同時通訳品質評価のための同時通訳者と翻訳者の評価比較分析

蒔苗 茉那*1, 須藤 克仁*1, 中村 哲*1, 松下 佳世*2, 山田 優*2
*1　奈良先端科学技術大学院大学
*2　立教大学

差異：同時通訳評価指標の欠如

	翻訳	同時通訳
人手評価	MQM[2]	なし
自動評価	BLEU, COMET	なし

同時通訳の自動評価に向けて
まずは人手評価の整備を行う

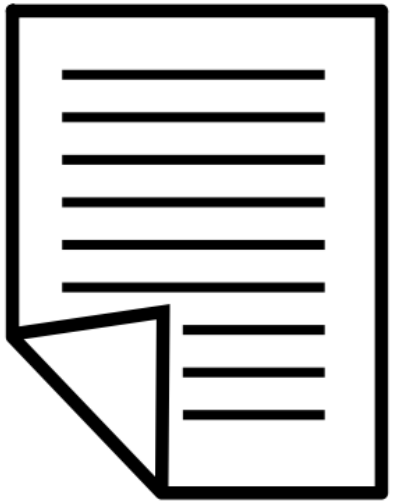
翻訳の評価

評価

- 原言語と目的言語が一対一の関係になっている

評価指標

- 人手評価：MQM[2]
  - マニュアルに基づいてエラーごとに「カテゴリ」と「重大度」で評価
- 自動評価：BLEU, COMET
  - BLEU：トークンレベルの一致数で評価
  - COMET：大規模言語モデルを用いる



同時通訳の評価

評価

- 原発話と目的発話のアライメントが取りにくい

評価指標

- 人手評価： NAATI[6], EU Metrics[7]
  - 同時通訳者の能力測定
  - 同時通訳の訳出結果の評価なし
- 自動評価：なし

動機：同時通訳の品質評価要件の明確化

同時通訳の質の安定

人間・機械による
通訳の質の改善

差異：同時通訳評価指標の欠如

	翻訳	同時通訳
人手評価	MQM[2]	なし
自動評価	BLEU, COMET	なし

同時通訳の自動評価に向けて
まずは人手評価の整備を行う

手法：同時通訳者と翻訳者の評価比較分析
動機：同時通訳の品質評価方法の検討

用いる指標：翻訳品質評価（JTF）ガイドライン[3]
- Multidimensional Quality Metrics (MQM)[2]を元に作成

MQM[2]とは
- 翻訳評価のためのフレームワーク
- 集計の段階で1文ずつ、エラーカテゴリと重大度を用いて体系的に評価
- WMTのShared Task[5]にて使用
  - 機械翻訳の性能評価タスク
  - MQMスコア、正解ラベルとして使用
  - MQMスコアと機械翻訳結果の相関関係を比較
    - 相関関係が高いほど、機械翻訳が優れている

手法：同時通訳者と翻訳者の評価比較分析
動機：同時通訳の品質評価方法の検討

MQMスコア計算手法

- エラー・カテゴリ設定の重み(依頼者が設定)　＊　重大度
- 1.33 = 1.33(正確さ「-誤訳-」)　＊　1 (軽度)

原発話 書き起こし	it's for the government to make sure that the private sector will be able to take part in the economic progress of our country in the right way within the rule of law.
同時通訳 書き起こし	ですから政府としては民間がきちっと経済の進歩に適切な形で法の支配のもとで関与できること参加できることを願っていると思います。
翻訳者評価 (エラー値)	「make sure」 を「願っていると思います。」 と訳すのは不適切 (1.33)

分析設定

- データ：日本記者クラブの記者会見の同時通訳収録データ(JNPC063)[9]

	同時通訳者	翻訳者
データの 提示方法	同時通訳の音声およびその書き起こし	原発話と同時通訳の書き起こしのみ
JTF ガイドライン[3] 使用経験	なし ＊事前にJTFガイドライン[3]に基づく評価方法の説明 ＊エラー重大度について通訳の観点で評価して良い旨を伝えた	あり

全体：同時通訳と翻訳の人手評価分析結果

共通点

MQMカテゴリ
「正確さ-誤訳」

↓

同時通訳にも
適用可

相違点

MQMカテゴリ
「正確さ-抜けと余分」
「流暢さ」

↓

同時通訳への適用不可
緩和する必要あり

結果分析　共通: カテゴリ「正確さ-誤訳」

原発話 書き起こし	it's for the government to make sure that the private sector will be able to take part in the economic progress of our country in the right way within the rule of law.
同時通訳 書き起こし	ですから政府としては民間がきちっと経済の進歩に適切な形で法の支配のもとで関与できること参加できることを願っていると思います。
同時通訳者 評価 (エラー値)	the government to make sure ..." → 「政府としては願っていると思います」は誤訳。「政府として～を確実にする」が正しい (26.6)
翻訳者評価 (エラー値)	「make sure」 を「願っていると思います。」 と訳すのは不適切(1.33)

