卒業論文

誤り情報注入型ピボット翻訳

指導教官　　村上 陽平 教授

立命館大学　情報理工学部
先端社会デザインコース　4 回生
260022034-6

細見 涼乃

2025 年度（秋学期）卒業研究 3（CH）
令和 8 年 1 月 30 日

誤り情報注入型ピボット翻訳

細見 涼乃

内容梗概

近年，大規模言語モデルを含む機械翻訳技術の進歩により，英語や中国語などの大規模な並列コーパスが豊富に存在する高資源言語ペアでは高品質な翻訳が実現されつつある．一方で，クメール語やミャンマー語などの低資源言語では，並列コーパスの不足により高品質な翻訳モデルの構築が依然として困難である．この課題に対する手法として，高資源言語を中間言語とするピボット翻訳が提案されている．ピボット翻訳は，高資源言語へ翻訳した後に目的言語へ翻訳する二段階の翻訳方式であり，中間言語には豊富な並列データを利用できる点から，一定の翻訳精度の向上が報告されている．しかしながら，中間翻訳で生じた情報損失や曖昧性の誤解釈が最終翻訳へ伝播する問題が指摘されており，翻訳品質の安定的な向上には限界がある．

そこで，本研究では，大規模言語モデルを用いてピボット翻訳における中間翻訳の誤りを検出し明示化することで，最終翻訳の品質向上を図る手法を提案する．具体的には，中間言語に対して AUTOMQM を適用して誤り情報を抽出し，それらが最終翻訳に残存しないよう制御することで，低資源言語における翻訳品質の向上を目指す．本手法の実現にあたり，取り組むべき課題は以下の 2 点である．

複数情報源混合プロンプトの設計

本手法では，原文，中間翻訳文，および AUTOMQM により抽出された誤り情報を同時に大規模言語モデルへ入力する．しかしながら，これらを適切に参照させなければ，特定の情報や重要度の低い誤りに過度に依存し，翻訳品質が低下する可能性がある．そのため，各情報源の役割を明確化するとともに，入力する誤り情報を重要度の高いものに絞るプロンプト設計が必要となる．

翻訳品質評価指標の統合

本手法での品質評価には，単一の指標ではなく，複数の観点を統合的に考慮することが必要である．具体的には，翻訳結果と参照訳との一致度，翻訳文全体の意味的妥当性，および翻訳過程において生じた誤りの修正状況という異なった側面が存在する．しかしながら，既存の自動評価指標は，これらの観点を同時に捉えることが難しい．そのため，翻訳品質をどのように統合的に評価するかが課題となる．

一つ目の課題に対しては，原文・中間翻訳文・AUTOMQM により出力された誤り情報の役割を明確化し，それに基づく翻訳プロンプトを構築した. 具体的には，中間翻訳文を主ソース，原文を曖昧性解消の補助ソースとし，誤り情報を再発防止の制約として与えた．さらに，誤り情報をすべて入力する場合と，重大度の高い誤りに限定する場合を比較し，提示方法の影響を検証した.

二つ目の課題に対しては，翻訳を単一の指標で評価するのではなく，評価観点ごとに適切な指標を割り当てる統合的な評価フレームワークを採用した．具体的には，翻訳結果と参照訳との一致度に対して BLEU，chrF++および TER を，翻訳文全体の意味的妥当性に対して GEMBA-DA と AUTOMQM を用い，これらの数値スコアは翻訳品質の全体傾向を把握するための補助的指標と位置付けた. 加えて，翻訳過程において生じた誤りの修正状況に対しては AUTOMQM の誤り出力を用いて，中間翻訳での誤り情報が最終翻訳に残存しているかを確認し，誤り単位での分析を行った. 以上の評価方法により，誤り認識型ピボット翻訳手法の有効性を多角的に検証した．本研究の貢献は以下の通りである.

複数情報源を適切に参照させるプロンプト設計

提案手法により，中間翻訳で検出された誤りのうち、中国語で 74.1%，インドネシア語で 76.5%，タイ語で 70.4%，ミャンマー語で 72.8%，クメール語で 64.2%が最終翻訳段階で修正されることを確認した. この結果から，本手法が中間翻訳に起因する誤り伝播の抑制に有効であることが示された．また，高資源言語では，重大度の高い誤りのみに限定して提示する手法が有効である一方，低資源言語では誤り情報を限定せずに提示した方が，より高い翻訳品質が得られる傾向が確認された.

翻訳品質評価指標の統合

AUTOMQM の誤り出力および対象言語を母語とする話者による人手評価を分析した結果，本手法により，中間翻訳で検出された誤りに対して，文体に関する表記上の問題および用語の不正確さにおいて，修正率を約 10〜40%向上させた. また，情報抜けについても，修正率を約 50%〜75%へ向上させた. さらに，従来のピボット翻訳手法と比較すると，低資源言語ほど修正率の増加幅が大きい傾向が見られ，低資源言語条件下で本手法が特に有効であることが示された.

Pivot Translation with Error Information Injection

Suzuno Hosomi

Abstract

In recent years, advances in machine translation technologies, including large language models, have enabled high-quality translation for high-resource language pairs such as English and Chinese. In contrast, for low-resource languages such as Khmer and Myanmar, the lack of parallel corpora still makes it difficult to build high-quality translation models. As an approach to this problem, pivot translation using a high-resource language as an intermediate language has been proposed. Although pivot translation can improve translation accuracy through a two-step process, information loss and misinterpretation of ambiguity in the intermediate translation may propagate to the final output.

To address this issue, this study proposes a method that improves translation quality by detecting and explicitly representing errors in the intermediate translation using large language models. Specifically, AUTOMQM is applied to the intermediate language to extract error information, which is used to control the final translation so that these errors do not remain. Through this approach, we aim to improve translation quality for low-resource languages and focus on the following two challenges.

Design of Multi-Source Prompting

In this method, the source sentence, the intermediate translation, and the error information extracted by AUTOMQM are simultaneously provided to a large language model. However, if these sources are not properly referenced, the model may over-rely on less important information or minor errors, leading to degraded translation quality. Therefore, it is necessary to clearly define the role of each information source and design prompts that focus on errors with higher importance.

Integration of Translation Quality Evaluation Metrics

In this method, translation quality is evaluated from multiple perspectives, including similarity to the reference translation, semantic adequacy, and error correction. However, existing automatic evaluation metrics have difficulty capturing these aspects simultaneously, highlighting the need for an integrated

evaluation approach.

To address the first challenge, we clarified the roles of the source sentence, the English intermediate translation, and the error information output by AUTOMQM, and designed translation prompts accordingly. The intermediate translation was treated as the primary source, the source sentence as an auxiliary source, and the error information as constraints to prevent error recurrence. We also examined the impact of different scopes of provided error information on translation quality.

To address the second challenge, we adopted an integrated evaluation framework that assigns appropriate metrics to each evaluation aspect. Translation quality was evaluated in terms of similarity to the reference translation, semantic adequacy, and error correction, using AUTOMQM error outputs to validate the effectiveness of the proposed method. The contributions of this study are summarized as follows.

Design of Multi-Source Prompting

Using the proposed method, we confirmed that 74.1% of the errors detected in the intermediate translation for Chinese, 76.5% for Indonesian, 70.4% for Thai, 72.8% for Myanmar, and 64.2% for Khmer were corrected in the final translation. These results indicate that the proposed method is effective in suppressing the propagation of errors originating from the intermediate translation. In addition, while limiting the input to high-severity errors is effective for high-resource languages, providing all error information tends to yield higher translation quality for low-resource languages.

Design of Multi-Source Prompting

Analysis using AUTOMQM and evaluations by native speakers showed that the proposed method improved the correction rate for stylistic and terminological errors by approximately 10-40%, and for missing information, it improved the correction rate from approximately 50% to 75%. This improvement rate was greater for low-resource languages, demonstrating the effectiveness of this method even in low-resource environments.

