MQM-Chat: 対話翻訳のための多次元品質指標

Yunmeng Li[1] 鈴木 潤[1,2] 森下 睦[3,1*] 阿部 香央莉[4,1,2*] 乾 健太郎[5,1,2]
[1] 東北大学 [2] 理化学研究所 [3] フューチャー株式会社
[4] マシンラーニング・ソリューションズ株式会社 [5]MBZUAI
li.yunmeng.r1@dc.tohoku.ac.jp

概要

対話翻訳におけるスタイライズされたコンテンツや対話の一貫性は，機械翻訳にとって重要な課題である．本研究では，これらの問題を評価する指標として「対話翻訳のための多次元品質評価基準（MQM-Chat）」を提案する．MQM-Chat は曖昧さ，流行語，対話不整合性などの 7 種類のエラータイプで構成される．5 つの機械翻訳モデルで生成した対話データに対し，人間によるアノテーションを行った結果，MQM-Chat は既存の評価基準よりも効果的にエラーを分類し，対話特有の問題を明確に示した．

1 はじめに

ニューラル機械翻訳（NMT）の進展により，形式化された書き言葉であるニュース記事や学術論文などの翻訳性能は著しく向上している [1]．一方，話し言葉である対話を翻訳する場合，対話に含まれる曖昧さ，話者の感情や個性，文化的ニュアンスなどを反映して適切に翻訳できるかが課題となる [2]．

対話翻訳性能を向上させるには，現行の機械翻訳モデルがどれだけ適切に対話に対応できるかを定量化する指針，すなわち評価指標が鍵となる．しかし，従来の評価指標では，対話中に頻出する表現を起因とする翻訳エラーをうまく捉えきれないという課題がある．具体的には，原文中のタイプミスやスラング，主語省略などによるエラーを引き起こすことがある（表 1）．このため，対話翻訳では話者の意図したニュアンスやスタイル，対話の一貫性を評価する枠組みが必要となる．

そこで本研究では，対話翻訳の評価に特化した「対話翻訳のための多次元品質評価基準（Multidimension Quality Metrics for Chat Translation, MQM-Chat）」を提案する．MQM-Chat は，既存の多次元品質評価基準

曖昧さと過度な訂正（*Ambiguity and Disambiguation*）	
原文	知ってｒ？昨日、ヘレンとあったよ
参照訳	u know waht, I saw Helen yesterday
エラー訳	You know what, I saw Helen yesterday
流行語・借用語の問題（*Buzzword or Loanword Issues*）	
原文	草 wwwwww
参照訳	lol
エラー訳	grass
対話の不整合性（*Dialogue Inconsistency*）	
原文	まどかは昨日買い物に行ったよ - 行った？聞いてないよ！
参照訳	Madoka went shopping yesterday. - She went? I didn't hear about it!
エラー訳	Madoka went shopping yesterday. - You went? I didn't hear about it!

表 1 MQM-Chat で定義された対話特有のエラーの例．

（Multidimension Quality Metrics, MQM）[1] [3, 4, 5] を基に，**誤訳，省略・追加，用語・固有名詞の誤り，不自然なスタイル，曖昧さ，流行語，対話不整合性**を重視した 7 つのエラータイプで構成される。

本研究では，5 つの機械翻訳モデルに対し，中英・日英の対話翻訳能力の評価を MQM-Chat で行った．結果，MQM-Chat は MQM より細分化された評価分類を提供し，標準 MQM でのエラーのうち約 30%が対話特有の問題として分類された．これは，既存の MQM では対話翻訳特有の課題をつぶさに評価できず，MQM-Chat を用いることで，それらの課題に対する現行モデルの強み・弱みを詳細に評価することができることを示している．また，few-shot 学習を用いた MQM-Chat による自動評価も試みた．結果として，few-shot 学習による MQM-Chat の自動評価は，システム全体評価において人手評価と一致する程度の性能を示したが，精度面では課題があることが判明した．

まとめると，本研究では (1) MQM-Chat の提案 (2)