結果分析　相違: カテゴリ「正確さ-抜けと余分」

原発話 書き起こし	And I am also confident that our relations with other countries would get better with time.
同時通訳 書き起こし	ただ他の国々との関係も時間が経つにつれさらに良くなっていくだろうと
同時通訳者 評価 (エラー値)	なし(0)
翻訳者評価 (エラー値)	「And I am also confident that」 部分が訳されていない(1.33)

結果分析　相違: カテゴリ「流暢さ」

原発話 書き起こし	It is now 30 years ago since I left Kyoto after a year which started in 1985 it ended in nineteen eighty six.
同時通訳 書き起こし	30 年ぶりですけれども1985年か1986年まで滞在したことがご ざいます。
同時通訳者評価 (エラー値)	なし(0)
翻訳者評価 (エラー値)	英文の語順通りに訳出されているため表現が不自然で流暢さに欠ける(0.67)

結果分析　相違: カテゴリ「流暢さ」

原発話 書き起こし	And you know that there have been attacks on police outposts and there was one attack on a police outpost as late as yesterday when one policeman was killed
同時通訳 書き起こし	警察の攻撃ということも言われておりますけれども例えば昨日でも警察官が 1人殺されてることがあります。
同時通訳者 評価 (エラー値)	なし(0)
翻訳者評価 (エラー値)	表現が不自然で流暢さに欠ける (0.67)

まとめ

考察

- 同時通訳評価にも応用可
  - カテゴリ「正確さ-誤訳-」
- 同時通訳評価応用不可
  - カテゴリ「流暢さ」「正確さ-抜けと余分-」
  - 翻訳に比べて評価の許容範囲を広めるべき

今後の方針

- 専門用語の訳出評価
  - 大規模言語モデルでは評価が適切に行われていないため

参考文献

- [1] Ryo Fukuda, Sashi Novitasari, Yui Oka, Yasumasa Kano, Yuki Yano, Yuka Ko, Hirotaka Tokuyama, Kosuke Doi, Tomoya Yanagita, Sakriani Sakti, Katsuhito Sudoh and Satoshi Nakamura. “Simultaneous Speech-to-speech Translation System with Transformer-based Incremental ASR, MT, and TTS,” Proc. Oriental COCOSDA 2021, 186-192, Nov. 18, 2021
- [2] A.Lommel et al., “Multidimensional Quality Metrics (MQM):A Framework for Declaring and Describing Translation Quality Metrics,” Revista Tradumàtica: tecnologies de la traducció, 2014
- [3] 一般社団法人 日本翻訳連盟(JTF), “JTF 翻訳 品質評価ガイドライン”: https://www.jtf.jp/pdf/jtf_translation_quality_guidelines_v1.pdf
- [4] Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. “Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation,” Transactions of the Association for Computational Linguistics, 9:1460–1474. 2021
- [5] Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021. Results of the WMT21 Metrics Shared Task: Evaluating Metrics with Expert-based Human Evaluations on TED and News Domain. In Proceedings of the Sixth Conference on Machine Translation, pages 733–774, Online. Association for Computational Linguistics
- [6] NAATI Metrics: https://www.naati.com.au/wp- content/uploads/2020/10/Certified-Interpreter- Assessment-Rubrics.pdf
- [7] EU Mettics: https://europa.eu/interpretation/doc/marking_criteria_ en.pdf
- [8] Antonios Anastasopoulos, Loïc Barrault, Luisa Bentivogli, Marcely Zanon Boito, Ondřej Bojar, Roldano Cattoni, Anna Currey, Georgiana Dinu, Kevin Duh, Maha Elbayad, Clara Emmanuel, Yannick Estève, Marcello Federico, Christian Federmann, Souhir Gahbiche, Hongyu Gong, Roman Grundkiewicz, Barry Haddow, Benjamin Hsu, et al.. 2022. Findings of the IWSLT 2022 Evaluation Campaign. In Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022), pages 98–157, Dublin, Ireland. Association for Computational Linguistics.
- [9] 松下佳世, 山田優, 石塚浩之, “英日・日英通訳 データベース(JNPC コーパス)の概要”, 日本翻 訳学会, vol. 22, pp. 87–94, 2020