誤り情報注入型ピボット翻訳

目次

第1章 はじめに

近年，大規模言語モデルを含む機械翻訳技術の進歩により，英語や中国語などの大規模な並列コーパスが豊富に存在する高資源言語ペアにおいては，自然で高品質な翻訳が実現されつつある．一方で，クメール語やミャンマー語などの低資源言語では，並列コーパスの不足が深刻な課題となっており，高品質な翻訳モデルの構築が困難である．この背景には，低資源言語と呼ばれる言語の多くが，話者人口が比較的に少ないことに加え，研究コミュニティや資金提供機関から十分な支援を受けてこなかった，あるいは社会的に周縁化・少数化された集団によって使用されてきたといった要因が存在する．その結果，大規模かつ高品質な言語資源の整備が進まず，機械翻訳を含む自然言語処理技術の発展から取り残されやすい状況にある[1]．この問題に対するアプローチとして，高資源言語を中間言語とするピボット翻訳が提案されている．ピボット翻訳とは，原言語から直接目的言語へ翻訳するのではなく，一度英語などの高資源言語へ翻訳したのち，そこから目的言語へ翻訳する二段階の翻訳方式である．中間言語には豊富な並列データが存在するため，低資源言語ペアに対しても翻訳精度の向上が期待できる．しかしながら，ピボット翻訳では，中間翻訳が最終翻訳の唯一の入力となるため，一度誤った意味表現が生成されると，後段の翻訳過程でそれを修正することが困難となり，翻訳品質の安定的な向上には限界がある[2]．

そこで，本研究では，大規模言語モデルを用いてピボット翻訳における中間翻訳の誤りを明示的に扱い，最終翻訳の品質向上を図る手法を提案する．具体的には，中間言語に対して AUTOMQM を適用し，中間翻訳で発生した誤りの種類や重大度，発生箇所といった誤り情報を抽出し，この誤りが最終翻訳に反映されないよう制御することで，低資源言語における翻訳品質の向上を目指す．本手法の実現にあたり，取り組むべき課題は以下の 2 点である．

複数情報源を適切に参照させるプロンプト設計

本手法では，日本語原文，中間言語である英語翻訳，および AUTOMQM により抽出された誤り情報を同時に大規模言語モデルへ入力する．しかしながら，これらの情報を適切に参照させなければ，モデルが特定の情報に過度に依存したり，誤り情報を不適切に反映してしまう可能性がある．また，AUTOMQM により抽出される誤り情報をすべて入力した場合，翻訳生成時に重要度の低い誤りまで過度に意識されることで，翻訳品質を低下させる可能性がある．そ

のため，各情報源の役割を明確化するとともに，入力する誤り情報を重要度の高いものに絞るプロンプト設計が必要となる．

翻訳品質評価指標の統合

本手法における翻訳品質の評価には，単一の指標に基づく評価ではなく，複数の観点を統合的に考慮することが重要である．翻訳品質は，翻訳結果と参照訳との一致度，翻訳文全体の意味的妥当性，さらに翻訳過程において生じた誤りが適切に修正されているかといった，異なる側面から評価される必要がある．しかし，既存の自動評価指標はいずれも特定の観点に強みを持つ一方で，これらすべての側面を同時に十分捉えることは難しく，それぞれに限界が存在する．このため，翻訳品質をどのように統合的に評価するかが，本研究における重要な課題となる．

以下，本論文では，第 2 章で関連研究について述べる．次に，第 3 章で提案手法である誤り情報注入型ピボット翻訳を述べ，第 4 章で実験環境を述べ，第 5 章で評価を述べ，第 6 章で母国語話者による人手評価を述べる．最後に，第 7 章では本稿をまとめ，今後の展望やさらなる課題について結論とする．

第2章 関連研究

本章では，本研究の関連研究となる機械翻訳技術および翻訳品質評価手法について述べる．まず，低資源言語翻訳において広く用いられてきたピボット翻訳手法の有効性と課題を整理する．次に，翻訳品質評価に関する従来の自動評価指標，人手評価，および大規模言語モデル（LLM）を用いた近年の評価手法について述べ，本研究で用いる誤り分析に基づく評価手法である AUTOMQM の特徴と位置づけを明確にする．

2.1 ピボット翻訳

近年の機械翻訳は，ニューラルネットワークを用いたニューラル機械翻訳（Neural Machine Translation; NMT）が主流となっており，大規模な並列コーパスを持つ高資源言語間の翻訳においては，人手翻訳に近い品質が達成されている．一方で，十分な並列コーパスを確保できない低資源言語では，翻訳品質が大きく低下することが知られている．また，Web 上で使用される言語の多様化に伴い，英語と非英語言語間の翻訳に加えて，非英語言語間の翻訳に対する需要も増加している．しかし，多数の言語すべての組み合わせに対して直接翻訳モデルを構築することは，開発コストの観点から現実的ではない．そこで，この課題に対する解決策として，高資源言語を中間言語（ピボット言語）として用いるピボット翻訳手法が提案されている．ピボット翻訳手法とは，図 1 に示す通り，原言語から中間言語，中間言語から目的言語へと段階的に翻訳を行うことで，高資源言語の豊富なデータ資源を活用し，低資源言語翻訳の精度向上を図る手法である．先行研究では，十分な並列データを持たない言語対において，ピボット翻訳手法が有効であることが報告されている．一方で，ピボット翻訳手法には，中間翻訳で生じた誤りが最終翻訳へと引き継がれる「誤り伝播問題」が存在する．特に，中間段階で生じた意味解釈の誤りや情報欠落が，後段の翻訳過程で修正されることなく最終翻訳に影響を及ぼしてしまう点が大きな課題である[3] ．さらに，英語を介して複数の翻訳サービスや翻訳モデルを連鎖的に用いる場合，各翻訳段階における語選択の違いにより，不整合性，非対称性，非推移性といった問題が生じ，語の意味が段階的に変化する意味ドリフトが発生しやすいことが指摘されている．このような語選択の不安定さは，機械翻訳を介したコミュニケーションにおいて，共通理解の形成を困難に

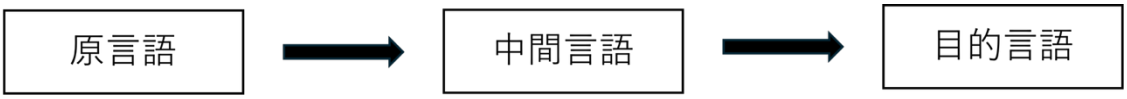

図 1　ピボット翻訳手法

する要因の一つである．これに対し，フレーズテーブルの統合や言語学的注釈を用いた手法など，中間翻訳の曖昧性を低減するアプローチも提案されている．しかしながら，これらの手法は大規模で信頼性の高い並列コーパスの構築や，翻訳システム内部への改変を前提とする場合が多く，既存の翻訳サービスに即時適用することは容易ではない．したがって，ピボット翻訳手法の精度向上のためには，翻訳モデルや翻訳サービスをブラックボックスとして扱いつつ，中間翻訳に由来する意味解釈の誤りや情報欠落が最終翻訳へ伝播することを制御する仕組みが重要な課題となる[4].