* 東北大学の学術研究員としての成果

1） https://themqm.org/

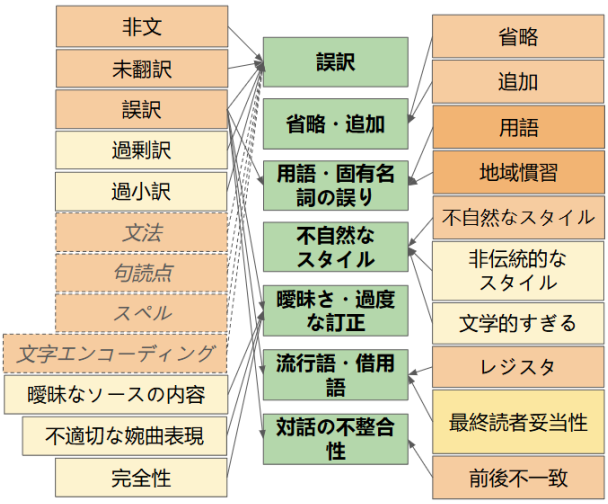

図 1 標準 MQM（橙: MQM-Core，黄: MQM-Full）と MQM-Chat（緑）のエラータイプの対応．色が濃いブロックは，対応するサブカテゴリ全てが MQM-Chat に含まれ，統合されている．灰色のテキストが表示されているブロックは，重大なエラーのみマークされるエラーである．

中英・日英の対話翻訳データの構築 (3) MQM-Chat を用いた人手・自動評価実験の 3 つを行った．これらの貢献は，現状の対話翻訳の問題点・理解を深化させ，今後の進展に役立つと考えられる．

2 関連研究

2.1 翻訳評価指標

従来指標の BLEU [6] は，n-gram や unigram を用いて参照訳との一致に基づく自動評価を生成する．しかし，BLEU では適切な評価ができないとして，人手評価に基づくフレームワークである多次元品質評価基準（MQM）の使用が推進されている [3, 4, 5]．MQM は，単語レベルのエラーや意味的正確性，スタイルのニュアンスなどを人手で詳細に評価する．MQM Core[2] および MQM Full[3] には 39 種以上のエラータイプが含まれ，これらのエラータイプとエラーの重大性（Major，Minor，Neutral）をアノテーションとして付加する．

近年は自動評価においても MQM による評価が試みられている．xCOMET [7] や GEMBA-MQM [8] は，LLM に MQM の説明を指示し，入力された翻訳文に対して自動で評価を行うという手法である．

2） https://themqm.org/the-mqm-typology/
3） https://themqm.org/the-mqm-full-typology/

2.2 対話翻訳タスク

対話はスラングや慣用表現，話者の個性を反映したスタイルを頻繁に含むため，翻訳タスクを複雑にする [9]．この複雑な課題を解決するため，対話翻訳に注力したワークショップが開催された [2]．このワークショップは主に顧客サービス対話に焦点を当て，文法的正確性の評価を重視した．また，Liang らは一貫性・流暢さ・話者の個性を考慮し，対話翻訳に特化したモデルを提案した [10]．対話の一貫性に焦点を当て，対話の翻訳文が誤訳かどうかを分類する誤訳判定器を提案した研究もある [11, 12]．

本研究では対話翻訳の評価プロセスを洗練し，曖昧さや流行語などの文化的ニュアンス，対話の一貫性を考慮する新たな評価指標である MQM-Chat を構築した．また，構築された評価指標を元に，中英・日英対話翻訳データセットを作成し，現行モデルの人手・自動評価を行った．

3 MQM-Chat

本研究は，高品質な対話翻訳を「正確さを維持しながら，話者の個性やスタイル，文化的ニュアンスを効果的に捉え，伝えること」と定義する．高品質な対話翻訳を実現するために，評価基盤を整備することが重要と考えた．そこで本研究では，既存の MQM を基に，対話翻訳特有のカテゴリを導入した MQM-Chat を新たに構築した．MQM-Chat と標準 MQM の対応は図 1 に示される．

