ISSN 1883-1818
No.83
December 2025

AAMT Journal

Asia-Pacific Association for Machine Translation

機 械 翻 訳

AAMT Journal
Asia-Pacific Association for Machine Translation 機械翻訳
No.83

目　次

C O N T E N T S

巻頭言

ことばの自由さ

須藤 克仁
奈良女子大学

筆者はこの 2 年ほどの間に専門である自動同時音声翻訳や機械翻訳に関する講演をする機会を幾度かいただいた。一般向けあるいは専門家向けに、機械翻訳研究者の目線からではあるが、翻訳がいかに難しく、また面白い問題であるかをお伝えしてきた（つもりである)。ことば、というのが十分深い問題であるというところ、それを別の言語で伝えるということがいかに深く、また広いか、そして創造的な営みであるか。本誌をご覧の方々であればそれぞれさまざまなお考えをお持ちのことであろう。ここでは筆者の私見を述べたい。

機械翻訳は大規模言語モデルの登場により過去最大の変革期を迎えたと言える。プロンプトでさまざまな制約条件を柔軟に追加できる上、形式上は長い入出力が扱え、文脈情報もある程度考慮されるようになったことから、文章をまるごと入力して機械翻訳することのハードルは大きく下がった。依然大規模言語モデル由来の課題は残るものの、今後も大規模言語モデルに基づく機械翻訳の利用は広がり続けるだろう。

一方で今後危惧されるのはことばの表現の平均化、あるいは画一化である。人が発することばの多様性は人の多様性でもあり、創造性にも繋がっている。大規模言語モデルをはじめとするいわゆる生成 AI は基本的には与えられた条件下でもっともらしい結果を得るものであり、人間の創造する営みを支援する存在ではあるが、多様性や創造性を拡張し得る存在ではない。生成 AI の便利さゆえに生成 AI に依存した創作が増加すれば、それをまた生成 AI が学習することで平均化や画一化が加速しかねない。便利なものは便利なものとして活用しつつも、独自の創造という「砦」は守られるべきものであろう。

では、ことばの多様性や創造性とは何であろうか。もちろん情報伝達のための記号としてのことばも書き手や話し手ごとにさまざまな違いがあるが、文字なら書き方、音声なら話し方もまた多様性や創造性の範疇であろう。そうした書き手や話し手の表出する記号以上の情報はコミュニケーションを豊かにし、同じ記号を基にしていても多様性や創造性が生まれる。しかしながら、機械翻訳では多くの場合ことばは計算機可読な記号として扱われ、書き方や話し方を積極的に反映させることはまだ主流とは言えない。書き手や話し手の意図を正しく伝達するためには、いわゆる字面の翻訳を超え、原言語と目的言語の知恵を総動員して適切なことばをひねり出す必要がある。近代における西洋の概念の輸入が新しいことばを生み出したように、ことばは静的なものでなく日々新しくなっていく。そうしたことばの営みを支えているのはことばの多様性や創造性、あるいは自由さと呼んでも良いかもしれない。その自由さは画一的なものへの収斂を是とすればたちまち失われてしまう。世界に存在する膨大な言語が徐々に喪われつつあることも無関係ではない。

大規模言語モデルは多言語を扱っているように見えて、言語の違いも含めた真の言語の多様性を扱いきれているのかは定かでない。大規模言語モデルが扱っているのは形を変えた Lingua Franca なのかもしれず、人のことばに影響を与え続けることは避けられないだろう。そうした中で我々のことばは呑み込まれてしまうのだろうか？あるいは引き続き自由に成長を続けていけるのだろうか？利便性一辺倒でなく、ことばの自由さを大切に守っていきたいものである。

Freedom of Language
Katsuhito Sudoh
Nara Women’s University

第 20 回 AAMT 長尾賞受賞記念

ユニバーサルコミュニケーションと機械翻訳

光安 渉
TOPPAN 株式会社 情報コミュニケーション事業本部

1. 謝辞

この度は、AAMT 長尾賞という大変名誉ある賞を賜り、誠に光栄です。

多くの皆様にご指導、ご支援いただき、このような素晴らしい賞を頂戴することができましたことを、開発を担当した弊社メンバーをはじめ、関係者の皆様方へ、心より感謝いたします。

2. ユニバーサルコミュニケーションと機械翻訳

“ユニバーサルコミュニケーション”と聞いて、翻訳アプリケーションや同時通訳システムを思い浮かべるでしょうか。

国籍や年齢、障がいなど、誰もがストレス無くコミュニケーションできること。それがユニバーサルコミュニケーションですが、現状では、「機械翻訳技術」が重視されているように感じています。

機械翻訳で『言葉の壁』を超えるだけが“ユニバーサルコミュニケーション”ではありません。『言語の壁』以外にも「聞こえにくい」、「見えにくい」、「話しにくい」などの課題を解決できるのが本当の意味での“ユニバーサルコミュニケーション”なのです。

3. VoiceBiz® UCDisplay®の開発経緯

弊社は VoiceBiz® UCDisplay®を様々な利用者に対して「誰もが分け隔てなくコミュニケーションを取れるように」という想いを込めて開発しました。

VoiceBiz® UCDisplay®は、VoiceBiz®という翻訳アプリケーションが前身にあります。2012 年に 5 か年計画で行った NICT の委託研究「自治体窓口向け音声翻訳サービスの研究」※1 の成果を社会実装したのが、VoiceBiz®なのです。VoiceBiz®は現在、多くの自治体や企業に導入して頂いており、海外の方とのコミュニケーションに役立っています。

しかし「相手が目の前にいるのに、目線が手元のタブレットやスマートフォンに行きがちで、コミュニケーションに不自然さを感じる」という、利用者からの声がありました。“目は口ほどに物を言う“という諺があるように、「相手の目を見て話す」ことはコミュニケーションにおいて重要なことなのです。

また、利用者から「耳の不自由な者との方とコミュニケーションで使う事は出来ませんか？」という問い合わせも、数多く頂きました。ユニバーサルコミュニケーションを実現にするには、VoiceBiz®では不十分なのではないかと感じていた時、世間では“COVID-19”が猛威を振るい、受付窓口ではパーティションが設置され、街中ではほとんどの人がマスクするという状況になりました。その時に「パーティションで声が聞こえにくい」、「口元が見えないから何を言っているか読み取れない」といった耳の不自由な方の声を聞きました。

そのような状況下でも、“誰とでもコミュニケーションが取れる窓口“。これを実現させるにはどうしたらよいかを考えていたところ、透明ディスプレイの存在を知り、「これをパーティションの代わりにしてそこに文字が出れば声が聞こえにくくても、コミュニケーションが取れるのではないか」という考えに至りました。

これにより翻訳サービスの VoiceBiz®をベースにアプリケーションをカスタマイズし、タブレット端末と透明ディスプレイ、それと高指向性マイクを組み合わせて VoiceBiz® UCDisplay®の開発に成功しました。

Universal Communication and Machine Translation
Wataru Mitsuyasu
Information & Communication Business Division, TOPPAN Inc.

4. VoiceBiz® UCDisplay®の活用

VoiceBiz® UCDisplay®の活用の用途は、大きく外国人支援と耳の不自由な方の支援に分かれています。外国人支援に重点に導入されているのは「鉄道の駅窓口・切符売り場」「ホテルのフロント」「スキー場・水族館などのレジャー施設」「百貨店・ショッピングモールなどの総合案内」また最近では「レンタカーの受付」など。

更には耳の不自由な方の支援目的では自治体の福祉課や、障がい者雇用を行っている企業において、面談や作業指示などで活用されています。

また、EXPO 2025 大阪・関西万博でも案内所、アクセシビリティセンター、メディアセンターに設置されており、様々に利用されています。

導入事例を１つ紹介します。長野県にある「歴史の宿 金具屋」は映画の舞台のモデルとしても知られる情緒あふれる老舗旅館で、外国人観光客の増加に伴い、言葉の壁を乗り越えることが、おもてなしの重要な課題となっていました。

今までは小型の端末型翻訳サービスを導入していましたが、操作が複雑で、旅館特有の言い回しや、日本の文化に根ざした表現がうまく伝わらず、意図しない訳になってしまうことが多々あり、これではせっかくの『おもてなしの心』が十分に伝わらないというジレンマを抱えていたとのことです。

VoiceBiz® UCDisplay®の導入後は、旅館での過ごし方や交通案内といった、お客さまが安心して滞在するための情報提供をスムーズに伝えることができるようになり、また、アレルギーや病気の名前など、お客さまの健康に関わるようなデリケートな内容のコミュニケーションにおいても非常に重宝し、「お客さまに大きな安心感を提供できるようになった」とのお声を頂いています。

5. おわりに

11 月開催の、東京で初めてのデフリンピックでは、「聞こえない・聞こえにくい方」とのコミュニケーションを支援する様々なサービスが活用される予定です。音声や文字だけではなく、手話も“一つの言語”として重要なコミュニケーション手段です。

VoiceBiz® UCDisplay®は手話にはまだ対応出来ていませんが、将来的には様々なパートナーとの共創で、手話にも対応したユニバーサルコミュニケーションサービスを開発・展開していて行ければと考えています。