2.2 翻訳品質評価

翻訳品質の評価は，機械翻訳研究において重要な課題の一つであり，これまでにさまざまな評価指標や評価手法が提案されてきた．本節では，従来の自動評価指標から，LLM を用いた近年の評価手法までを述べ，本研究の立ち位置を明確にする．

翻訳品質の評価には，BLEU[5]，chrF++[6]，および TER といった自動評価指標が広く用いられてきた．これらの指標は，参照訳と翻訳文の文字列レベルでの一致度を数値化することで，大規模な翻訳結果を効率的に比較できる点に特徴がある．BLEU は翻訳文と参照訳の n-gram 一致率に基づく指標であり，brevity penalty によって過度に短い翻訳を抑制する．一方，chrF++は文字 n-gram と単語 n-gram を組み合わせることで，語形変化が豊富な言語や分かち書きが困難な言語に対しても安定した評価を可能にしている．また，TER は翻訳文を参照訳に変換するために必要な挿入・削除・置換・並べ替えといった編集操作の割合を算出する指標であり，翻訳文が参照訳からどの程度修正を要するかを表す．翻訳後編集の観点から解釈しやすいという利点を持つ．しかしながら，これらの指標はいずれも文字列一致に基づくため，意味的妥当性や文全体の自然さを十分に評価できないという限界を有する．特にピボット翻訳手法においては，中間翻訳に由来する意味のずれや曖昧性が最終翻訳に伝播するという特性があり，こう

した誤りの影響を参照訳依存の指標のみで直接的に捉えることは困難である．

このような自動評価指標の限界を補う方法として，人手評価は翻訳品質評価において最も信頼性の高い方法として位置づけられてきた．中でも，翻訳文全体の品質を連続値で評価する Direct Assessment（DA）は，WMT（機械翻訳全般の国際会議）における標準的な人手評価手法として用いられている．しかしながら，人手評価は高コストであり，大規模評価への適用が難しいという課題を有する．近年では，この課題に対するアプローチとして，LLM を用いて人手評価を代替・補完する手法が提案されている．その代表例が GEMBA（GPT Estimation Metric Based Assessment）であり，翻訳品質評価の手順をプロンプトとして LLM に与えることで，人手評価に近い判断を自動的に行う枠組みである．特に GEMBA-DA は，DA 形式を LLM によって実現し，翻訳文全体の意味的妥当性や自然さを総合的に評価できる．先行研究では，GEMBA-DA が従来の自動評価指標と比較して，参照訳を与えた評価手法が人手評価との相関が最も高いことが報告されている．一方で，翻訳文中のどの部分にどのような誤りが存在するかを特定する詳細な誤り分析には適さないという課題が指摘されている[7].

この課題に対し，翻訳品質をより分析的に捉える枠組みとして，MQM（Multidimensional Quality Metrics）が提案されている．MQM は，翻訳文中の誤りを位置情報とともに特定し，誤り種別や重大度に基づいて分類することで，翻訳品質を詳細に分析する評価フレームワークである．これにより，翻訳品質を単一の数値として評価するのではなく，どの種類の誤りがどの程度発生しているかを明示的に把握することが可能となる．この MQM フレームワークを LLM によって自動化する手法として，AUTOMQM が提案されている．AUTOMQM は，翻訳文中の誤りの位置，内容，重大度，および誤り種別を MQM の基準に従って付与し，結果を可視化できる形で出力する点に特徴がある．これにより，翻訳品質を誤り情報に基づいて分析することが可能となる．しかしながら，既存研究の多くは，AUTOMQM によって検出された誤りを評価・分析の対象とすることにとどまっており，誤り情報を翻訳生成そのものに活用する試みは十分に検討されていない．そこで本研究では，LLM によって検出された誤り情報をピボット翻訳手法の生成過程に明示的に反映させることで，中間翻訳に由来する誤り伝播の軽減を目指す点に新規性を有する[8].

第3章 誤り情報注入型ピボット翻訳

本章では, 本研究で提案する AUTOMQM による誤り情報注入型ピボット翻訳手法について述べる. はじめに, 誤り伝播問題に対処するために提案手法の概要と特徴を示す. さらに, 中核技術である AUTOMQM の役割を明確にした上で, 誤り情報を翻訳生成過程に反映させる処理手順について述べる.

3.1 AUTOMQM による誤り情報の注入

本研究では, ピボット翻訳手法の誤り伝播問題に対処するために, 中間翻訳文に対して AUTOMQM を適用し, 検出された誤り情報を最終翻訳生成に反映させる誤り情報注入型ピボット翻訳手法を提案する. 本手法の処理フローを図 2 に示す. まず, 原言語文を中間言語へ翻訳し, 中間翻訳文を生成する. 次に, 中間翻訳文に対して AUTOMQM を適用し, 誤りの位置, 種類, 重大度を構造化された形式で取得する. 最終翻訳生成では, 中間翻訳文, 原言語文, および AUTOMQM により検出された誤り情報をすべてプロンプトに明示する. このうち, 中間翻訳文を翻訳生成の主たる情報源とし, 原言語文は中間翻訳文の表現が曖昧または不正確な場合に限り補助的に参照される. また, AUTOMQM の誤り情報は, 再発してはならない誤りを示す制約情報としてプロンプト内に付与される. これにより, ピボット翻訳手法の枠組みを維持しつつ, 中間翻訳に由来する誤りの修正を可能とする. 本手法は, 翻訳モデルの再学習を必要とせず, プロンプト設計のみで誤り情報を活用できる点に特徴がある.

なお, 中間翻訳誤り検出手法として LLM を用いた GEMBA-DA が代替可能であるが, GEMBA-DA は翻訳文中の誤り位置や重大度を構造化して出力する設計ではないため, 誤り情報を翻訳生成に直接反映させることは困難である. これに対し, AUTOMQM は翻訳文中の誤りの位置, 種類, 重大度を明示的に取得できるため, 中間翻訳誤りの伝播を抑制するという本研究の目的に適している.

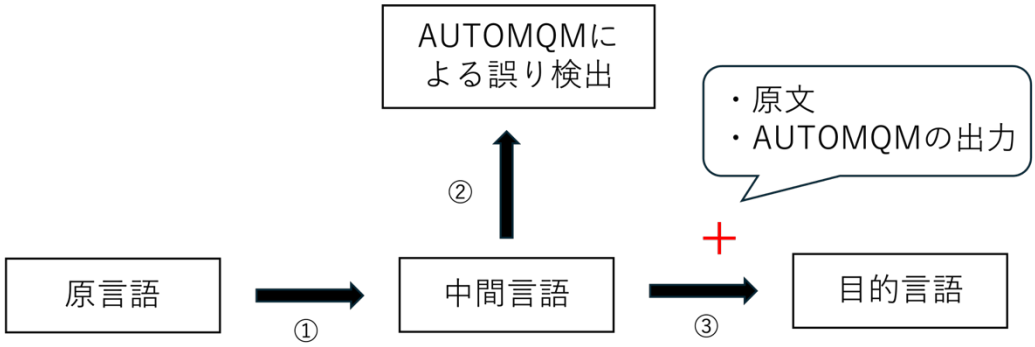

図 2　誤り情報注入型ピボット翻訳手法

3.2 誤り情報注入型ピボット翻訳の手順

本節では，誤り情報注入型ピボット翻訳手法の処理手順について詳述する．

はじめに，ピボット翻訳手法と同様に原言語から中間言語への翻訳を行う．本研究では比較の公平性を確保するため，日本語から英語への中間翻訳文は，ピボット翻訳手法と提案手法で同一のものを用いる．これにより，日本語→英語段階の翻訳品質や誤りの差異が評価結果に影響することを防ぎ，後段の翻訳生成方法の違いのみを比較可能とする．

次に，中間翻訳文に対して AUTOMQM を適用し，翻訳誤りの検出を行う．その際に, AUTOMQM は中間翻訳文に含まれる誤りを誤りの位置, 種類, 重大度, および誤り内容の説明を構造化された形式で出力する．本研究では，この出力を中間誤りメモとして扱い，後続の翻訳生成における制約情報として利用する．図 3 に中間誤りメモの例を示す．SRC は原言語文，HYP は翻訳生成文を表し，OUTPUT_RAW には翻訳品質の総合スコア(overall_score）に加え，誤り種別（error_type），誤りの重大度(severity)，原文中の誤り位置(source_span)，翻訳文中の誤り位置(target_span），および誤り内容の記述(description)が含まれる．一文に対して複数の誤りが検出される場合があるため，各誤りには識別子 id が付与され，二つ目以降の誤りは E2，E3 のように区別される．

続いて，中間翻訳文を基に目的言語への最終翻訳を生成する．本段階では，中間翻訳文，原言語文，および AUTOMQM により検出された誤り情報をすべてプロンプトに明示する．ここから示す各プロンプトは，翻訳生成に用いた実装プロン

プトを，設計意図に基づいて機能別に分解した例である．実際の翻訳生成では，これらの要素を統合した単一のプロンプトを用いている．なお，説明の明確化のため，原言語を日本語，中間言語を英語，目的言語を中国語とした翻訳設定を例として用いる．まず，中間翻訳文を翻訳生成の主たる情報源とし，原言語文は中間翻訳における英語表現が曖昧または不正確な場合に限り，補助的な参照情報として用いる．この参照優先度を制御するプロンプトを図 4 に示す．これにより，原言語から直接翻訳することによるピボット翻訳手法の枠組みからの逸脱を防ぎつつ，中間翻訳に由来する誤りの補正を可能とする．

