第一回 AAMT 翻訳通訳研究会
2026年3月18日

翻訳品質の自動評価は何を評価対象としているのか？

萱野陽子[1,2] 菅原朔[1,2]
[1]総合研究大学院大学　[2]国立情報学研究所
yokokayano@nii.ac.jp saku@nii.ac.jp

問題提起

「翻訳品質の自動評価は何を評価対象としているのか？」

- 翻訳品質は、何を基準に判断されているのか
- 自動評価は、何を「正解」とみなしているのか
- その基準は、翻訳の目的が変わっても妥当なのか

質の高い翻訳とは？

📍本発表では、
この問いを**機械翻訳の評価史**と**翻訳理論**の両方から考える

NLPにおける機械翻訳評価の流れ

- **BLEU**（2002）
  参照訳との**表層的一致**を測る
- **chrF**など（2010年代）
  文字列一致を改良し，**表層一致**をより柔軟に測る

参照訳との一致

- **BLEURT / COMET**（2020）
  原文・参照訳との**意味的な一致**を測る
- **MQM / MetricX / LLM judge**（2021～）
  エラー分析を行い、人手判断に近づけて評価する

原文との一致

→ 主に「**原文との等価性**」をより正確に測ろうとしてきた

NLPの機械翻訳評価の前提

- 良い翻訳 ＝ **原文**と等価な翻訳
- エラー ＝ **原文**との不一致（意味・文体・形式など）
- 評価 ＝ **原文**との一致をより正確に測ること

評価方法に変化はあっても評価の前提にあるのは**原文との等価性**

翻訳学における翻訳品質の考え方

等価性理論

"Translating consists in **reproducing the closest natural equivalent** of the source-language message."— Nida and Taber[1](1969)

スコポス理論

"The dominant factor of every translation is **its purpose**."— Vermeer[2] (1989)

→ 翻訳品質は、原文との等価性だけでなく**目的にも依存する**

[1]Nida, Eugene A., and Charles R. Taber. *The Theory and Practice of Translation*.1969.

[2] Vermeer, Hans J. "Skopos and Commission in Translational Action." 1989

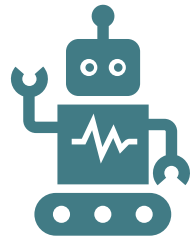
本研究の目的とアプローチ

本研究では以下を検討する：

- 翻訳は**目的**によって変わる
- しかし現在の自動評価は主に原文との等価性を評価している

- そこで本研究では、
  - 翻訳の目的を明示した翻訳生成
  - 翻訳の目的に沿った評価

  を行い、**自動評価が実際に何を評価しているのか**を検討する



自動評価では何が高く評価されるのか

- 上場企業時価総額上位33社のIR資料の英訳を対象とした
- COMETKiwi（原文と翻訳文を比較するQE指標）で評価した

Type	Official	Google	GPT	GPT+Sp	GPT PE+Sp
Mean	*0.783*	**0.830**	0.822	0.821	0.810
SD	0.043	0.031	0.039	0.033	0.037