デフリンピックをきっかけに、ユニバーサルコミュニケーションサービスの認知が広がり、近い将来には、利用者が誰とでもコミュニケーション可能なユニバーサルな窓口が、当たり前のように設置されている社会になることを期待するとともに、弊社もその一助になるべく今後も活動していく所存です。

※１国立研究開発法人情報通信研究機構の委託研究 自治体向け音声翻訳システムに関する研究開発 1800101

第 20 回 AAMT 長尾賞受賞記念

英文ニュース作成への機械翻訳導入と LLM を用いた自動校正の開発

川上貴之、林直也、朝賀英裕
株式会社時事通信社

1. はじめに

このたびは、栄えある第20回AAMT長尾賞を賜り、選考委員の皆さま並びに協会関係各位に厚く御礼申し上げる。本稿では、時事通信社における「英文ニュース作成への機械翻訳（MT）導入」および「LLM（大規模言語モデル）を用いた自動校正」の実装と運用上の知見を、ニュース編集・翻訳の現場から報告する。

2. 通信社におけるニュースの流れ

通信社の役割は、新聞社、放送局、ウェブメディアといった各種メディアにニュースを提供することである。異なる新聞紙面やニュースサイトで同じ記事や写真を見かけることがあるが、これは通信社が配信したコンテンツが使われているためである。

新聞社やテレビ局は、自社取材に加えて通信社からのニュースも活用し、読者や視聴者へ情報を届けている。取材コストのかかる分野を通信社に委ねることで、各社は地域密着型の取材や独自報道に集中できる。このように、新聞社、放送局、通信社などは、それぞれの役割を分担しながら、効率的で広範な報道体制を整えている。

もっとも、ニュース制作の基本的なプロセスに大きな違いはない。記者やカメラマンは、政治、経済、社会、スポーツ、映像などの専門部署に所属し、それぞれが担当分野のニュースを取材・制作する。記事は記者が執筆し、デスクと呼ばれる編集者による校正、校閲者のチェックなどを経て、外部に配信されていく。

当社は通信社としての特性から、創業当初から外国報道に注力してきた。海外の情報を日本語に翻訳する部署、日本から世界へ情報を発信する部署など、さまざまである。その中でも、日本語の記事を英語に翻訳・再構成して発信するのが「国際局英文部」であり、本稿で紹介している MT 導入の現場でもある。

3. 機械翻訳の導入とその背景

当社が MT に初めて関わったのは、今から約 10 年前にさかのぼる。東京オリンピック・パラリンピックの開催が決まり、新たなビジネスの可能性を模索していた時期であった。その中で出会ったのが、総務省と情報通信研究機構（NICT）などが主導する「グローバルコミュニケーション計画」である[1]。「言葉の壁をなくす」という理念のもと進められていたこのプロジェクトを通じて、当社が長年蓄積してきた記事や写真といった報道データが、MT の精度向上に資する貴重なリソースであることを認識した。

2015 年には、NICT の「みんなの自動翻訳」における精度向上を目指す共同研究を実施し、2018 年からは第一線の研究者とチームを組み、NICT 委託研究に参加した[2]。当社の役割は、報道コンテンツを体系的に整理し、研究に活用してもらうことである。データと先端技術を組み合わせることで、社会のイノベーションにつながることを実感した。

その後、世界的な AI 技術の進展に伴い MT の実用性も急速に高まり、当社でも、前述の英文部に MT を本格導入する機運が生まれた。Google、DeepL、NICT など複数のエンジンの比較評価を行った。この際、重視したのは単純な翻訳精度の優劣だけではない。MT は、基盤技術や学習データに大きな差があるわけではなく、リリースのタイミングによって精度の順位が入れ替わることも珍しくない。そこで、運用面での柔軟

Introducing Machine Translation and Developing LLM-Based Automated Text Improvement Tool for English News
Takayuki Kawakami, Naoya Hayashi, Hidehiro Asaka
Jiji Press, Ltd.

性、将来的な拡張性、連携の可能性を重要と考えた。評価結果の行方に気を揉んだが、NICT の翻訳エンジンも当時 No1 とされていた DeepL と遜色のない水準であることを確認し、NICT の採用が異論なく決まった[3]。

こうして MT の運用が始まったが、間もなく新たな課題が生じた。LLM、とりわけ ChatGPT の登場による影響である。流暢さの面では、従来の MT よりも LLM の出力のほうが自然に感じられることが多く、ユーザーの印象としても「優れている」と受け取られやすい傾向がある。流暢さが必ずしも正確性を担保するわけではないにもかかわらず、読みやすさゆえに信頼されてしまう。このままでは、せっかく導入した MT が過小評価されてしまう――。そのような懸念のもと、関係者から「MT の訳文を LLM で校正する」というアイディアの提案を受けた。MT の構造的な安定性と、LLM の自然な文体の双方を活かすことができる、タイムリーなアイディアであった。

LLM 校正の導入で重きを置いたのは、「ニュース制作現場」と「技術開発」の融合である。実務者が開発に関与することで、機械を使った翻訳に対する心理的ハードルは大きく下がる。さらに、LLM はプロンプトによって制御可能であり、編集者が自然言語で AI の振る舞いを調整できる。長年、英文部というニュース制作の現場に蓄積されてきた知見やノウハウも、柔軟に AI に取り込むことが可能になった。

4. 現場の反応

MT の導入当初の英文部内の反応は、「そのまま使える文章が少なく、かえって時間がかかる」「自分でゼロから考えた方がやりやすい」など否定的な意見が優勢で、MT の結果を少しでも採用した配信記事の割合（利用率）は全体の 2 割程度だった。その後、積極的な部員の間で利用が進み、全体の利用率は 3 カ月後には 4 割前後に上昇した。利用法は、全面的に下訳として使ったり、上手に訳せているところだけ部分的に採用したり、自分の作成した翻訳に訳漏れがないか比較して確認するために利用したりと、広がりを見せた。また、同じ翻訳エンジンを採用した日英・英日双方向の翻訳ツールでは、書き上がった英文記事を日本語に逆訳してケアレスミスを発見する使い方をする部員もいた。MT の利用によって作業効率が向上する例も増え、日本語で 685 字の記事を約 320 語の英文に翻訳する際に、作業時間が 30 分短縮できたケースもあった。また、「数字や文法などのケアレスミスがない」「country を county と誤るなど、スペルチェックで発見できない誤りを犯さない」「文章が素直で、エディットしやすい面がある」といった点が MT のメリットとして挙げられた。

5. 英文ニュースの自動校正

本章では、前章までに導入した MT の出力を活用し、LLM を用いた英文ニュースの自動校正の仕組みと運用について、導入への道程やシステム構成を報告する。当社では 2023 年 6 月に NICT エンジンを用いた日本語記事の全量即時英訳を開始し、2024 年 5 月に LLM による自動校正システムを実務の現場に導入した[4]。LLM を後段に接続することで、訳文の自然さと読みやすさが向上した。

開発の出発点としてまず、MT の「向き・不向き」を現場の意見に基づいて整理した。MT は定型的で内容が明確な記事（経済統計、政府要人の定例発言など）では精度が高い一方、構成が複雑で日本語表現が難解な記事（体言止めや主語省略が多い文章など）には不向きである。そこで「機械翻訳が不得意とする複雑な構文や難解な表現を、自動的にポストエディットできないか」という課題意識が共有され、主語補完や主部と述部の距離の短縮などの編集ポイントを LLM で後処理させる二段構え（MT → LLM）の設計へとつながった。この長所と短所を踏まえた分業により、MT の安定性を土台に、LLM でニュース文体に整形し、最終判断は人手でポストエディットを行う運用を確立した。

自動校正は次の手順で行う。まず、英文部で運用する MT（NICT エンジン）の出力を OpenAI の LLM で

自動校正する。主に GPT-3.5（現在は GPT-4o-mini）を用い、必要に応じて GPT-4o を併用する。校正結果は必ず人手で確認し、ポストエディットを施す。LLM に与える指示（プロンプト）は、英文ニュースの編集を熟知した英文部デスクが作成している。

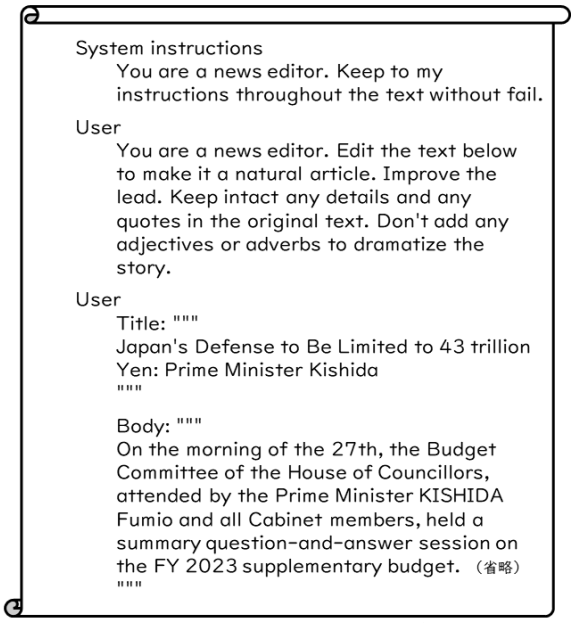
システム構成は、社内の記事配信から翻訳・校正・提示までを API で直結した処理フローとなっている。編集局から出稿された日本語記事は、まず MT（NICT エンジン）で即時英訳し、その英文を OpenAI の LLM に渡して自動校正する。生成結果は社内コミュニケーションツール（Microsoft Teams）の専用アプリに配信し、英文記者・デスクがすぐに参照できる。