MQM-Chat は，以下の 7 種類の分類で構成される．後者の 3 つは対話翻訳に特化した分類である．

誤訳 翻訳で意味を歪める不適切な語彙選択や文法エラー．これらは，翻訳の理解可能性や正確性に直接影響するため重要である．

省略・追加 原文にある情報の欠落，原文にない情報の追加．原文の意図が誤解され，文章の一貫性が乱される可能性がある．

用語・固有名詞の誤り 専門用語や固有名詞の不正確な翻訳．特に，専門的文脈で信頼性を損なう可能性がある．

不自然なスタイル 文法的に正しいが自然でない表現．これにより，翻訳の理解しやすさが損なわれることがある．

曖昧さ・過度な訂正 原文の曖昧さや誤りが翻訳において訂正され，正確に反映されていない場合．表 1 に示した通り，「知って r」のタイプミスを，

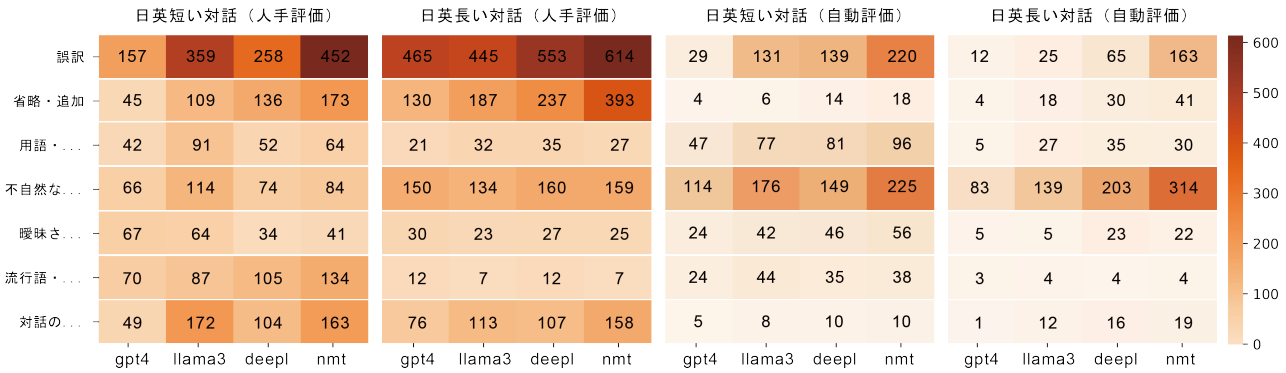

図 2　MQM-Chat の人手評価で判断された各エラータイプの数のヒートマップ．色が濃いほど数値が高いことを示す．

“You know what” と正しい綴りのまま翻訳した場合，本分類と見なされる．本研究では，このような表現は対話のニュアンスとして保持し，タイプミスのまま正確に翻訳すべきであり，原文の欠陥を保持することが重要であると考える．

流行語・借用語　流行語等の用語が正確に翻訳されていない場合．表 1 のように，日本語のミーム「草」（笑い）は，現行の翻訳モデルでは話者の意図と異なる形に誤訳されることが多い．

対話の不整合性　話者が変わる際，あるいは同一話者の発話で文脈の一貫性を欠いている場合．表 1 に示されるように，日本語の「まどか」（彼女）が省略されていることから，翻訳時にエラーが発生している．指示代名詞や主語省略の不適切な処理がこれに当たる．翻訳の一貫性を保ち，読者の混乱を防ぐために重要である．

4　実験: MQM-Chat 人手評価

MQM-Chat の有効性を検証するために，中英・日英の対話翻訳を行い，5 種類のモデル出力に対して MQM-Chat による人手評価を行った．

4.1　実験設定

データセット　短い対話として，日英翻訳には，Open 2ch Dialogue Corpus [13] から，曖昧な内容と流行語・スラング等を多く含む 200 件を使用した．中英翻訳には，LCCC-base データセット [14] から 200 件を抽出した．また長い対話として，日英翻訳には BPersona-chat [15, 11] から 100 件，中英翻訳には NaturalConv [16] から 100 件の対話を使用した．

翻訳モデル　LLMs である GPT-4 [17] と LLaMA3（70B-Instruct） [18]，商用システムの DeepL[4)]，sentence-to-sentence Transformer モデル [19] である Facebook@WMT21（中英用） [20] と SKIM@WMT23（日英用） [21] を用いた．GPT-4 と LLaMA3 は zero-shot 学習の設定で，Hendy らの方法論 [22] に基づいたプロンプトを使用した．詳細なプロンプトは以下となる．

> You are a professional {source_language} to {target_language} translator.This is a {source_language} to {target_language} chat translation task. Please translate each line of the chat from {source_language} to {target_language}. Each line of the chat is considered a message sent by a different speaker.