また，AUTOMQM により検出された誤りは，「再発してはならない誤り」を示す制約情報としてプロンプト内に明示的に付与される．このプロンプトは図 5 に示す．翻訳モデルはこれらの誤り情報を制約条件として解釈し，中間翻訳に起因する誤りを回避または修正する方向で翻訳を生成する．

最後に，翻訳結果の形式的な不整合を防ぐため，目的言語の表記体系および言語的慣習に基づく制約条件を翻訳生成プロンプトに付与する．このプロンプトを図 6 に示す．具体的には，目的言語である中国語の標準的な表記体系に従った翻訳を生成することを求め，日本語仮名の混入や不必要なローマ字表記を含めないこと，ならびに中国語として一般に用いられる語彙や固有表現を優先的に使用することを制約条件として設定する．翻訳結果がこれらの条件を満たさない場合には，翻訳生成が十分に達成されていないと判断し，制約条件を強化したプロンプトを用いて最大 1 回に限り再生成を行う．これは，形式的制約を初期段階から一律に強く適用すると，翻訳内容の過度な単純化や意味情報の欠落を招く可能性があるためであり，まずは通常の制約下で翻訳生成を行い，明らかな形式的不整合が確認された場合にのみ補正を行うことを目的としている．この再生成処理は，評価に用いる翻訳文の最低限の品質を保証するための補助的手続きであり，反復的な試行による性能向上を目的とするものではない．

Sample#1:
SRC: イタリアはハーフタイムの時点で16対5とリードしていたが、後半ではポルトガルと互角になった。
HYP: Italy was leading 16 to 5 at halftime, but in the second half, they were evenly matched with Portugal.
OUTPUT_RAW: {
 "overall_score": 95.0,
 "errors": [
  {
   "id": "E1",
   "error_type": "Fluency",
   "severity": "minor",
   "source_span": [0, 27],
   "target_span": [0, 44],
   "description": "The phrase '16 to 5' is understandable but '16-5' or '16 against 5' would be more natural in sports context."
  }
 ]
}

図 3　中間誤りメモ

You are a professional translator.
Translate the ENGLISH source into Chinese.
Use the ENGLISH text as the primary information source.
Consult the JAPANESE text only when the English expression is ambiguous or incorrect.

User:
Translate the following into Chinese.

ENGLISH (primary source):
<Intermediate English translation>

JAPANESE (auxiliary reference):
<Original Japanese sentence>

図 4　参照優先度を制御するプロンプト

System:
You are a professional translator.
Avoid repeating the translation errors indicated by AUTOMQM.

User:
Translate the following into Chinese.
Do NOT repeat the errors listed below.

AUTOMQM Error Notes (Do NOT repeat):
- <error_type> (<severity>): <description>
- <error_type> (<severity>): <description>

ENGLISH:
<Intermediate English translation>

図 5　再発防止プロンプト

System:
You are a professional translator.
Produce a fluent Chinese translation that follows standard Chinese writing conventions.

User:
Translate the following into Chinese.

Constraints:
- Write using standard Chinese characters.
- Do not include Japanese kana.
- Avoid unnecessary Latin characters.
- Use established Chinese terminology where applicable.
- Output only the final Chinese translation.

ENGLISH:
<Intermediate English translation>

図 6　制約条件プロンプト

第4章 実験環境

本章では，第 3 章で提案した AUTOMQM による誤り情報注入型ピボット翻訳手法の有効性を検証するために構築した実験環境について述べる．具体的には，評価に用いるデータセット，性能比較のために設定した翻訳手法，および評価指標を示す．本研究では，同一条件下で直接翻訳手法，ピボット翻訳手法，および提案手法を比較することで，言語資源量の異なる条件下における翻訳品質の差異と，中間翻訳誤りが最終翻訳に与える影響を検証する．

4.1 データセット

本研究では信頼性の高い参照訳を含む対訳データセットとして Asian Language Treebank（ALT）Project を用いた．ALT Project は，アジア言語間の機械翻訳および自然言語処理研究の促進を目的として構築された並列コーパスであり，高品質な文整列およびアノテーションを備えている．英語文を基準に複数のアジア言語が収録されており，全体で約 20,000 文規模の対訳データが提供されている．本研究では，ALT Project に収録されている言語のうち，日本語，英語，中国語（簡体字），インドネシア語，タイ語，ミャンマー語，クメール語の 7 言語を対象とした．翻訳設定としては，日本語を原言語，中間言語を英語とし，目的言語を中国語，インドネシア語，タイ語，ミャンマー語，クメール語の 5 言語とした．これらは，高資源言語（中国語），中資源言語（インドネシア語，タイ語），低資源言語（ミャンマー語，クメール語）を含んでおり，言語資源量の違いが翻訳品質および中間翻訳誤りの影響に与える差異を体系的に比較できる．実験では，日本語文を原文とする 100 文を抽出し，参照訳との比較に基づいて翻訳品質の評価を行った．

4.2 比較手法

翻訳手法の有効性を検証する上では，異なる翻訳戦略との比較を通じて，相対的な性能差を明らかにする必要がある．そこで本研究では，以下の三つの翻訳手法を比較対象として設定した．いずれの手法も，前節で述べたデータセットおよび翻訳設定を共通条件として用いた．一つ目は，日本語から各目的言語へ翻訳する直接翻訳手法である．本手法では，日本語原文のみを入力として翻訳生成を行い，中間言語を介さずに目的言語への翻訳を実施する．翻訳生成には，原言語

である日本語を目的言語へ直接翻訳するよう指示したプロンプトを与えた．このプロンプトでは，翻訳内容の正確性を維持しつつ，説明文や注釈を含まない最終的な翻訳結果のみを出力することを指示している．また，提案手法およびピボット翻訳手法と同様に，翻訳結果の形式的な不整合を防ぐため，日本語仮名の混入や過度なローマ字表記を抑制する制約条件を設定した．翻訳結果がこれらの条件を満たさない場合には，翻訳生成が十分に達成されていないと判断し，制約条件を強化したプロンプトを用いて最大 1 回に限り再生成を行う．ただし，本再生成処理は翻訳品質の向上を目的とした反復的な最適化ではなく，評価に用いる翻訳文として最低限の形式的妥当性を保証するための補助的手続きである．二つ目は，原言語→中間言語→各目的言語の二段階翻訳を行うピボット翻訳手法である．三つ目は，ピボット翻訳手法に加え，中間翻訳に対する AUTOMQM の誤り検出結果を最終翻訳に反映する提案手法である．ピボット翻訳手法と提案手法の相違点は，中間翻訳に対する誤り検出の有無，および誤り情報を最終翻訳生成に利用するか否かにある．これら三手法の比較により，提案手法がピボット翻訳手法および直接翻訳手法と比べて有効であるかを検証する．

4.3 評価指標

本研究では，翻訳品質を多角的に評価するため，評価観点ごとに異なる自動評価指標を併用した．具体的には，参照訳との表層的一致度および修正量に基づく従来の自動評価指標と，LLM を用いた意味・誤り分析に基づく自動評価手法を組み合わせて用いた．

まず，従来の自動評価指標として BLEU，chrF++および TER を用いた．BLEU および chrF++は，翻訳文と参照訳との文字列レベルの一致度に基づいて翻訳品質を評価する指標であり，翻訳品質の全体的な傾向を把握することを目的とする．一方，TER は，翻訳文を参照訳に変換するために必要な編集操作量を測定する指標であり，翻訳文の修正量や編集容易性を反映する．これらの指標は大規模な翻訳結果を効率的に比較する上で有用であるが，翻訳文の意味的妥当性や自然さを十分に捉えることは難しい．