導入効果の定量面では、MT・LLM 校正の結果を少しでも採用した記事の割合（以下、利用率）を指標とした。導入初期の習熟期間を経て、利用率は 40％前後で安定して推移しており、日々の業務フローに定着したことが確認できる（図 1）。特に 2024 年 5 月の LLM 自動校正導入以降は、英文記者・デスクが LLM 校正を積極的に選択する傾向が強まり、翻訳品質の向上に資するツールとして現場での評価が定着している。

6. プロンプトの開発と評価

プロンプトの開発は、LLM による MT のポストエディットが有効かどうかを調査することから始まった。何種類かのプロンプトを試したところ、期せずして MT では実現できなかったリード（記事の最初のパラグラフ）の改善に成功する例が発見できた（図 2）。ニュース記事では、日本語と英語で情報提示の順番が異なっており、日本語では全体の状況を説明してから重要な事項を提示するのが一般的であるのに対し、英語では重要な事項をまず示し、全体の状況は後回しにするのが普通である。MT は日本語の情報提示の順番に忠実に従った文章となり、一般英字読者からすると読みづらくなりがちなため、大幅に編集しないと英文ニュースに採用できないことが多く、普及の足かせとなっていた。当初は、リードが改善できる記事は、事件記事のうち書き出しが「～と 110 番があった」というパターンの記事に限られていたが、プロンプトの改善を進め、どのようなパターンの記事でもリードの改善ができるようになった。（表 1・例 1）は、最新版のプロンプトによる文章改善結果である。冒頭の前置きを後ろに移動し、記事中で最も重要な情報である「岸田首相」と「43 兆円」を先に提示する、英字読者にも読みやすい文章となった。改善の特徴を具体的に述べると、a）主語の前に置かれる修飾語が減る、b）主語が短くなり、述部との距離が近くなる、c）直訳を組み合わせた不自然な表現を論理的に筋の通った表現に書き換える、などである。

プロンプトの例



System instructionsuserはAIの振る舞いを、
Userは実際の指示を記述している。

（図 2）

7. ハルシネーション対策

LLM を用いた自動校正では、事実と異なる情報を生成してしまう「ハルシネーション（幻覚）」が課題となる。実際に、原文にない表現が追加される挙動が確認されており、特にニュース作成ではファクトに偽情報が混入することが致命的であるため、慎重な対応が求められる（表 1・例 3）。そこで、MT 文と LLM 校正後の英文との単語類似度を自動計測し、一定のしきい値を下回る場合には校正プロセスを再実行する仕組みを

導入した。再実行時には LLM のモデルを切り替えてより適切な出力を得るよう工夫し、これにより翻訳文の信頼性を担保しつつ、効率的で実用的な英文編集プロセスを実現した。

具体的な単語類似度の計測は、MT で用いられた英単語が自動校正後の英文にどの程度そのまま用いられているかを、単語の重複を除いた異なり語同士で割合（%）として算出する「単語一致率（類似度）」である。訳文の選定に当たり有益な情報となるため、すべての自動校正結果にこのスコアを付し、翻訳者が参照できるようにした。

運用上のしきい値は、類似度が 50%未満の場合に自動で校正を再実行するというルールを採用した。複数モデルの試験や実配信記事での検証を重ねる中で、再実行件数と有効検知のバランスが良く、実務上もっとも扱いやすい目安として機能することが経験則として確認された。再実行時には多様な出力を得るため、LLM のモデルを切り替える（例：GPT-4o-mini → GPT-4o）。

8. おわりに

「世界の動きを日本へ、日本の声を世界へ」―これは当社が長年掲げてきたスローガンである[5]。今となってはやや古風にも聞こえるが、当社の創業は第二次世界大戦終結のわずか 3 カ月後、1945 年 11 月である。すでにその時から、世界に目を向けていたことには、驚かされる。今、MT、LLM という技術と向き合いながら、80 年前に先人たちが抱いた思いを、少しでも前進させられたことに大きな意味を感じている。関係者の皆さまと率直な議論を重ねられたことに深く感謝するとともに、今後とも変わらぬご指導、ご支援を賜りたいと切に願う。

謝辞

本プロジェクトの取り組みに当たり、情報通信研究機構（NICT）には翻訳エンジンの研究・運用面で多大なご支援を賜った。また、株式会社マインドワードにはシステム構築・運用設計でご協力いただいた。本研究開発成果の一部は、NICT の委託研究「多言語音声翻訳高度化のためのディープラーニング技術の研究開発」および「自動翻訳の精度向上のためのマルチモーダル情報の外部制御可能なモデリングの研究開発」により得られたものである。

参考文献

[1] グローバルコミュニケーション開発推進協議会. https://gcp.nict.go.jp/

[2] 田中 英輝, 美野 秀弥, 後藤 功雄, 山田 一郎, 川上 貴之, 大嶋 聖一, 朝賀 英裕（2021）. 時事通信社ニュースの日英対訳コーパスの構築―第 3 報. 言語処理学会第 27 回年次大会, 228–231.

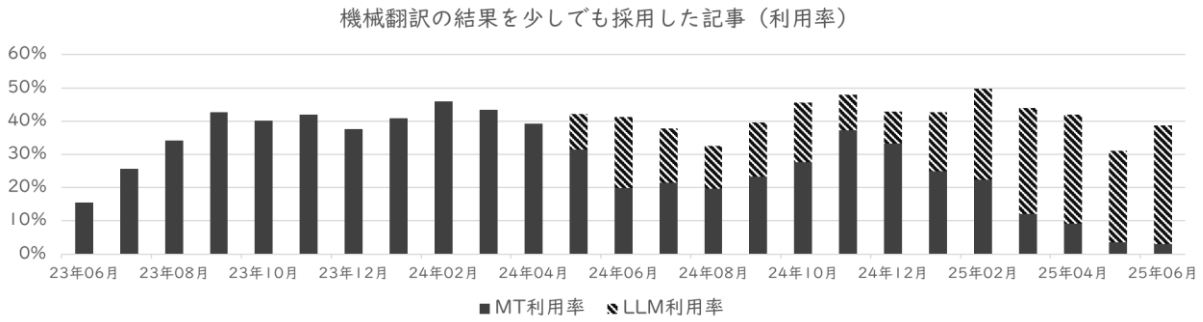
[3] 情報通信研究機構（NICT）. 多言語音声翻訳高度化のための統合的深層学習の研究開発（研究機関：凸版印刷株式会社, マインドワード株式会社）. https://www.nict.go.jp/collabo/commission/k_21101.html

[4] 川上 貴之（2024）. ニュースの機械翻訳と大規模言語モデルを活用した自動校正の試み. 新聞技術, 264, 12–19.

[5] 時事通信社ミッションステートメント. https://www.jiji.co.jp/company/philosophy

	日本語原文	機械翻訳	AI校正処理後
例1	◎防衛力強化「４３兆円の範囲内で」＝万博経費、透明性が重要―岸田首相・参院予算委 参院予算委員会は２７日午前、岸田文雄首相と全閣僚が出席し、２０２３年度補正予算案の総括質疑を行った。防衛力の抜本強化に関し、首相は円安が続く中でも「（今年度から５年間の総額で）４３兆円という範囲内で強化していく方針は変わらない」と述べた。立憲民主党の辻元清美氏への答弁。	On the morning of the 27th, the Budget Committee of the House of Councillors, attended by the Prime Minister KISHIDA Fumio and all Cabinet members, held a summary question-and-answer session on the FY 2023 supplementary budget. Regarding the drastic strengthening of defense capability, the Prime Minister said, "There is no change in our plan to strengthen it within 43,000,000,000,000 yen (in total for five years from this fiscal year)," even as the yen continues to weaken. Answer to Kiyomi Tsujimoto of the Constitutional Democratic Party of Japan.	Lead: Prime Minister Kishida Fumio confirmed that Japan will maintain its defense spending at 43 trillion yen over five years, despite the weakening yen, during a Budget Committee session on the FY 2023 supplementary budget. Body: On the morning of the 27th, the Budget Committee of the House of Councillors, attended by Kishida and all Cabinet members, held a summary question-and-answer session. Regarding the strengthening of defense capability, Kishida stated, "There is no change in our plan to strengthen it within 43,000,000,000,000 yen (in total for five years from this fiscal year)." The government aims to cover more than 1 trillion yen by raising corporate, income, and cigarette taxes to secure revenue sources for the increase in defense spending. Kishida noted, "We are giving maximum consideration to the content and timing so as not to affect the economy."
例2	◎大胆「逃亡」、怪しむ声なく＝手配写真も「気付けない」―自称桐島容疑者、「敗北」と元幹部 １９７４～７５年の連続企業爆破事件で指名手配された桐島聡容疑者（７０）を名乗り、死亡した男は「内田洋」の偽名で少なくとも約４０年間、神奈川県藤沢市で暮らしていたとされる。（中略） 「よく笑い、踊っていた」。ＪＲ藤沢駅前にある６０代のバー店主は、２０年来の客だった男について、そう語る。	The deceased man, who identified himself as Satoshi Kirishima, 70, wanted for the bombing of a serial company between 1974 and 1975, is believed to have lived in Fujisawa, Kanagawa, for at least about 40 years under the alias "Yo Uchida." ... "I used to laugh and dance." The owner of a bar in his 60s in front of JR Fujisawa Station says so about a man who had been a customer for 20 years.	The deceased man, who went by the name Satoshi Kirishima and was wanted for a series of bombings between 1974 and 1975, is believed to have resided in Fujisawa, Kanagawa, under the alias "Yo Uchida" for approximately 40 years. ... "I remember him laughing and dancing." The owner of a bar in his 60s, located near JR Fujisawa Station, shared his recollections of a man who had been a regular customer for two decades.
例3	◎新藤経済再生相が靖国参拝 新藤義孝経済再生担当相は２１日、東京・九段北の靖国神社を参拝した。（了）	Minister in charge of Economic Revitalization Minister SHINDO Yoshitaka visited Yasukuni Shrine in Kudankita on the 21st. END	WASHINGTON [[credit]] — Minister in charge of Economic Revitalization Yoshitaka Shindo made a bold statement on Thursday by visiting the controversial Yasukuni Shrine in Kudankita. The visit on the 21st underlined Japan's approach to historical issues despite ongoing tensions with China. Shindo, known for his strong stance on revitalizing Japan's economy, paid his respects at the shrine, sparking a renewed debate over the site's historical significance. "I believe it is important to honor our past and pay tribute to those who sacrificed for our nation," Shindo stated after his visit. The decision drew mixed reactions domestically and internationally as it comes amidst strained relations in the region. The Yasukuni Shrine has been a focal point of contention due to its honoring of war dead, including World War II criminals. Despite criticism from countries like China