人手評価　日本語・英語ペア，中国語・英語に通じたプロの評価者を各 6 名（計 12 名）募集し，Label Studio [23] を用いて MQM および MQM-Chat 評価指標に基づいて各モデル出力を評価した．評価者には詳細なガイドラインが提供され，一貫性を持たせるため，目視でレビューを実施した．レビューでは，エラーの範囲やラベルの一致を重点的に確認し，不適切な範囲は修正した．MQM-Chat のエラータイプに精通し，中国語，日本語，英語に堪能な 2 名のレビュアが評価結果を最終確認した．

4.2　分析: MQM-Chat のエラー分布

結果，MQM-Chat の分類に従うと，誤訳が多く見られるものの，特に日英翻訳において多様なエラー分布を見せている．日英翻訳では，中英翻訳よりもエラー全体が多く見られた．誤訳が多い理由としては，翻訳モデルが対話翻訳データに特化して訓練されていない可能性が考えられる．MQM-Chat が示す日英翻訳エラー分布を図 2 に示す．対話特有のエラーに対して，MQM-Chat はいくつかの洞察を提供した．また誤訳や不自然なスタイルのエラーは，短い対話より長い対話で頻繁に起こる傾向があり，長

4）https://www.deepl.com/translator

	zh⇒en		*ja⇒en*	
	再分類 (%)	対話特有 (%)	再分類 (%)	対話特有 (%)
S G	39 (50.7)	23 (59.0)	92 (57.9)	41 (44.6)
L	46 (48.9)	29 (63.0)	201 (74.2)	80 (39.8)
D	46 (52.9)	18 (39.1)	172 (81.1)	77 (44.8)
N	94 (67.6)	35 (37.2)	329 (86.6)	120 (36.5)
L G	16 (27.6)	3 (18.8)	33 (25.2)	10 (30.3)
L	21 (26.3)	10 (47.6)	69 (30.5)	17 (24.6)
D	71 (49.7)	34 (47.9)	92 (43.8)	26 (28.3)
N	128 (55.4)	48 (37.5)	278 (61.5)	49 (17.6)

表 2　標準 MQM でエラー判定された例のうち，MQM-Chat で別種の例に再分類された割合を示す．S は短い対話，L は長い対話データ，G は GPT-4，L は LLaMA3，D は DeepL，N は NMT モデルを示す．対話特有（%）は，全標準 MQM エラーに対して**曖昧さ，流行語，対話の不整合性**に再分類された例の割合である．

い対話では省略や追加も顕著に見られた．対照的に，曖昧さと流行語の問題は短い対話でより頻繁に見られる一方で，対話の不整合性は対話の長さやソース言語にかかわらず一貫して存在した．日英翻訳では，特に省略または追加，不自然なスタイル，対話の不整合性が頻繁に確認された．逆に，曖昧さと解消のエラーは中英翻訳において他のエラーと同様の頻度で発生した．これは中国語対話における句読点省略による曖昧さが要因と考えられる．短い日本語対話では用語や固有名詞の誤り，曖昧さ，流行語のエラーが長い対話以上に発生している．翻訳モデルに関して，GPT-4 は短い対話翻訳において他モデルよりも少ない誤訳を生成したが，長い対話では同じ量を示した．GPT-4 は特に長い対話で流行語や借用語の問題を多く出し，LLaMA3 は短い対話で一貫性の問題を抱え，特に短い日本語対話で用語やスタイルの問題を多く生成した．全体として，MQM-Chat は翻訳課題の傾向をより深く理解するための有用なツールであることを示している．

4.3　分析: MQM → MQM-Chat の再分類

対話特有の翻訳エラーを捉える能力において，MQM-Chat が標準 MQM より優れていることを示すため，標準 MQM と MQM-Chat 間で人手評価の結果にどのような変化があったかを調査した．具体的には，MQM-Chat において，標準 MQM と対応するもの以外のエラータイプとしてラベル付けされたもの[5)]の割合を評価した（表 2）．結果，MQM-Chat では 25.2 ~ 86.6% のエラーが再分類されており，しかもその多くが対話特有のエラーであることが判明した．これらの比較結果から，MQM-Chat は対話翻訳特有の問題を，既存の評価指標である MQM よりも正確に検出し分析できることが示された．3 つの対話特有エラータイプと判断された実際のモデル出力の例を，Appendix A に記載する．