そこで，本研究では，これらの指標では捉えきれない翻訳文の意味的妥当性や誤りの性質を評価するため，LLM を用いた自動評価手法として GEMBA-DA および AUTOMQM を導入した．先行研究において，これらの手法は参照訳を用いた評価設定の方が，より人手評価に近い精度の高い評価結果が得られることが

報告されている．GEMBA-DA は，原文・翻訳文・参照訳を入力とし，翻訳文全体の意味保持を人手評価に近い形式で単一のスコアとして出力する評価手法である．本研究では，GEMBA-DA が翻訳文全体を単一のスコアとして出力する点を踏まえ，局所的な誤りの分析ではなく，翻訳品質の全体的な傾向を把握するための補助的な指標として位置づけた．一方，AUTOMQM は，翻訳文に含まれる誤り情報に基づいて翻訳品質を数値として算出するとともに，誤りタイプや重大度に基づく詳細な誤り情報を出力する評価手法である．本研究における本節の目的は，中間翻訳に含まれる誤りが最終翻訳においてどの程度修正されているかを分析することである．このため，中間翻訳を介さない直接翻訳は評価対象に含めず，ピボット翻訳手法および提案手法のみを対象として評価を行った．まず AUTOMQM が出力する数値スコアを用いて翻訳品質の全体的な傾向を把握した．その上で，中間翻訳に含まれる誤りが最終翻訳においてどの程度修正されているかを，誤り単位で分析することで，翻訳過程における誤り修正の効果を詳細に検討した．

以上のように，本研究では，各評価指標が捉える翻訳品質の側面を明確に区別した上で複数の評価指標を併用することで，単一の数値や単一の評価観点に依存しない総合的な翻訳品質評価を行う．

第5章 評価

本章では，第 3 章で提案した誤り情報注入型ピボット翻訳手法の有効性を検証するために実施した自動評価および人手分析の結果について述べる．また，翻訳生成に用いる誤り情報として，すべての誤りを用いる場合と，重大度が高い誤りのみに限定する場合とを比較し，その違いが翻訳品質に与える影響を分析する．

5.1 従来の自動評価指標

本節では，自動評価指標として BLEU・chrF++・TER を用いて評価を行った結果を表 1 に示す．各言語について，第 4 章で説明したデータセットを参照訳とし，直接翻訳，ピボット翻訳手法，および提案手法の三手法を比較した．

はじめに，高資源言語である中国語では，BLEU・chrF++の両指標においてピボット翻訳手法が最も高いスコアを示した．一方，提案手法はこれらの指標ではピボット翻訳手法を上回らなかったが，TER においては最も低い値を示した．この結果は，中間翻訳に含まれる誤りを抑制する効果が，翻訳文の修正量の低減として一部反映された可能性を示している．

次に，中資源言語であるインドネシア語およびタイ語では，BLEU において三手法間の差は比較的小さかった．chrF++に着目すると，インドネシア語では提案手法が最も高い値を示した一方，タイ語では三手法間のスコア差が小さく，明確な優劣は確認されなかった．また，TER においてはインドネシア語で提案手法が最も低い値を示したものの，タイ語では顕著な改善には至らなかった．

最後に，低資源言語であるミャンマー語およびクメール語では，BLEU・chrF++のいずれにおいても，直接翻訳よりもピボット翻訳手法および提案手法が高いスコアを示した．しかし，TER に着目すると，ミャンマー語ではピボット翻訳手法が，クメール語では直接翻訳が最も低い値を示しており，提案手法がすべての指標において一貫して優位であるとは言えない．

表 1　自動評価指標

		BLEU	chrF++	TER
中国語	直接翻訳	36.74	34.53	57.84
	ピボット翻訳	**39.92**	**37.36**	56.29
	提案手法	39.01	36.54	**55.48**
インドネシア語	直接翻訳	**16.34**	56.10	68.03
	ピボット翻訳	15.51	55.65	68.80
	提案手法	16.06	**56.11**	**66.75**
タイ語	直接翻訳	**15.90**	**46.62**	**69.15**
	ピボット翻訳	15.18	45.07	71.34
	提案手法	15.03	45.33	72.13
ミャンマー語	直接翻訳	25.40	41.08	66.93
	ピボット翻訳	**26.82**	41.15	**64.19**
	提案手法	26.76	**41.69**	66.11
クメール語	直接翻訳	8.15	40.02	**84.10**
	ピボット翻訳	8.42	**40.93**	85.81
	提案手法	**8.52**	40.60	85.40

5.2 LLM を用いた自動評価手法

本節では，LLM を用いた自動評価手法による評価結果について述べる．

5.2.1 GEMBA-DA による自動評価

まず，LLM を用いた自動評価手法である GEMBA-DA による評価結果とその考察について述べる．評価に用いたプロンプトは図 7 に示すとおりであり，原文・翻訳文・参照訳を入力として，翻訳品質を単一の数値で出力する構成となっている．

GEMBA-DA による評価結果を表 2 に示す．高資源言語である中国語では，三手法間で大きな差は確認されず，いずれの翻訳手法においても高い評価が得られた．これは，高資言語において翻訳品質が全体的に安定しているためであると考えられる．一方，中資源言語であるインドネシア語および低資源言語であるミャンマー語では，直接翻訳が最も高い評価を示した．ただし，三手法の差はいずれも小さく，顕著な性能差は確認されなかった．ミャンマー語においては，提案手法がピボット翻訳手法を上回る結果も見られた．これに対し，中資源言語であるタイ語および低資源言語であるクメール語では，提案手法が最も高い評価を示した．これらの結果は，中間翻訳段階で生じやすい意味的な誤りや解釈のずれを明示的に認識し，翻訳生成に反映させる提案手法が，低資源条件においても翻

訳文全体の意味保持の改善に寄与する可能性を示している．以上の結果から，GEMBA-DA に基づく評価では，言語や資源量に応じて最も高い評価を示す翻訳手法は異なるものの，全体として各手法間の差は比較的小さかった．また，評価プロンプトは図 5 に示すとおり原文・翻訳文・参照訳を入力とする構成であるが，評価基準は一定の抽象性を含むため，得られたスコアは絶対値としてではなく，相対比較の指標として解釈することが適切である．このため，本研究では GEMBA-DA の結果を翻訳品質の参考指標として位置づける．

```
Score the following translation from {source_lang} to {target_lang} with respect to
human reference on a continuous scale 0 to 100 where score of zero means "no
meaning preserved" and score of one hundred means "perfect meaning and grammar"
```

図 7　GEMBA-DA のプロンプト

表 2　GEMBA-DA の結果

		GEMBA-DA
中国語	直接翻訳	**91.38**
	ピボット翻訳	90.84
	提案手法	**91.38**
インドネシア語	直接翻訳	**91.14**
	ピボット翻訳	90.79
	提案手法	89.95
タイ語	直接翻訳	90.19
	ピボット翻訳	89.05
	提案手法	**90.27**
ミャンマー語	直接翻訳	**88.96**
	ピボット翻訳	86.89
	提案手法	88.10
クメール語	直接翻訳	86.89
	ピボット翻訳	85.67
	提案手法	**86.90**

5.2.2 AUTOMQM による自動評価

本節では，LLM を用いた自動評価手法である AUTOMQM による評価結果とその考察について述べる. 評価に用いたプロンプトは図 8 に示すとおりである.