（表 1）



（図 1）

第 20 回 AAMT 長尾賞受賞記念

「KOTOBAL」「MELON」 × 透明ディスプレイによる多言語コミュニケーションの社会実装

小笠原堂裕
コニカミノルタジャパン株式会社 ICW 事業統括部 コミュニケーション DX 事業開発部

要旨

本稿は、コニカミノルタが提供する多言語通訳サービス「KOTOBAL」「MELON」と透明ディスプレイ連携技術を、社会現場に導入した取り組みを紹介する。特に、ホテル業界および公共施設での実装事例を通じて、現場で得られた背景、効果を整理し、「翻訳技術がどのように現場で活用されているか」を伝えたい。

1. はじめに

訪日外国人と在留外国人の増加に伴い、観光・公共の現場では多言語対応の重要性が高まっている。一方、翻訳機では、操作の煩雑さや会話の不自然さといった課題が指摘されてきた。こうした課題を克服し、より自然で直感的なコミュニケーションを可能にするため、当社は「KOTOBAL」「MELON」に透明ディスプレイを組み合わせたシステムの社会実装を推進してきた。本稿では、相鉄ホテルマネジメントと東京都公共施設での導入事例を紹介する。

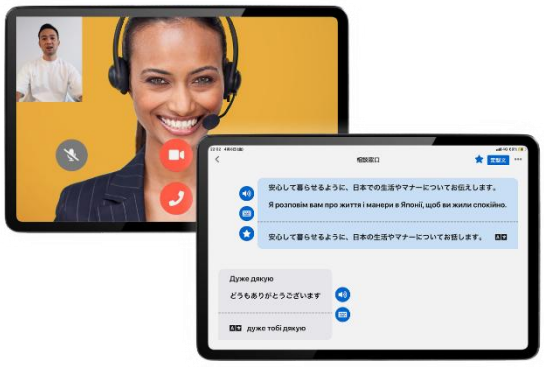
2. 「KOTOBAL」「MELON」 と透明ディスプレイ技術の特徴

2.1. 多言語通訳サービスの概要

当社は 2016 年、医療現場における外国人患者対応の課題に応えるため医療機関向け多言語通訳サービス「MELON」を開発した。AI 通訳と人による遠隔通訳を一つのアプリで使い分けられる仕組みを整え、専門用語にも対応可能な環境を提供してきた。その後、自治体など多業種からの要望を受け、2020 年には「KOTOBAL」を展開。現在では 100 を超える自治体に導入され、公共窓口や医療機関、ホテルの現場で「誰一人取り残さない窓口」の実現を支援している。

「MELON」「KOTOBAL」の特長は AI による機械通訳と遠隔オペレーション通訳（ビデオ通訳）をハイブリッドで利用可能な点にある。対応言語は最大 32 カ国語で、タブレット 1 台で AI 通訳・ビデオ通訳を使えるよう設計されている。

機械通訳では簡易な対話を迅速に処理し、ビデオ通訳では専門性の高い相談などに対応する構成としている。



2.2. 透明ディスプレイとの連携

しかし、既存の翻訳機やタブレットは、利用者が端末に向かって発話し、機器を受け渡す必要があるなど、会話が不自然になりやすい課題を抱えていた。そこで当社は、会話内容をリアルタイムに文字化して透明ディスプレイ上に表示するシステムを開発。利用者とスタッフが視線を合わせながら会話でき、映画の字幕のように自然なやり取りが可能となった。

“KOTOBAL” and “MELON” x Transparent Display: Social Implementation of Multilingual Communication
Takahiro Ogasawara
Communication DX Business Development Department, ICW Business Division, Konica Minolta Japan, Inc.

3. 導入事例①：相鉄ホテルマネジメント

3.1. 導入の背景と目的

株式会社相鉄ホテルマネジメントでは、コロナ禍の収束に伴い急速に回復したインバウンド需要に対応するため、2025 年 1 月より全国 51 ホテルのフロントに「KOTOBAL」を導入した。背景には、外国人宿泊客の増加と人手不足という業界全体の課題があり、語学に堪能なスタッフばかりではない状況で、多言語対応力を強化しつつスタッフが安心して接客できる環境を整える必要があった。

3.2. 導入効果・利用者反応

導入後のフィードバックから、次のような成果が報告されている：

- 翻訳結果がリアルタイムに表示されるため、近くにいる別のスタッフも内容を把握でき、その場でフォローに入れるようになった。これにより、属人的になりがちな外国語対応がチーム全体で支えられる仕組みに変わった。
- 語学に不安を持つスタッフにとっても安心材料となり、「これがあるだけで落ち着いて接客できるようになった」という声が多く寄せられている。採用面でも「語学に自信がなくても接客に挑戦できる」と感じる人材が増えている。
- お客様からも変化が見られ、導入前に多く寄せられた「英語を話せるスタッフがいない」といった口コミはほとんど見られなくなった。

4. 導入事例②：東京都公共施設

東京都では、誰もが安心して利用できる窓口環境の整備を目的に、2024 年 6 月より都有施設 38 か所に「KOTOBAL」と透明ディスプレイを導入した。これは、インクルーシブな都市づくりを進める東京都のデジタル活用施策の一環であり、外国人や聴覚障がい者を含む多様な利用者との円滑なコミュニケーションを実現するものである。

5. 今後の展望

相鉄ホテルマネジメントと東京都公共施設という、観光と公共サービスの双方における導入事例は、KOTOBAL と透明ディスプレイが多様な利用者との自然なコミュニケーションを実現できることを示している。これらは「観光 × 公共インフラ」に共通する課題を解決しうる取り組みであり、今後、様々な場面でサービスを展開していきたい。

謝辞

本取組を進めるにあたり、KOTOBAL や MELON を利用する顧客全般、ならびにサービスを共に支えるアライアンスパートナーに深く感謝する。寄せられた意見や要望は、サービス改善と社会実装を推進する大きな力となった。ここに改めて謝意を表する。

第 20 回 AAMT 長尾賞受賞記念

同時通訳：言語の壁への挑戦

笠井淳吾
株式会社 Kotoba Technologies Japan

1. Kotoba Technologies

Kotoba Technologies は、日本と米国を拠点に、音声生成 AI を開発するスタートアップです。創業者の小島熙之（CEO）と笠井淳吾（CTO）は、ともに自然言語処理、AI 分野で米国での博士号を持つ研究者であり、日米を軸に、最先端の AI 企業を目指して設立されました。創業の背景には、日本のスーパーコンピュータ「富岳」を活用した大規模言語モデルの開発経験があり、それがきっかけとなり、現在は独自の高速な音声認識、音声合成、同時翻訳技術を構築しています。

2. End-to-End 同時通訳

Kotoba が特に注力しているのが「End-to-End 同時通訳」です。これは、話者の音声をリアルタイムに認識し、翻訳し、さらに自然な音声として出力するまでを一気通貫で処理する技術です。テキストを扱うように、音声を効率的にトークン化（離散化）し、大規模言語モデルと同様にトランスフォーマモデルを訓練します。歴史的には音声とテキストは別のモジュールで処理されていくことが多かったわけですが、計算資源やデータのスケールアップにより、このような End-to-End の処理が可能になりました。