5)　例えば，MQM の**誤訳**が，MQM-Chat の**誤訳**として分類されて**いない**場合，「再分類」に相当する．

5　実験: MQM-Chat 自動評価

GEMBA-MQM [8] のプロンプトを基に，標準 MQM のエラータイプの説明を MQM-Chat のエラータイプに置き換え，自動 MQM-Chat 評価を実装した．MQM-Chat の人手評価結果を例として付与し，few-shot 学習を行った GPT-4 で，4 種類のモデル出力を評価した．ここでは，正解として人手評価の結果を用い，Pairwise accuracy とピアソンの相関を計算した．結果，Pairwise accuracy は 79.17%，ピアソン相関は 0.774 を示し，自動評価の結果がある程度人手評価と一致する傾向が示された．提案された MQM-Chat のエラースパンを検証し，アノテーションがどの程度エラーを詳細に検出できているかを調査した．MQM-Chat 自動評価が示す日英翻訳エラー分布を図 2 の右に示す．ただし，自動評価結果は，人手評価よりも**誤訳**および**不自然なスタイル**に過剰に分類する傾向があった．また，自動評価で判定することができた日英のエラー数は中英よりも少なかった．これらの結果は，GPT の訓練データセットにおける日本語対話データの限界が影響していると考えられる．

6　結論

本研究では，対話翻訳タスクの評価基盤を整えるため，MQM-Chat を提案し，その有効性を一連の実験で評価した．エラー分布分析により，MQM-Chat が既存の評価指標よりも対話翻訳におけるモデルの弱点を効果的に特定できることが示された．また，MQM-Chat による自動評価の実験も行い，自動評価が人間によるシステムランキングと一致することが確認された．しかし，詳細なエラー検知の自動評価精度についてはまだ改良の余地がある．今後は，カスタムサービスなど他ドメインの対話翻訳データで MQM-Chat を評価し，MQM-Chat の有効性をより検証していく．また，自動 MQM-Chat を改良し，対話翻訳タスクの効果的な評価基準として機能させることを目指す．

謝辞

本研究は，JST 科学技術イノベーション創出に向けた大学フェローシップ創設事業 JPMJFS2102，JST 次世代研究者挑戦的研究プログラム JPMJSP2114，JSPS 科研費 22H00524，JST CREST JPMJCR20D2，JST ムーンショット型研究開発事業 JPMJMS2011-35 (fundamental research) の支援を受けたものです．アノテーションツール Label Studio（https://labelstud.io/）へ感謝を申し上げます．

参考文献

[1] Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. Findings of the 2020 conference on machine translation (WMT20). In **Proceedings of the Fifth Conference on Machine Translation**, pp. 1–55, Online, November 2020. Association for Computational Linguistics.

[2] M. Amin Farajian, António V. Lopes, André F. T. Martins, Sameen Maruf, and Gholamreza Haffari. Findings of the WMT 2020 shared task on chat translation. In **Proceedings of the Fifth Conference on Machine Translation**, pp. 65–75, Online, November 2020. Association for Computational Linguistics.

[3] Aljoscha Burchardt. Multidimensional quality metrics: a flexible system for assessing translation quality. In **Proceedings of Translating and the Computer 35**, London, UK, November 28-29 2013. Aslib.

[4] Valerie R Mariana. **The Multidimensional Quality Metric (MQM) framework: A new framework for translation quality assessment**. Brigham Young University, 2014.

[5] 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**, Vol. 9, pp. 1460–1474, 2021.

[6] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics**, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[7] Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. xcomet: Transparent machine translation evaluation through fine-grained error detection, 2023.

[8] Tom Kocmi and Christian Federmann. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In **Proceedings of the Eighth Conference on Machine Translation**, pp. 768–775, Singapore, December 2023. Association for Computational Linguistics.

[9] Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. How noisy social media text, how diffrnt social media sources? In **Proceedings of the Sixth International Joint Conference on Natural Language Processing**, pp. 356–364, Nagoya, Japan, October 2013. Asian Federation of Natural Language Processing.

[10] Yunlong Liang, Fandong Meng, Yufeng Chen, Jinan Xu, and Jie Zhou. Modeling bilingual conversational characteristics for neural chat translation. In **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)**, pp. 5711–5724, Online, August 2021. Association for Computational Linguistics.

[11] Yunmeng Li, Jun Suzuki, Makoto Morishita, Kaori Abe, Ryoko Tokuhisa, Ana Brassard, and Kentaro Inui. Chat translation error detection for assisting cross-lingual communications. In **Proceedings of the 3rd Workshop on Evaluation and Comparison of NLP Systems**, pp. 88–95, Online, November 2022. Association for Computational Linguistics.