- 平均スコアでは **Google翻訳が最高**
- **企業による公式翻訳が最低**
- 33社すべてで **Google > 公式翻訳**

GPT=GPT4

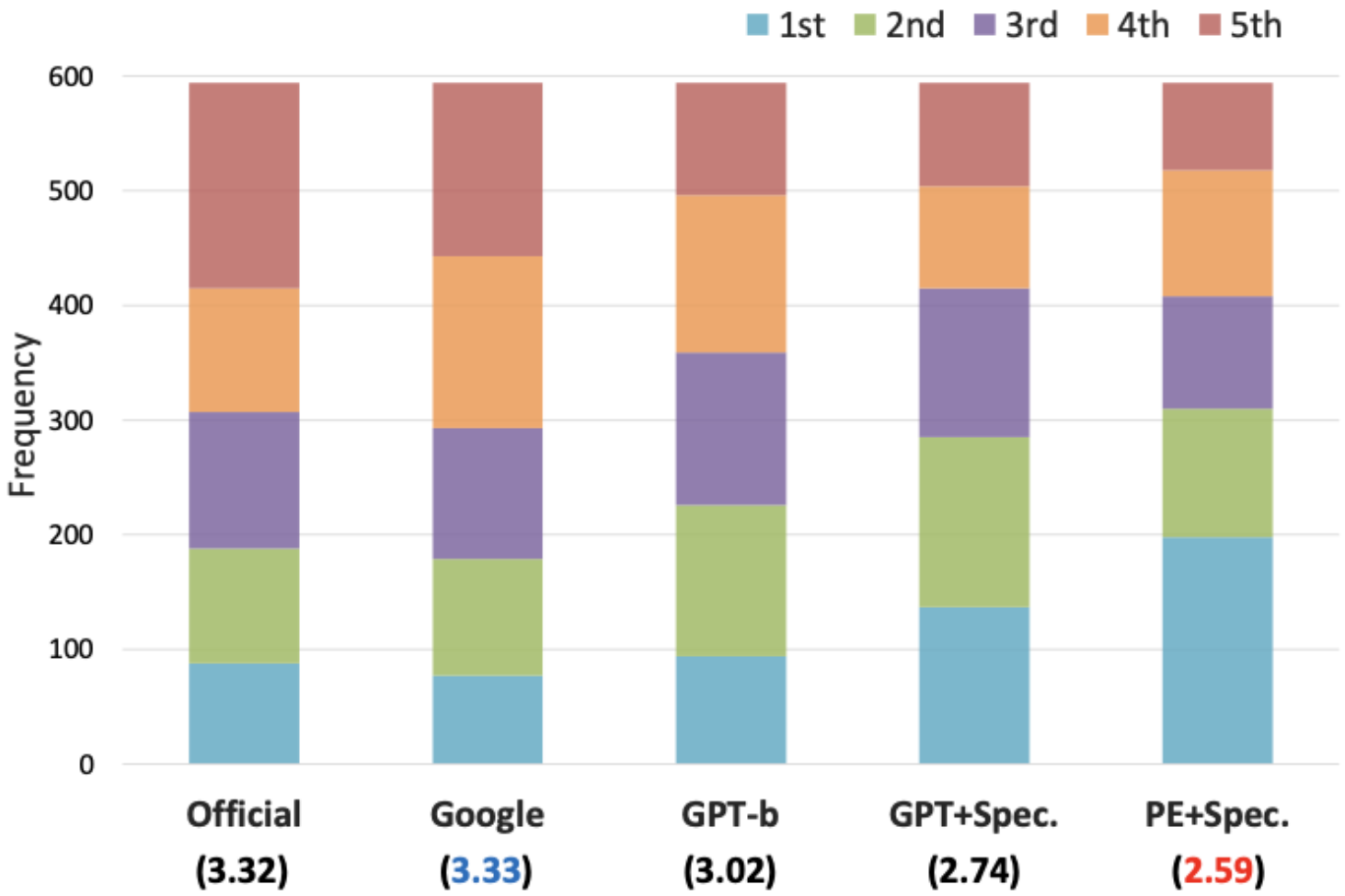
Sp= Specifications

→ この実験では、原文に近い対応関係を保つ訳ほど高く評価される可能性

なぜGoogle翻訳が高く評価されるのか：ANAの例

- **原文**
  “ワクワク” は、人を動かすエネルギー。
  それは人から人へと伝わり、世界をあかるく元気にする。（全日本空輸株式会社）
- **公式英訳** (COMET スコア：**0.70**)
  “Waku waku” is what moves people to push what’s possible. It’s Japanese for the joy and excitement of discovering the unknown. And when passed from person to person, becomes a force that creates a brighter world, united in wonder.
- **Google翻訳** (COMET スコア：**0.85**)
  “Excitement” is the energy that moves people. It spreads from person to person, making the world brighter and more energetic.