AUTOMQM による評価結果を表 3 に示す．中間翻訳に対する AUTOMQM スコアは，全言語において同一である．これは，日英翻訳を両手法で共通の翻訳結果として使用しているためである. 最終翻訳に対する AUTOMQM スコアを見ると，高資源言語である中国語および中資源言語であるインドネシア語では，提案手法がピボット翻訳手法を上回る結果は確認されなかった．一方，中資源言語であるタイ語では，提案手法がピボット翻訳手法を上回り，低資源言語であるミャンマー語およびクメール語においても，同様に提案手法が高い評価を示した．この結果から, 提案手法は, 低資源言語条件下で有効的な手法であると考えられる. しかしながら，全体として見ると，AUTOMQM に基づく自動評価は，提案手法が全ての言語において一貫してピボット翻訳を上回る結果は得られなかった.

```
＜System プロンプト＞
You are an expert translation evaluator following the MQM (Multidimensional Quality Metrics) spirit.
Evaluate the translation quality precisely and return a strict JSON with the requested fields only.

＜User プロンプト（参照訳あり / WITH_REF）＞
Score the quality of translation from {src_lang} to {tgt_lang} with respect to the human reference.
Return STRICT JSON with the following schema (no extra text):

{
 "overall_score": <float 0-100>,
 "errors": [
  {
   "id": "E1",
   "error_type": "Mistranslation|Omission|Addition|Grammar|Fluency|Terminology|Style|Other",
   "severity": "minor|major|critical",
   "source_span": [start_index, end_index]
   "target_span": [start_index, end_index]
   "description": "short explanation"
  }
 ]
}

Constraints:
- "overall_score" MUST be a number (0-100). Use decimals allowed, e.g., 86.5.
- "errors" can be empty [] if no errors are found.
-  Keep explanations concise.

{src_lang} source: "{source}"
{tgt_lang} human reference: "{reference}"
{tgt_lang} translation: "{translation}"
```

図 8　AUTOMQM のプロンプト

表 3　AUTOMQM の結果

		AUTOMQM（中間翻訳）	AUTOMQM（最終）
中国語	ピボット翻訳	92.54	**91.41**
	提案手法		91.27
インドネシア語	ピボット翻訳		**90.45**
	提案手法		90.01
タイ語	ピボット翻訳		89.01
	提案手法		**90.01**
ミャンマー語	ピボット翻訳		85.56
	提案手法		**87.45**
クメール語	ピボット翻訳		85.53
	提案手法		**86.48**

5.2.3 AUTOMQM の出力に対する人手判断

前節までの AUTOMQM による自動評価では, 提案手法の効果が必ずしも一貫して確認されなかった．そこで本節では，自動評価を補完する目的で，AUTOMQM の出力内容に対する人手評価を行い，提案手法の効果をより詳細に検証する．

まず，中間翻訳に対して AUTOMQM が出力した誤り指摘の妥当性を確認した．AUTOMQM は自動評価手法であるため，その出力には一定の誤判定が含まれる可能性がある. そこで, 本研究では, 中間翻訳文, 参照訳, および AUTOMQM の誤り指摘内容を参照し，指摘内容が参照訳と一致している場合を「正しい指摘」，参照訳と中間翻訳文が一致している場合を「誤った指摘」として人手で検証した．その結果，中間翻訳に対する誤り指摘は全 104 件存在し，そのうち 23 件が誤判定であることが確認された．例えば，人名表記に関する指摘において，中間翻訳および参照訳が一致しているにもかかわらず，AUTOMQM が誤りとして判定する例が確認された．このような誤った誤り情報が翻訳生成に用いられると，かえって翻訳品質を低下させる可能性がある．そこで，提案手法の効果を正確に評価するため，これらの誤判定 23 件を除外し，残り 81 件のみを対象として再評価を行った．その上で，再度 AUTOMQM による 0-100 スケールの自動評価を実施した．その結果は，表 4 に示す通り，すべての対象言語において，提案手法はピボット翻訳手法を上回る性能を示した．特に，言語資源量が少ない言語ほど両手法の差が大きく，提案手法が低資源言語翻訳において有効であることが確認された．

表 4　AUTOMQM の誤判定を除いた結果

		AUTOMQM（中間翻訳）	AUTOMQM（全部）	AUTOMQM（ミス抜き）
中国語	ピボット翻訳	92.54	**91.41**	91.4
	提案手法		91.27	**91.5**
インドネシア語	ピボット翻訳		**90.45**	90.4
	提案手法		90.01	**90.7**
タイ語	ピボット翻訳		89.01	89.2
	提案手法		**90.01**	**90.1**
ミャンマー語	ピボット翻訳		85.56	86.0
	提案手法		**87.45**	**87.5**
クメール語	ピボット翻訳		85.53	84.9
	提案手法		**86.48**	**86.7**

次に，中間翻訳で AUTOMQM により検出された誤りが，最終翻訳において実際に修正されているかを人手で分析した．AUTOMQM の出力には誤判定や未検出が含まれる可能性があるため，本分析では，最終翻訳の AUTOMQM の指摘結果のみに依存せず，中間翻訳に対する指摘内容と，提案手法およびピボット翻訳手法による最終翻訳文を直接照合し，誤りが最終翻訳に残っているかどうかを確認した．確認には，Google 翻訳などの既存の機械翻訳システムや，ChatGPT，Gemini などの LLM を補助的に用いた．以上の手順に基づいて分析を行い，中間翻訳で検出された誤りが最終翻訳に残っていない場合を「修正可能」，残っている場合を「修正不可能」と定義して分類した．その結果，表 5 で示す通り，すべての対象言語において，ピボット翻訳手法と比較して，提案手法の方が多くの誤りを修正できていることが確認された．この結果から，すべての言語に対して，提案手法が有効であると考えられる．

表 5　修正可能数

	日英のエラー数	提案（修正可能数）	従来（修正可能数）
中国語	81	**60**	47
インドネシア語	81	**62**	32
タイ語	81	**57**	30
ミャンマー語	81	**59**	33
クメール語	81	**52**	32

さらに，修正可能・修正不可能の結果に基づき，AUTOMQM が付与したエラータイプ（Mistranslation，Omission，Fluency，Terminology，Style）別の分析を行った．なお，AUTOMQM が付与するエラータイプは誤りの性質に基づいて分類される．本研究では，原文の意味が正確に反映されていない誤りを Mistranslation，原文情報の欠落を Omission とする．また，句読点や綴り，文法，文体など翻訳文の流暢さや表記に関する誤りを Fluency，用語の不適切または不一貫な使用を Terminology と定義する．さらに，内容は正しいものの表現が不自然な誤りを Style として扱う．その結果を表 6 に示す．

Mistranslation については，高資源言語および中資源言語ではピボット翻訳手法の方が修正可能数が多かった一方，低資源言語では提案手法の方が多くの誤りを修正できていた．提案手法で改善できなかった例としては，原文自体の語解釈に起因する誤りが挙げられるが，固有名詞の不正確さや場所表現の曖昧さ，文構造の不自然さといった誤りは改善される例が確認された．

Omission については，全体の件数が少なく大きな差は見られなかったものの，提案手法の方が修正可能数が多い傾向を示した．特に，人物名や時点・関係を示す情報の欠落は，提案手法によって修正される例が確認された．

Fluency では，すべての対象言語において提案手法の修正可能数がピボット翻訳手法を上回っており，提案手法の効果が最も明確に確認された．一方で，不適切な言い回しや冗長な表現が残る例も見られたが，語順や構文の分かりにくさ，表現の曖昧さといった問題は多くの場合で改善されていた．

Terminology においても，すべての対象言語で提案手法の修正可能数が従来手法を上回っていた．提案手法で改善できなかった例としては，語彙選択がやや不自然な場合や，原文に存在しない性別情報の付加が挙げられる一方，不自然な専門用語の選択や固有名詞の誤用，冗長な用語表現は改善される傾向が確認された．

Style については，高資源言語を除く言語において提案手法の方が修正可能数が多かった．情報配置や強調の仕方が不自然な場合や直訳的な表現が残る例も見られたが，不自然な固有名詞表現や冗長な言い回しについては改善が確認された．

これらの結果を修正可能率の観点から整理すると，本手法により，Fluency および Terminology において，修正率を約 10〜40%向上した．また，Omission についても，一部言語において修正率を約 50%〜75%向上した．さらに，低資源言

語ほど修正可能率の増加幅が大きい傾向が見られ，低資源言語条件下で本手法が特に有効であることが示された．さらに，エラータイプごとの差の統計的有意性を確認するため，修正可能・修正不可能の分布に対してカイ二乗検定を適用した（表 7）．その結果，Fluency では中資源言語および低資源言語において，Terminology では中資源言語および一部の低資源言語において有意差が確認された．一方，Mistranslation，Omission，Style については，多くの言語で有意差は確認されなかった．