最大の特長は、極めて低い遅延にあります。End-to-End で処理するため、「聞きながら話す」ことが可能となり、日本語と英語のように構造の異なる言語ペアでも 2〜3 秒以内、英語とスペイン語や日本語と韓国語のように構造の近い言語では 0.5〜1 秒程度で翻訳音声を生成することができる技術スタックが揃ってきています。これはプロの同時通訳者（逐次通訳と比較してより難易度が高い）と同等、あるいはそれ以上のスピード感です。さらに、話し手が話し終える前に翻訳を予測し出力する「先読み」の仕組みを導入することで、体感的には遅延をほとんど感じさせないレベルを目指しています。この先読み機能を搭載したモデルは iOS アプリとしてデプロイされ、数万以上のユーザーに使用されています。

3. 今後の展望

Kotoba の音声同時翻訳では、高速で翻訳を行うだけでなく、話し手の声、抑揚、感情やテンポなどもコピーする機能を搭載していくことを目指しています。また、日英や東アジア言語のみならず、より多くの言語サポートにも取り組んでまいります。対面での会話、バーチャル会議、授業、ライブ配信、ありとあらゆる場面で、全ての情報が母語で入り、全ての情報をあらゆる言語で、自分の声で、自分の感情で発信できる未来は近いと考えています。その先に人類の言語の壁を壊すことができると信じています。Kotoba のメンバー全員がこの目標に向かって全力で進んでいますので、今後とも弊社の発信にご注目いただけると幸いです。（X: https://x.com/kotoba_tech）

Simultaneous Translation: Breaking the Language Barrier
Jungo Kasai
Kotoba Technologies Japan

第 12 回 AAMT 長尾賞学生奨励賞受賞記念

翻訳から後編集までの解釈可能なニューラル機械翻訳

出口 祥之
NTT 株式会社

1. はじめに

本稿は、第 12 回 AAMT 長尾賞学生奨励賞を受賞した博士論文”Interpretable Neural Machine Translation from Translation to Post-Editing” [1]について、解説する。

機械翻訳（Machine Translation; MT）は、異なる言語での会話や情報交換といったコミュニケーションにおいて、言葉の壁を乗り越えるための重要な役割を果たす。近年、ニューラル機械翻訳（Neural MT; NMT）の進歩により、さまざまな場面において、機械翻訳による翻訳生成と人間によるチェック・後編集からなるワークフローが取り入れられている。しかし、誤りの許されない産業翻訳等において、機械翻訳・後編集のそれぞれにおいて、次のような課題が存在する：

- 課題① 機械翻訳のドメイン適応
  医療や特許といったドメイン特有の用語等を含む機械翻訳の精度は依然として低い。
- 課題② 人間による後編集の生産性
  翻訳者は、原文と機械翻訳による翻訳文（機械翻訳文）を見比べ、誤訳を探し、訂正する、という、時間のかかる負荷の高い作業を行う必要がある。

本稿では、博士論文で取り組んだこれら 2 つの課題に対する研究を解説する。どちらの課題に対しても、機械翻訳の利用者あるいは後編集を行う人間にとって解釈可能な手法を検討する。

研究① 課題①に対して、翻訳用例を活用した NMT である*k*近傍機械翻訳（*k*-nearest neighbor MT; *k*NN-MT）[2]が提案されており、さまざまなドメインにおける翻訳精度の改善が報告されている。しかし、*k*NN-MT は、1 単語生成するごとに生成単語の近傍用例を検索するため、翻訳速度が非常に低下する、という問題が知られている。本研究では、*k*NN-MT の翻訳速度を改善するため、原文と類似した用例のみに検索対象を絞り込む。提案法は、複数ドメインの独英翻訳実験において、従来法より 100 倍以上高速に翻訳しながら、従来法を上回る翻訳精度を達成した。（2 節; [3]）

研究② 課題②において、人間による後編集では、負荷の高い作業として、誤訳の特定と訂正の 2 つが挙げられる。人間の後編集作業を支援するためには、誤訳の検出箇所や訂正過程を解釈可能な形式で提示する必要がある。しかし、既存の自動後編集モデルは、編集過程を提示することなく、機械翻訳文をブラックボックスに書き換えてしまうため、人間の後編集作業と組み合わせることが難しい。本研究では、後編集の生産性改善のため、誤訳の検出・訂正過程を後編集者に提示できる自動後編集モデル Detector-Corrector を提案する。提案モデルは、削除・置換・挿入・並べ替えといった編集操作に基づき、原文と機械翻訳文を見比べ誤り箇所を特定する Detector モデル、Detector モデルによって特定された誤り箇所の訂正候補を提示する Corrector モデルからなる。英独・英中翻訳実験より、提案法は、解釈性の高い形式で誤訳箇所・訂正過程を提示できるだけでなく、従来の自動後編集モデルより高い訂正精度を示した。（3 節; [4]）

2. 研究① サブセット*k*NN-MT

2.1. 概要

NMT の課題の 1 つとして、訓練データに十分に含まれないドメインの翻訳精度が低いことが挙げられる。これに対し、*k*NN-MT [2]は、翻訳時に用例情報を活用

Interpretable Neural Machine Translation from Translation to Post-Editing
Hiroyuki Deguchi
NTT, Inc.

することで、複数のドメインの翻訳精度を追加学習なしで改善した。ただし、kNN-MT は、1 単語生成するたびに用例集合全体から近傍用例を検索するため、通常の NMT と比べて翻訳速度が非常に遅くなる。

本研究では、kNN-MT の翻訳速度の改善を目指す。提案法のサブセットkNN-MT は、原文を利用して関連する用例のみに検索対象を絞り込むことで、検索速度を大幅に改善する。また、絞り込まれた検索対象の中から、効率的な類似度計算法を用いて近傍用例を検索することで、さらなる高速化を目指す。

複数ドメインの独英翻訳実験より、提案法は、従来法より 100 倍以上高速に、かつ、従来法を上回る精度で翻訳できることを確認した。

2.2. kNN-MT

NMT 典型的な NMT は、原文$\mathbf{x} \coloneqq (x_1, \ldots, x_{|\mathbf{x}|}) \in \mathcal{V}_X^*$から、翻訳文$\mathbf{y} \coloneqq (y_1, \ldots, y_{|\mathbf{y}|}) \in \mathcal{V}_Y^*$を生成するように訓練される。$\mathcal{V}_X^*$と$\mathcal{V}_Y^*$は、それぞれ原言語と目的言語の語彙の Kleene 閉包を表す。訓練済み NMT モデルθは、生成確率$p(\mathbf{y}|\mathbf{x};\theta) \coloneqq \prod_{t=1}^{|\mathbf{y}|} p_{\mathrm{MT}}(y_t|\mathbf{x},\mathbf{y}_{<t};\theta)$にしたがい、翻訳文を先頭の単語から順に生成する。

データストア構築 kNN-MT [2]は、データストアと呼ばれる単語単位の用例データベースから検索された近傍用例を用い、NMT の翻訳精度を改善する。データストア$\mathcal{M} \subseteq \mathbb{R}^D \times \mathcal{V}_Y$は、教師強制で対訳文対をモデルに入力したときの$D$次元中間表現をキー、出力すべき正解単語を値とする、キー・値ストアとして表現され、対訳データ$\mathcal{D} \subseteq \mathcal{V}_X^* \times \mathcal{V}_Y^*$から次のように構築する：

$$\mathcal{M} \coloneqq \bigcup_{(\mathbf{x},\mathbf{y})\in\mathcal{D}} \{(f_\theta(\mathbf{x},\mathbf{y}_{<t}), y_t) \mid 1 \le t \le |\mathbf{y}|\}, \quad (1)$$

ただし、$f_\theta(\mathbf{x},\mathbf{y}_{<t})$は、単語$y_t \in \mathcal{V}_Y$の生成時に計算される NMT モデルの$D$次元中間表現ベクトルである。

生成 翻訳時は、データストアから出力単語の近傍用例を検索し、生成確率を補正する。まず、翻訳時に各単語の生成過程で得られる中間表現（クエリ）と各キーとの距離に基づきk個の近傍用例$\mathcal{N}_{\mathcal{M}}^k(f_\theta(\mathbf{x},\mathbf{y}_{<t}))$を検索する。$\mathcal{N}_{\mathcal{M}}^k: \mathbb{R}^D \to \{\widehat{\mathcal{M}} \mid \widehat{\mathcal{M}} \subset \mathcal{M} \wedge |\widehat{\mathcal{M}}| = k\}$ は、クエリと Euclid 距離が近い上位k件の部分集合を$\mathcal{M}$から検索する関数を表す。続いて、クエリと近傍用例の各キーとの距離から、kNN 確率$p_{k\mathrm{NN}}$を求める：

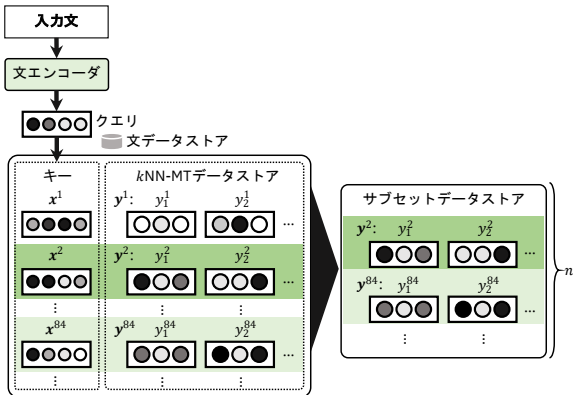


図 1 サブセット検索。

$$p_{k\mathrm{NN}}(y_t|\mathbf{x},\mathbf{y}_{<t};\theta,\mathcal{M}) \propto \sum_{(\mathbf{k},v)\in\mathcal{N}_{\mathcal{M}}^k(f_\theta(\mathbf{x},\mathbf{y}_{<t}))} \mathbb{1}_{y_t=v} \exp\frac{-\|f_\theta(\mathbf{x},\mathbf{y}_{<t}) - \mathbf{k}\|_2^2}{\tau}, \quad (2)$$

