエラー分類を明示	スコア提示のみ	不明確

CATERの潜在的な課題と今後の展望

スコアの安定性向上

LLMのモデルや設定 (例: Temperature) によってスコアが変動する可能性がある。そこで、実装時には (1) Temperatureを下げて出力を安定させる、(2) 複数のLLMを利用して評価の範囲を取る、(3) 評価を複数回行い平均を取るといった工夫を検討することが可能であろう。

ツールとしての普及

システム化することが極めて容易であり、幅広く翻訳評価のツールとして普及させることが可能である。右のQRコードは、その一例として筆者がノーコードツールで実装したCATERである。