表 6　エラータイプ分類

			Mistranslation	Omission	Fluency	Terminology	Style
中国語	ピボット翻訳	修正可能	**8**	2	22	12	**7**
		修正不可能	2	2	5	19	2
	提案手法	修正可能	7	**3**	**25**	**14**	5
		修正不可能	3	1	2	17	4
インドネシア語	ピボット翻訳	修正可能	4	2	14	9	2
		修正不可能	6	2	13	22	7
	提案手法	修正可能	**7**	**3**	**23**	**21**	**6**
		修正不可能	3	1	4	10	3
タイ語	ピボット翻訳	修正可能	3	2	14	9	2
		修正不可能	7	2	13	22	7
	提案手法	修正可能	**6**	**3**	**24**	**20**	**5**
		修正不可能	4	1	3	11	4
ミャンマー語	ピボット翻訳	修正可能	5	2	11	4	3
		修正不可能	5	2	16	27	6
	提案手法	修正可能	**8**	**3**	**22**	**21**	**5**
		修正不可能	2	1	5	10	4
クメール語	ピボット翻訳	修正可能	5	2	13	10	2
		修正不可能	5	2	14	21	7
	提案手法	修正可能	**7**	2	**22**	**15**	**6**
		修正不可能	3	2	5	16	3

表 7　カイ二乗検定

	Mistranslation	Omission	Fluency	Terminology	Style
中国語	0.60	0.46	0.22	0.60	0.31
インドネシア語	0.17	0.46	**0.008**	**0.002**	0.05
タイ語	0.17	0.46	**0.002**	**0.005**	0.14
ミャンマー語	0.15	0.46	**0.002**	**0.00001**	0.34
クメール語	0.36	1	**0.01**	0.19	0.05

以上の結果から，自動評価のみでは一貫して確認できなかった提案手法の効果が，人手評価を通じて明確に示された．特に，Fluency および Terminology といった表現品質に関わる誤りに対して，提案手法が有効であることが統計的にも裏付けられた．一方で，原文解釈や文脈理解に強く依存する誤りについては，誤り情報の提示のみでは十分な改善が困難である場合があることも示唆された．これらの結果は，AUTOMQM の出力には一定の限界が存在するものの，誤り情報を適切に選択・活用することで，誤り情報注入型ピボット翻訳手法の性能を効果的に向上させられる可能性が示された．今後は，AUTOMQM に限らず，中間翻訳の誤りをより高精度に捉える他の誤り検出手法や評価手法を導入することで，さらなる性能向上が期待される．

5.3 注入する誤り情報の比較

前節の分析結果を踏まえ，AUTOMQM によって検出される誤りには，翻訳品質への影響度に応じて critical，major，minor の三段階の重大度が付与される．AUTOMQM において，critical 誤りは翻訳文の意味理解を著しく損なう重大な誤り，major 誤りは意味理解に影響を与える可能性の高い誤りとして位置付けられている．一方，minor 誤りは意味理解に致命的な影響を与えない表現上や文体上の軽微な問題を含む場合が多い．すべての誤り情報を一律に翻訳生成時の制約として与えると，翻訳モデルに過剰な制約を与え，意味的に重要な誤りへの対応が十分に行われない可能性がある．そこで本章では，提案手法の有効性をより詳細に分析するため，AUTOMQM が出力する誤りのうち，重大度が major と判定された誤りのみを中間誤り情報として用いる誤り情報注入型ピボット翻訳手法を比較条件として検討する．本研究で対象とした中間翻訳結果においては critical 誤りが確認されなかったため，本比較手法では major 誤りのみに着目する．これにより，誤り情報の量が翻訳品質に与える影響を検証し，誤り情報注入型ピボット翻訳手法において有効な誤り情報の範囲を明らかにすることを目的とする．

本節では，AUTOMQM による誤り情報注入型ピボット翻訳において，誤り情報をすべて提示する手法と，major 誤りのみに限定して翻訳生成に反映する手法の比較を行う．表 8 に示す結果から，翻訳生成に反映させる誤り情報の範囲が翻訳品質に与える影響は，言語資源量によって異なることが確認された．まず，高資源言語である中国語では，BLEU，chrF++，TER の各指標において，major

誤りのみに限定した手法が高い評価を示した. この結果から, 高資源条件下では, 重大度の高い誤りに絞った制御によって翻訳品質を十分に維持・向上できることが示唆される．一方，GEMBA-DA では誤り情報をすべて提示した手法が高い評価を示しており，軽微な誤りの補正が文全体の意味的評価に影響を与える可能性がある．次に，中資源言語であるインドネシア語およびタイ語では，評価指標によって優位となる手法が異なり，一貫した傾向は確認されなかった．インドネシア語では，BLEU，chrF++，TER においては誤り情報をすべて提示した手法が良好であった一方，GEMBA-DA および AUTOMQM では major 誤りのみに限定した手法が高い評価を示した．一方，タイ語では，BLEU，chrF++，TER，GEMBA-DA の多くの指標において major 誤りのみに限定した手法が安定して高い評価を示しており，過度な誤り制御を避けることが有効である可能性が示された．最後に，低資源言語であるミャンマー語およびクメール語では，すべての指標において，誤り情報を限定せずに提示した手法が高い評価を示した．このことから，低資源条件下では，中間翻訳に多様な誤りが含まれやすく，軽微な誤りを含めた包括的な誤り情報の活用が翻訳品質の向上に有効であると考えられる．

以上より，AUTOMQM による誤り認識型ピボット翻訳では，すべての言語に対して一律に誤り情報の粒度を設定するのではなく，言語資源量や評価指標の特性を考慮し，major 誤りのみに限定するか，あるいは誤り情報をすべて活用するかを切り替えることが重要であることが示された．

表 8　注入した誤り情報の比較結果

		BLEU	chrF++	TER	GEMBA-DA	AUTOMQM
中国語	提案手法（全部）	39.01	36.54	55.48	**91.38**	91.27
	提案手法（majorのみ）	**40.57**	**37.93**	**54.90**	89.85	**91.53**
インドネシア語	提案手法（全部）	**16.06**	**56.11**	**66.75**	89.95	90.01
	提案手法（majorのみ）	15.84	56.10	67.62	**90.35**	**90.64**
タイ語	提案手法（全部）	15.03	45.33	72.13	90.27	**90.01**
	提案手法（majorのみ）	**15.30**	**45.55**	**71.65**	**90.62**	89.56
ミャンマー語	提案手法（全部）	**26.76**	**41.69**	**66.11**	**88.10**	**87.45**
	提案手法（majorのみ）	26.33	41.30	72.97	87.27	86.23
クメール語	提案手法（全部）	**8.52**	**40.60**	**85.40**	**86.90**	**86.48**
	提案手法（majorのみ）	8.38	40.16	86.45	86.67	86.38

第6章 母国語話者による人手評価

本章では，第 3 章で提案した誤り情報注入型ピボット翻訳手法の有効性を検証するため，母語話者または高い運用能力を有する評価者による人手翻訳品質評価を実施する．前章までで用いた自動評価手法は，翻訳品質を定量的に比較する上で有用である一方で，翻訳文の自然さや意味の分かりやすさといった，人間の読解に基づく評価を十分に反映できない場合がある．そこで本章では，人手評価を通じて，自動評価では捉えきれない翻訳品質の側面を補完的に分析する．本章では，まず人手評価の方法を説明し，その後，得られた評価結果に基づいて提案手法の有効性を考察する．

6.1 評価方法

本節では，本研究で実施した人手評価の方法について述べる．評価対象文は，第 5 章で用いたデータセットから抽出した同一の日本語原文に対する翻訳結果であり，直接翻訳，ピボット翻訳手法，および提案手法の三手法によって生成された翻訳文を対象とした．人手評価は，中国語，インドネシア語，タイ語の各言語に対して実施し，すべての言語で同一の評価方法および同一の 20 文を用いた．これら三言語を評価対象としたのは，各言語について母語話者または高い運用能力を有する評価者を確保でき，人手評価の信頼性を担保できたためである．評価における先入観を防ぐため，評価者には翻訳手法の種類を明示せず，三手法の翻訳結果をランダムな順序で提示した．評価者は，各原文に対して提示された三つの翻訳文を比較し，翻訳品質が高いと感じた順に順位付けを行った．翻訳品質が同程度であると判断された場合には，同順位として評価することを許容した．評価の観点は，翻訳文の自然さおよび原文の意味が適切に保持されているかの二点とした．また，各翻訳文について，評価理由を自由記述形式で記載するよう求めた．評価は，対象言語を母語とする，または高い運用能力を有する留学生および教員に依頼し，事前に評価基準および手順を説明した上で実施した．