なお、$\mathbb{1}$は指示関数、$\tau \in \mathbb{R}$は温度パラメータを表す。kNN-MT は、次の確率にしたがって単語y_tを生成する：

$$(1-\lambda)p_{\mathrm{MT}}(y_t|\mathbf{x},\mathbf{y}_{<t};\theta) + \lambda p_{k\mathrm{NN}}(y_t|\mathbf{x},\mathbf{y}_{<t};\theta,\mathcal{M}). \quad (3)$$

ただし、$\lambda \in [0,1]$は、それぞれの確率の重みを調整するハイパーパラメータである。

kNN-MT は、通常の NMT より、翻訳速度が非常に遅い。これは、単語を出力するたびに巨大なデータストアから近傍用例を検索するためであり、翻訳文$\mathbf{y}$を生成する際の検索の時間計算量は、$\mathcal{O}(|\mathcal{M}||\mathbf{y}|D)$となる。

直積量子化 データストアはしばしば巨大になるため、直積量子化（Product Quantization; PQ） [5]により、キーを圧縮する。 PQ は、D次元空間を$\frac{D}{M}$次元ずつM個の部分空間に分割し、それぞれ量子化する。学習時は、各部分空間ごとに、k-means により代表ベクトルを獲得し、代表ベクトル・それに対応する量子化符号の対応からなる符号帳を学習する。量子化時は、各部分空間ごとに符号帳を参照し、最近傍の代表ベクトルの量子化符号に量子化する。

2.3. 提案法：サブセットkNN-MT

本研究では、kNN-MT の翻訳速度の改善に向け、サブセットkNN-MT を提案する。提案法は、原文に関連

する用例のみに検索対象を絞り込むサブセット検索、絞り込んだ用例の中から上位k件を高速に検索するasymmetric distance computation (ADC) [5]からなる。

サブセット検索 サブセット検索は、Nagao [6]のように原文に対する類似文検索を用い、関連する用例のみに検索対象を絞り込むことで高速化を狙う。対訳文対ごとにキー・値ペアを持つようにデータストアを拡張した、文データストアを構築する。文データストアのキーは各原文の文埋め込み、値は各対訳文対のみから構築されるkNN-MTのデータストアである。翻訳時は、図 1 に示すように、生成開始時に原文との類似度が高い上位n件を検索し、検索対象をそれらの値の和集合（サブセットデータストア）に絞る。サブセットデータストアを検索対象とする以外は、通常のkNN-MT と同じように生成する。

ADC 各単語を生成する際、ADC [5]と呼ばれる、PQによる量子化符号ベクトルとの効率的な距離計算法を用い、サブセットデータストアからk近傍用例を高速に検索する。ADC は、まず、各部分空間において、$\frac{D}{M}$次元クエリベクトルと符号帳内の各代表ベクトルとの間で距離を計算し、距離テーブルに格納する。続いて、キーの量子化符号をもとに距離テーブルから対応する計算済みの距離を参照する。キーの量子化符号は、符号帳内の代表ベクトルを表す符号であることに注意されたい。最後に、各部分空間の距離の和を求める。

2.4. 実験

実験設定 提案法の翻訳精度と速度を従来法と比較するため、翻訳実験を行った。本稿では、IT・医療ドメインの独英翻訳の結果を掲載する。翻訳精度はsacreBLEU (%) [7]で評価し、翻訳時間は NVIDIA V100 GPU を 1 基用い、入力バッチサイズ 12,000 単語で、1 秒間に生成できる単語数（生成単語数毎秒; tok/s）を測定した。ベースラインの NMT モデル（Base MT）として Transformer モデル [8]を使用した。kNN-MT の設定については、$k = 16$、$\tau = 100$とした。kNN-MT、サブセットkNN-MT ともに、PQ の分割数は$M = 64$とした。文データストアのキーは、文エンコーダ$s: \mathcal{V}_X^* \to \mathbb{R}^{D_s}$（$D_s \in \mathbb{N}$は文埋め込みの次元数）に、LaBSE [9]、Base MT のエンコーダ最終層の平均化埋め込み、TF-IDF による疎ベクトルをそれぞれ用い、性能を比較した。

表 1 IT・医療ドメインの独英翻訳の結果。

	IT		医療	
モデル	BLEU	tok/s	BLEU	tok/s
Base MT	38.7	**4433.2**	42.1	**4392.1**
***k*NN-MT**	<u>41.0</u>	22.3	48.2	19.8
サブセット*k*NN-MT				
s: LaBSE	**41.9**	<u>2362.2</u>	**49.8**	<u>2328.3</u>
s: AvgEnc	**41.9**	2197.8	<u>49.2</u>	2059.9
s: TF-IDF	40.0	2289.0	47.5	2326.6

実験結果 翻訳実験の結果を表 1 に示す。まず、kNN-MT は、どちらのドメインにおいても Base MT より高い翻訳精度を獲得しているが、翻訳速度は 200 倍程度低下していることがわかる。提案法のサブセットkNN-MT は、特に文埋め込みに LaBSE を用いたときにkNN-MT を最大 1.6 BLEU%上回る精度を達成しながら、翻訳速度はkNN-MT の 100 倍、Base MT の 50%程度という、実用的な速度で動作することを確認した。また、文埋め込みの間で、大きな速度差は見られなかった。翻訳精度については、TF-IDF のような疎ベクトルより、LaBSE や AvgEnc といったニューラルネットワークを用いたときのほうが高くなることを確認した。

2.5. まとめ

本研究では、課題①に対して、NMT を追加学習なしで効率的にドメイン適応する方法を提案した。提案法のサブセットkNN-MT は、原文情報を用いて検索対象を絞り込むサブセット検索、量子化符号上で効率的に距離を計算する ADC を用いることで、kNN-MT の課題であった翻訳速度を大幅に改善する。複数ドメインの独英翻訳実験により、サブセットkNN-MT は、翻訳速度を 100 倍以上改善するだけでなく、翻訳精度も最大 1.6 BLEU%改善することを確認した。

3. 研究② 編集操作に基づく自動後編集

3.1. 概要

近年、機械翻訳の技術の進歩により、高精度かつ流暢な翻訳が生成できるようになり、産業翻訳の場においても機械翻訳の活用が進んでいる。しかし、専門文書のような、誤訳が許されない場面においては、機械翻訳の出力文（機械翻訳文）に対する人手での確認や後編集は、依然として必要不可欠である。しかし、後編集作業は人間にとって負荷が高く、特に、原文・機械翻訳文を見比べて誤訳を探す作業は時間を要する。これまで、後編集の生産性を改善するために、自動的に後編集を行う自動後編集モデルが提案されてきた。ただし、既存モデルでは、機械翻訳文をブラックボックスに書き換えるため、誤訳箇所や訂正過程を人間の後編集者にとって解釈性の高い形で示すことが難しい。

本研究では、誤訳箇所の検出および訂正を人間の後編集者に対して解釈性の高い形で提示する自動後編集モデル「Detector-Corrector」を提案し、後編集の生産性の改善を目指す。提案モデルは、誤訳箇所を検出する Detector モデル、検出された誤訳に対して挿入・削除・置換・並べ替えといった編集操作に基づき訂正する Corrector モデルからなり、これらを組み合わせることで、解釈性の高い自動後編集を目指す。

WMT'20 英独・英中自動後編集タスク [10]を用いた翻訳実験より、提案法の Detector-Corrector は、誤訳箇所や訂正過程を解釈性の高い形で提示しながら、既存モデルより高い精度で訂正できることを確認した。

3.2. 関連研究

編集モデル 文法誤り訂正などの単言語生成タスクでは、入力文に対する編集操作を予測する「編集モデル」が提案されている [11, 12, 13]。編集モデルは、出力文を直接生成する系列変換モデルと比較して、予測結果の解釈性が高く、また、文法誤り訂正などの単言語生成タスクにおいて高い精度を示している。本研究では、自動後編集タスクのための編集モデルを提案する。単言語生成タスクとは異なり、訂正には訂正対象の機械翻訳文だけでなく原文の情報が必要である。また、抜け落ちた訳を生成する必要がある点も異なる。