[12] Yunmeng Li, Jun Suzuki, Makoto Morishita, Kaori Abe, and Kentaro Inui. An investigation of warning erroneous chat translations in cross-lingual communication. In **Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: Student Research Workshop**, pp. 10–16, Nusa Dua, Bali, November 2023. Association for Computational Linguistics.

[13] Michimasa Inaba. A example based dialogue system using the open 2channel dialogue corpus. In **Proceedings of SIG-SLUD-B902-33**, pp. 129–132, 2019.

[14] Yida Wang, Pei Ke, Yinhe Zheng, Kaili Huang, Yong Jiang, Xiaoyan Zhu, and Minlie Huang. A large-scale chinese short-text conversation dataset. In **NLPCC**, 2020.

[15] Hiroaki Sugiyama, Masahiro Mizukami, Tsunehiro Arimoto, Hiromi Narimatsu, Yuya Chiba, Hideharu Nakajima, and Toyomi Meguro. Empirical analysis of training strategies of transformer-based japanese chit-chat systems, 2021.

[16] Xiaoyang Wang, Chen Li, Jianqiao Zhao, and Dong Yu. Naturalconv: A chinese dialogue dataset towards multi-turn topic-driven conversation, 2021.

[17] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. **arXiv preprint arXiv:2303.08774**, 2023.

[18] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. https://ai.facebook.com/blog/large-language-models-llama-3, 2024.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. **Advances in neural information processing systems**, Vol. 30, , 2017.

[20] Chau Tran, Shruti Bhosale, James Cross, Philipp Koehn, Sergey Edunov, and Angela Fan. Facebook AI's WMT21 news translation task submission. In **Proceedings of the Sixth Conference on Machine Translation**, pp. 205–215, Online, November 2021. Association for Computational Linguistics.

[21] Keito Kudo, Takumi Ito, Makoto Morishita, and Jun Suzuki. SKIM at WMT 2023 general translation task. In **Proceedings of the Eighth Conference on Machine Translation**, pp. 128–136, Singapore, December 2023. Association for Computational Linguistics.

[22] Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. How good are gpt models at machine translation? a comprehensive evaluation, 2023.

[23] Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. Label Studio: Data labeling software, 2020-2022. Open source software available from https://github.com/heartexlabs/label-studio.

A　対話特有エラータイプの例

Example 1	
source	我在那滴 (a typographical error of “那地”) 吃的 饭。。
NMT	I had my meals in that drop....
(ref)	theere (a typographical error of “there”)
標準 MQM	誤訳 - Critical
MQM-Chat	**曖昧さ・過度な修正 - Major**
Example 2	
source	... 我只是为了凹造型人 艰勿拆
DeepL	I just for the sake of the shape of the people hard not to break down
(ref)	Ren Jian Wu Chai (Chinese transliteration) or life is already so hard or arduous, so don’t judge me. (the meaning)
標準 MQM	誤訳 - Major
MQM-Chat	**流行語・借用語 - Major**
Example 3	
source	... 結婚して早く家を出ろって母がうるさくて。 - そうだったのかぁ。うち [1] とは逆だね。うちは一緒にいてほしいみたいだよ。[2]
NMT	...my mother insisted that I get married and leave the house as soon as possible. - I see, it’s the opposite of my house[1]. I want you to stay with me.[2]
(ref)	(1) my family, (2) She want me to stay with her.
標準 MQM	(1) 誤訳 - Major, (2) 誤訳 - Critical
MQM-Chat	(1) 誤訳 - Major, (2) **対話の不整合性 - Major**

表 3　標準 MQM で誤訳の分類と評価されたが，MQM-Chat で別種の対話特有エラータイプ（曖昧さ，流行語，対話の不整合性）として分類された例.

表 3 は，3 つの対話特有エラータイプとして分類された実際のモデル出力例を示す．全ての例において，標準 MQM では一元的に「誤訳」と分類されているが，MQM-Chat ではさらに詳細に細分化されていることが読み取れる．例 1 では，翻訳モデルが「那地」だと誤認識した例，「那滴」が MQM-Chat では曖昧さとしてラベル付けされた．例 2 では，人 艰勿拆というスラングを，MQM-Chat では流行語の問題としてラベル付けされている．例 3 では，「うち」という表現に対する共参照の問題を，MQM-Chat では対話の不整合性としてラベル付けしている．