6.2 人手評価の結果・考察

表 9 に示す人手評価の結果から，全体として提案手法は多くの文において高い評価を得る傾向が確認された．特に，中国語およびタイ語では，提案手法が 1 位と評価される文が最も多く，人間の読解に基づく評価において一定の有効性

が示唆された. 評価理由としては,「自然な文章である」「表現が洗練されている」「原文の意味を正しく反映している」といった肯定的な意見が多く挙げられている．一方，インドネシア語では，順位付けの結果として直接翻訳が高く評価される文が比較的多く見られた．しかし，自由記述による評価理由を分析すると，提案手法に対して「自然な文章である」といった評価が多く与えられており，自然性の面では一定の評価を得ていることが確認された. 以上の結果から，本研究で提案した誤り情報注入型ピボット翻訳手法は，中間翻訳に含まれる誤り情報を活用することで，翻訳文の不自然さや意味のずれを抑制し，自然な表現を生成するという点において，人手評価の観点から一定の効果を示したと考えられる．ただし，順位データに対してフリードマン検定を適用した結果，全ての言語で，p 値が 1 という結果になり，手法間の差について統計的に有意な差は確認されなかった．この要因の一つとして，本研究で用いた評価データセットが主にニュース記事由来の文で構成されており，文体や構文が比較的定型的である点が挙げられる．このような文では翻訳手法間の差異が表れにくく，順位付けにおけるばらつきが抑えられた可能性がある．今後，より文脈依存性や多義性を含む文章を対象とした評価を行うことで，提案手法の効果がより明確に示される可能性がある．

表 9　人手評価の結果

		1位	2位	3位
中国語	直接翻訳	**10**	7	3
	ピボット翻訳	8	7	5
	提案手法	8	3	9
インドネシア語	直接翻訳	11	6	3
	ピボット翻訳	8	8	4
	提案手法	**13**	1	6
タイ語	直接翻訳	**12**	6	2
	ピボット翻訳	7	4	9
	提案手法	9	4	7

第7章 おわりに

本研究では，ピボット翻訳手法の誤り伝播問題に対処するために，LLM を用いてピボット翻訳における中間翻訳の誤りを明示的に扱い，最終翻訳の品質向上を図る手法を提案した. 具体的には, 中間言語に対して AUTOMQM を適用し, 中間翻訳で発生した誤りの種類, 重大度, および発生箇所といった誤り情報を抽出し，これらの誤りが最終翻訳に反映されないよう翻訳生成を制御することで, 特に低資源言語における翻訳品質の向上を目指した．本研究の貢献は以下の通りである．

複数情報源を適切に参照させるプロンプト設計

提案手法では, 中間翻訳文を主たる情報源, 原文を曖昧性解消の補助的情報源として位置付け, さらに AUTOMQM により抽出された誤り情報を再発防止のための制約として与えることで, 翻訳生成を制御した．その結果，中間翻訳で検出された誤りの多くが最終翻訳段階で修正されることが確認され，中間翻訳に起因する誤り伝播を抑制できることが示された．また, 言語資源量が少ない言語ほど, 提案手法とピボット翻訳手法との差が大きく, 低資源言語翻訳において本手法が特に有効であることが確認された．

翻訳品質評価指標の統合

翻訳品質を多角的に捉えるための統合的な評価を行い，提案手法の有効性を詳細に分析した点が挙げられる．評価の結果，不自然な表現や文体上の問題，用語の不正確さ，および情報抜けといった実用上重要な誤りに対して，一貫した改善効果が確認された．また，ピボット翻訳手法と比較すると，低資源言語条件下において誤り削減効果がより顕著であり，中間翻訳に含まれる誤りを明示的に扱うことが，ピボット翻訳の品質向上に有効であることが示唆された．

本研究では, 中間翻訳の誤り検出手法として AUTOMQM の出力に基づいて検討を行ったが, 誤り検出の精度や網羅性には依然として改善の余地がある．このため，今後の課題としては，AUTOMQM に限らず，中間翻訳の誤りをより高精度に捉える他の誤り検出手法や評価手法を導入することが挙げられる．これにより, 誤り情報の信頼性向上や適用範囲の拡張が期待され，さらなる翻訳性能の向上につながると考えられる．

謝辞

本研究の遂行にあたり、多大なるご指導ならびに貴重なご助言を賜りました、指導教官の村上陽平教授、Mondheera Pituxcoosuvarn 講師に心より感謝申し上げます. 加えて、研究を進める上で多くの示唆をいただいた富田悠斗先輩にも深く御礼申し上げます. さらに、日頃より有益な議論や温かいご支援をいただいた社会知能研究室の皆さまに、厚く感謝の意を表します.

参考文献

[1] Constantine Lignos，Nolan Holley，Chester Palen-Michel and Jonne Saleva：“Toward More Meaningful Resources for Lower-resourced Languages”，*Findings of the Association for Computational Linguistics: ACL 2022*，pp.523-532(2022).

[2] Hua Wu and Haifeng Wang：“Revisiting Pivot Language Approach for Machine Translation”，*Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*，pp.154-162(2009).

[3] 三浦　明波，Graham Neubig，Sakriani Sakti，戸田　智基，中村　哲：“中間言語情報を記録するピボット翻訳手法”,自然言語処理，Vol.23，No.5，pp.499-528(2016).

[4] Rie Tanaka，Yohei Murakami，Toru Ishida：“Context-Based Approach for Pivot Translation Services”，*Proceedings of International Joint Conference on Artificial Intelligence(IJCAI-09)*，pp.1555-1561(2009).

[5] Kishore Papineni，Salim Roukos，Todd Ward and Wei-Jing Zhu：“BLEU : aMethod for Automatic Evaluation of Machine Translation”，*Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*，pp.311-318(2002).

[6] Ananya Mukherjee, Manish Shrivastava：“chrF-S : Semantics is All You Need”，*Proceedings of the Ninth Conference on Machine Translation*，pp.407-474(2024).

[7] Tom Kocmi and Christian Federmann：“Large Language Models Are State-of-the-Art Evaluators of Translation Quality”，*proceedings of the 24th Annual Conference of the European Association for Machine Translation*，pp.193-203(2023).

[8] Oatrick Fernandes，Daniel Deutsch，Mara Finkelstein，Parker Riley，AndreF.T.Martins, Graham Neubig,Ankush Garg, Jonathan H.Clark, Markus Freitag，Orhan Firat：“The Devil is the Errors: Leveraging Large Language Model for Fine-grained Machine Translation Evaluation”，*Proceedings of the Eighth Conference on Machine Translation*，pp.1066-1083(2023).

付録

AUTOMQM を用いて行った減点法の結果を掲載する.

A.1 減点法

critical を-25，major を-5，minor を-1 で減点法を行った.

		critical	major	minor	mqm_score
日本語→英語	中間翻訳	0	3	101	**-1.16**
中国語	直接翻訳	0	16	183	**-2.63**
	ピボット翻訳	0	17	182	-2.67
	提案手法	0	20	178	-2.78
インドネシア語	直接翻訳	0	19	173	**-2.68**
	ピボット翻訳	0	21	188	-2.93
	提案手法	2	24	179	-3.49
タイ語	直接翻訳	1	36	187	-3.92
	ピボット翻訳	0	34	194	-3.64
	提案手法	0	22	185	**-2.95**
ミャンマー語	直接翻訳	0	46	180	**-4.10**
	ピボット翻訳	0	61	185	-4.90
	提案手法	0	51	178	-4.33
クメール語	直接翻訳	0	62	167	**-4.77**
	ピボット翻訳	0	64	162	-4.82
	提案手法	0	61	172	**-4.77**