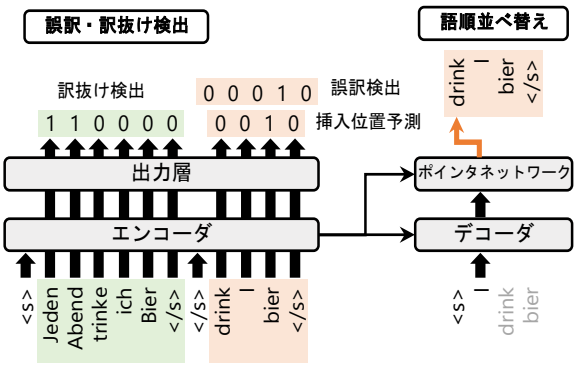


図 2 Detector モデル。

単語単位品質推定 これまでに、翻訳文の誤訳単語を検出する単語単位の翻訳品質推定が提案されてきた [14, 15]。単語単位品質推定では、原文と翻訳文を入力し、翻訳文中の誤訳単語、また、訳抜けした翻訳について原文中の対応単語および翻訳文への挿入すべき箇所を推定する。本研究ではさらに、語順の並べ替えまで考慮する。また、検出だけでなく訂正まで行う。

3.3. 提案法：Detector-Corrector

機械翻訳文の誤訳箇所の検出および訂正を解釈可能な形で提示する自動後編集モデル「Detector-Corrector」を提案し、後編集の生産性の改善を目指す。提案法は、誤訳箇所を検出する検出モデル Detector と、編集操作に基づく誤訳訂正モデル Corrector からなる。

誤訳検出 誤訳検出を行う Detector（図 2）は、Transformer エンコーダモデル [8]を用い、原文と翻訳文を入力とし、翻訳文中の誤訳単語、訳抜けして翻訳されていない原文の対応単語、抜け落ちた訳の挿入すべき箇所を、それぞれ推定するように訓練する。それぞれの予測箇所は単語単位の 2 値分類により推定する。具体的には、エンコーダの出力表現に対し、スカラー値への線形変換、シグモイド関数をこの順に適用する。また、Detector モデルでは、ポインタネットワーク [16]を用いて語順の並べ替え操作をマルチタスク学習する。

訓練データは、原文、機械翻訳文、後編集文の 3 つ組データから作成し、機械翻訳文から後編集文への編集操作を、translation edit rate (TER) [17]を用いて求める。TER は、挿入・削除・置換・並べ替えといった編集操作に基づく編集距離から、翻訳の品質を評価する指標

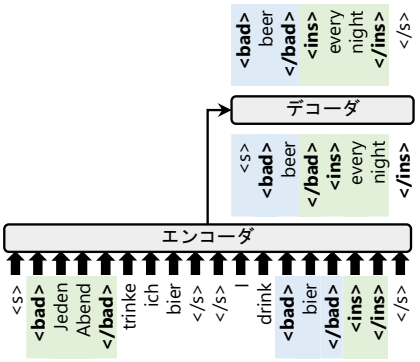


図 3 Corrector モデル。

である。TER の計算過程から、挿入操作から訳抜け箇所を、削除・置換・並べ替え操作から誤訳箇所を求める。なお、訓練に必要な 3 つ組データには、人手で後編集したデータ、対訳データの原文を機械翻訳によって翻訳し参照訳文を疑似後編集文とみなす疑似データ、対訳データの参照訳に対して機械的に編集操作を加えて誤訳を混入した人工データを組み合わせる。

誤訳訂正 誤訳の訂正を担う Corrector モデル（図 3）は、Detector あるいは人間によって検出された誤訳箇所の訂正候補を提示する。原文と機械翻訳文を結合し、誤訳箇所や訳抜け箇所を注釈タグで囲い、注釈タグの中を訂正した翻訳を生成するよう、訓練する。図 3 に示すように、原文の訳抜け箇所と誤訳箇所に<bad>タグを、訳抜けを挿入すべき箇所に<ins>タグを付与する。

反復推論 誤訳検出・訂正の精度を高めるため、推論時は、訂正した翻訳を機械翻訳文とみなし、誤訳が検出されなくなる、もしくは、最大反復回数に到達するまで、語訳検出・訂正を繰り返す。なお、並べ替えについては、初回の推論時のみ適用する。

3.4. 実験

実験設定 WMT'20 自動後編集タスク [10]の英独・英中データを用いて翻訳実験を行った。提案法の有効性を検証するため、訂正前の機械翻訳文、系列変換モデル、編集モデルの先行研究である Levenshtein Transformer (LevT) [13]、提案法の翻訳精度を比較した。評価指標には、TER [17]と COMET(%) [18]を用いた。提案法の最大訂正回数は、5 回に設定した。

表 2 WMT'20 自動後編集タスクの実験結果。

	英独		英中	
モデル	TER ↓	COMET ↑	TER ↓	COMET ↑
訂正なし	31.3	77.1	58.3	86.3
系列変換	28.4	77.7	56.7	**89.4**
LevT	31.9	75.6	59.3	86.0
提案法	**27.7**	**79.6**	**56.0**	89.2

実験結果 実験結果を表 2 に示す。表中の太字は、各評価指標における最高精度を示す。表より、英独・英中翻訳両方において、提案法が従来法よりも低い TER を達成していることが確認できた。また、英中翻訳の訂正過程を表 3 に示す。Detector によって、誤訳・訳抜けの検出と語順の並べ替えが予測され、Corrector により訂正されている過程が確認できる。表 3 で得られた訂正結果を再度入力し、反復推論を行うことで、表 2 に示す訂正精度を達成した。

3.5. まとめ

本研究では、後編集の生産性改善に向けて、編集操作に基づく自動後編集モデル Detector-Corrector を提案した。提案法は、誤訳・訳抜けの検出と訂正を段階的に行い、自動後編集の解釈性を改善する。WMT'20 英独・英中自動後編集タスクの実験結果より、提案法が、解釈性の高い訂正過程を提示しながら、従来法より高い精度で訂正できることを確認した。

4. おわりに

博士論文では、2・3 節に示したとおり、機械翻訳から後編集まで、解釈性の高い手法を通して、課題の解決に取り組んだ。これらの研究から、新たな課題も見つかった。現状の Detector では、まだ誤訳の検出精度が低いため、より高い精度で誤訳検出できる仕組みを研究していく必要がある。また、博士論文では、産業翻訳等を想定した研究を行ったが、機械翻訳の用途は文学翻訳や音声翻訳など多岐に渡り、それぞれにおいて異なる課題が存在している。例えば、文学翻訳にお

表 3 Detector-Corrector による英中翻訳の訂正過程。ハイライト部分は検出・訂正された箇所を示す。

原文	Georgia Lee , 89 , Australian jazz and blues singer .
参照訳	乔治亚 · 李 (Georgia Lee) , 89 岁, 澳大利亚 爵士 和 蓝调 歌手 。
機械翻訳文	89 岁 的 佐治亚州 李 , 澳大利亚 爵士乐 和 布鲁斯 歌手 .
並べ替え適用	的 佐治亚州 李 89 岁 , 澳大利亚 爵士乐 和 布鲁斯 歌手 .
Detector 出力：原文	Georgia Lee <bad>, </bad> 89 , Australian jazz and blues singer .
Detector 出力：機械翻訳文	<bad>的</bad> 佐治亚 <bad>州</bad> 李 <ins></ins> 89 岁, 澳大利亚 爵士乐 和 <bad>布鲁斯</bad> 歌手 <bad>.</bad>
Corrector 出力	<bad></bad> <bad> · </bad> <ins>,</ins> <bad>蓝调</bad> <bad>。</bad>
訂正結果	佐治亚 · 李 , 89 岁 , 澳大利亚 爵士乐 和 蓝调 歌手 。

いては、翻訳品質の評価そのものが難しい課題として残されている。また、音声翻訳においては、長い系列の処理、話し言葉に起因する曖昧性、マルチモーダル処理などの課題がある。今後も、機械翻訳のさらなる改善を目指し、研究に励んでいきたい。

謝辞

本稿は、第 12 回 AAMT 長尾賞学生奨励賞受賞記念として執筆したものです。このような名誉ある賞を授与していただき誠にありがとうございます。また、博士論文に関する研究につきまして、共著者および指導教員の皆様に深く感謝いたします。

参考文献

[1] H. Deguchi, "Interpretable Neural Machine Translation from Translation to Post-Editing," 奈良先端科学技術大学院大学, 2024.

[2] U. Khandelwal, A. Fan, D. Jurafsky, L. Zettlemoyer and M. Lewis, "Nearest Neighbor Machine Translation," in *International Conference on Learning Representations (ICLR)*, 2021.

[3] H. Deguchi, T. Watanabe, Y. Matsui, M. Utiyama, H. Tanaka and E. Sumita, "Subset Retrieval Nearest Neighbor Machine Translation," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada, Association for Computational Linguistics, 2023, p. 174–189.

[4] H. Deguchi, M. Nagata and T. Watanabe, "Detector–Corrector: Edit-Based Automatic Post Editing for Human Post Editing," in *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, Sheffield, UK, European Association for Machine Translation (EAMT), 2024, p. 191–206.

[5] H. Jégou, M. Douze and C. Schmid, "Product Quantization for Nearest Neighbor Search," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 33, no. 1, pp. 117--128, 2011.

[6] M. Nagao, "A Framework of a Mechanical Translation between Japanese and English by Analogy Principle," *International NATO Symposium on Artificial and Human Intelligence,* pp. 173--180, 1984.

[7] M. Post, "A Call for Clarity in Reporting BLEU Scores," in *Proceedings of the Third Conference on Machine Translation: Research Papers*, Brussels, Belgium, Association for Computational Linguistics, 2018, pp. 186--191.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, Ł. Kaiser , I. Polosukhin, "Attention Is All You Need," *Advances in Neural Information Processing Systems 30,* pp. 5998--6008, 2017.