ただし人手評価の結果は異なる



括弧内は平均順位（低いほど良い）

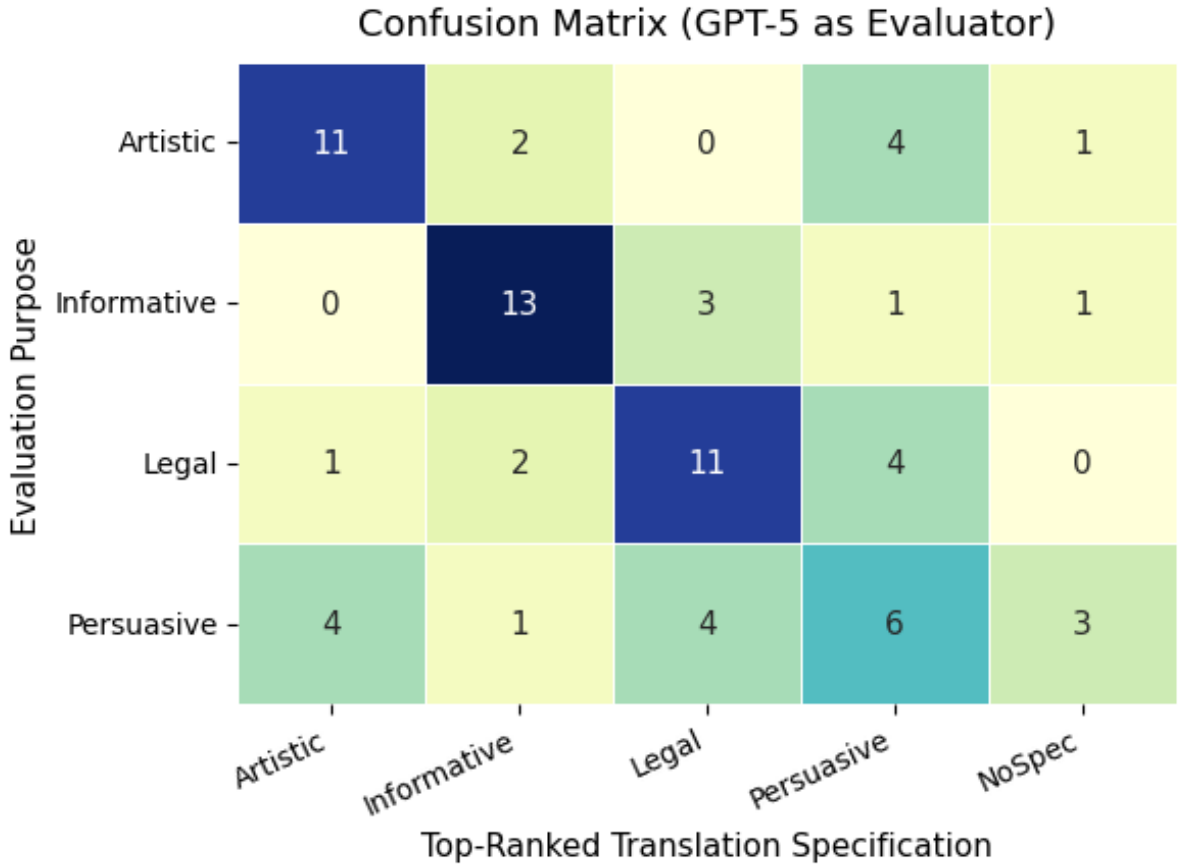
英語母語話者（会計学・金融学専攻、学士以上）17名と**翻訳者**1名による**主観評価**（いずれもProlificで募集）

- 最も高く評価されたのは、Specification（簡単な仕様）ありでGPTがポストエディットした翻訳（PE+Spec）

- Specification：IR資料として投資家などのステークホルダーに魅力的にうつる英語に翻訳する

大規模言語モデル

LLMは目的に沿った翻訳を高く評価できるか？



設定
1つの原文に対して**4つの目的別の翻訳**と**目的なし翻訳**を、5つのモデルで生成（**468翻訳**）

タスク
GPT-5 に 原文と**翻訳目的** を与え「どの翻訳が最も目的に沿った翻訳か」を選択させる

図の見方
行：**翻訳目的＝評価目的**
列：**最上位に選ばれた翻訳タイプ**
対角線上の数字が大きいほど、目的に合った訳を選べている

GPT-5 は、多くの場合、翻訳目的に対応する翻訳を最上位に選んだ

MQM(2013)が本来示していたこと

MQM: 機械翻訳研究で広く活用されている人手評価フレームワーク

“…translation quality can only be assessed in terms of whether or not a translation meets specified requirements and meets its communicative purpose.”ーLommel[3] (2013)

→翻訳品質は、要件と目的に照らして評価される

- NLPで広く使われたのは主にエラー分類のフレームワーク
- MQMの品質の考え方は自動評価では十分に主流化してこなかった

[3] Arle Richard Lommel, Aljoscha Burchardt, and Hans Uszkoreit. “Multidimensional quality metrics: a flexible system for assessing translation quality.” 2013.

自動評価の再検討

翻訳品質は、単一の固定的基準ではなく、**要件や目的に依存する**

したがって、自動評価も**何を質が高い翻訳とみなすか**から設計する必要がある

LLM によって、プロンプトなどを用いることで、**目的を明示した評価設計** の可能性が広がっている[4]

[4] Yamada, Masaru. Optimizing machine translation through prompt engineering: An investigation into ChatGPT's customizability. Proceedings of Machine Translation Summit XIX, Vol. 2: Users Track, 195-204. 2023.

現在取り組んでいる課題

- 評価する翻訳が定められた目的に沿っているかどうかをどのようにLLMに評価させ、妥当性を保証するのか
- 翻訳の目的を、大規模実験で扱える形にどのように設計するか
- 目的に沿った表現の変更と、普遍的な翻訳エラーをどのように区別して評価するか

📍こうした観点を取り入れた自動評価の枠組みを模索中

本発表の問題意識は以下の研究に基づく：

- Kayano, Yoko., & Sugawara, Saku. Specification-Aware Machine Translation and Evaluation for Purpose Alignment. WMT 2025.
- 萱野陽子・菅原朔. 「機械は仕様を考慮して翻訳できるか: 統合報告書の英訳の場合」 情報処理学会研究報告 NL-262. 2024.

ありがとうございました