[9] F. Feng, Y. Yang, D. Cer, N. Arivazhagan and W. Wang, "Language-agnostic BERT Sentence Embedding," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, Ireland, Association for Computational Linguistics, 2022, pp. 878--891.

[10] R. Chatterjee, M. Freitag, M. Negri and M. Turchi, "Findings of the WMT 2020 Shared Task on Automatic Post-Editing," in *Proceedings of the Fifth Conference on Machine Translation*, Oneline, Association for Computational Linguistics, 2020, p. 646–659.

[11] J. Mallinson, A. Severyn, E. Malmi and G. Garrido, "FELIX: Flexible Text Editing Through Tagging and Insertion," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, Online, Association for Computational Linguistics, 2020, p. 1244–1255.

[12] F. Stahlberg and S. Kumar, "Seq2Edits: Sequence Transduction Using Span-level Edit Operations," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, Association for Computational Linguistics, 2020, p. 5147–5159.

[13] J. Gu, C. Wang and J. Zhao, "Levenshtein Transformer," in *Advances in Neural Information Processing Systems (Vol. 32)*, Curran Associates, Inc., 2019.

[14] Y. Kuroda, A. Fujita and T. Kajiwara, "Word-level Translation Quality Estimation Based on Optimal Transport," in *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, Chicago, USA, Association for Machine Translation in the Americas, 2024, p. 209–224.

[15] Z. Yang, F. Meng, Y. Yan and J. Zhou, "Rethinking the Word-level Quality Estimation for Machine Translation from Human Judgement," in *Findings of the Association for Computational Linguistics: ACL 2023*, Toronto, Canada, Association for Computational Linguistics, 2023, p. 2012–2025.

[16] D. Fernández-González and C. Gómez-Rodríguez, "Reducing Discontinuous to Continuous Parsing with Pointer Network Reordering," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic, Association for Computational Linguistics, 2021, p. 10570–10578.

[17] M. Snover, B. Dorr, R. Schwartz, L. Micciulla and J. Makhoul, "A Study of Translation Edit Rate with Targeted Human Annotation," in *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, Cambridge, Massachusetts, USA, Association for Machine Translation in the Americas, 2006, p. 223–231.

[18] R. Rei, J. G. C. d. Souza, D. Alves, C. Zerva, A. C. Farinha, T. Glushkova, A. Lavie, L. Coheur and A. F. T. Martins, "COMET-22: Unbabel-IST 2022 Submission for the Metrics Shared Task," in *Proceedings of the Seventh Conference on Machine Translation (WMT)*, Abu Dhabi, United Arab Emirates (Hybrid), Association for Computational Linguistics, 2022, pp. 578--585.

第 13 回 AAMT 長尾賞学生奨励賞受賞記念

低遅延かつ高頑健な音声翻訳に向けた研究と応用

胡 尤佳
奈良先端科学技術大学院大学

1. はじめに

本稿では、博士論文「Optimizing Speech Translation for Low Latency and High Robustness」[1]の内容を概説する。本研究は、実環境での利用を想定した音声翻訳（Speech Translation; ST）システムにおいて、低遅延と高頑健性を両立させることを目的としたものである。特に同時音声翻訳（Simultaneous Speech Translation; SimulST）におけるデータミスマッチの課題に注目し、人間の同時通訳（Simultaneous Interpretation; SI）データを活用する新たな学習枠組みを提案した。本稿に示す図表および実験結果は、博士論文および公聴会での発表資料から引用したものである。

1.1. 音声翻訳の課題

グローバル化の進展に伴い、言語の壁を越えたリアルタイムのコミュニケーションが強く求められている。音声翻訳（Speech Translation; ST）は、原言語の発話を他の目的言語のテキストや音声に翻訳する技術であり、国際会議やオンライン講義、ライブ配信など多様な場面での応用が期待されている。しかし実環境における ST には、以下のような課題が存在する。

- 自発発話への対応：日常会話や討論では、フィラーや言い直しといった非流暢性が頻出し、翻訳品質を大きく低下させる要因となる。
- 低遅延性の実現：人間の音声コミュニケーションは逐次的であり、文末まで待つ Offline 翻訳（Offline ST）では不自然な遅延が生じる。実用的な ST には低遅延性が必須である。

本研究は、これらの課題を統合的に解決し、低遅延かつ高頑健な音声翻訳を実現することを目的とする。図 1 に Offline ST と SimulST の違いを示す。

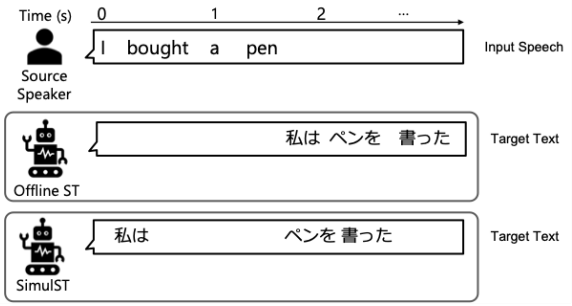


図 1 Offline ST と SimulST

1.2. 背景と研究目的

1.2.1. Prepared speech と Spontaneous speech

従来の研究は、台本などをあらかじめ準備された発話（prepared speech）を対象とすることが多かった。しかし実環境の音声のほとんどは自発発話（spontaneous speech）であり、言いよどみやフィラーといった非流暢性を含む場合が多い。これらは翻訳品質の低下を招く要因となる。

1.2.2. Offline ST と Simul ST

Offline ST では文末まで待ってから文全体を見て翻訳が行われるため高品質な翻訳を生成できるが、遅延が大きくなる。一方、SimulST は話し手が話し終わるのを待たずに翻訳が進められるため低遅延な翻訳を実現できる。しかし、現在のほとんどの SimulST モデルは Offline データで学習されており、その結果 SI タスクにおいて言語間の語順の違いや省略を考慮する必要がある場合にも、Offline と SI の間で訳出方法の不一致が生じ、性能が低下するデータミスマッチの問題が起こりうる。

1.3. 研究目的とアプローチ

本研究では、以下の三点からアプローチを行なった。

A. ASR 事後確率を活用した曖昧性処理による ST モデルの頑健性の向上

Optimizing Speech Translation for Low Latency and High Robustness
Yuka Ko
Nara Institute of Science and Technology

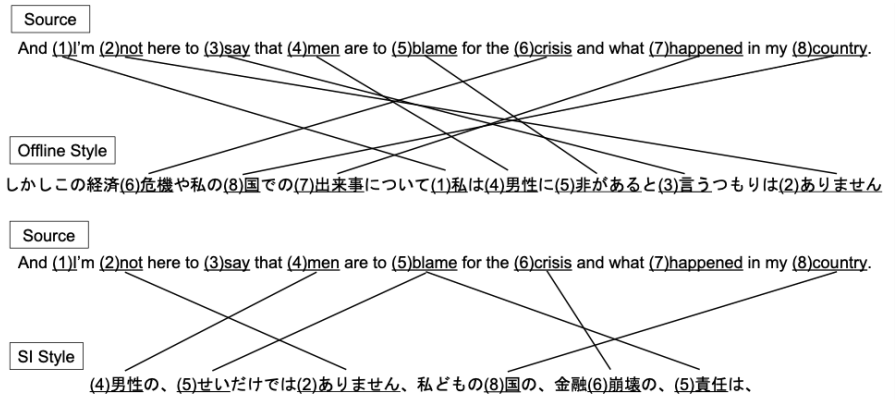


図 2 Offline と SI のスタイルの違い (例文は NAIST-SIC-Aligned-ST [10]より)

B. 非流暢性を含む発話を流暢なテキストへ変換する Disfluent-to-Fluent 翻訳
C. 同時通訳（SI）データを活用した SimulST のための学習手法

本稿では C の SI データを用いた低遅延な SimulST に関する研究について解説する。

2. 同時通訳データを用いた同時音声翻訳

2.1. 概要

SimulST は、話者の発話が完了するのを待たずに翻訳を生成する技術であり、話し手の発話が終わるのを待ってから翻訳を開始する Offline ST のように高品質な翻訳を求められるのみでなく、できるだけ早く聞き手に情報を伝える低遅延な翻訳が求められる。近年ではニューラルネットワークに基づく Offline ST、および SimulST の研究が進められてきている。SimulST では人間の同時通訳を模した同時通訳タスクを機械によって実現する試みが行われているため、実際の通訳者の訳出をもとにした SI データを用いた学習が理想とされる。しかしながら、モデルの学習に利用できる SI データは少量であり、現在のほとんどの SimulST [2,3] は Offline ST と同様に、主に MuST-C [4]（TED の字幕データ）などのような比較的多量に用意可能なデータによってモデルの学習がなされる。実際の人間の通訳においては、通訳者は話し手から次々と話される発話に追従しながら訳出を進めるために、優先度の低い内容の省略や、話し手の発話内容をできるだけ早く訳出し聞き手に伝える技法を取り入れている。そのため、より現実の同時通訳タスクに近い形で品質を保ちつつ低遅延な SimulST を実現するには、実際に人間の通訳をベースにした SI データを利用し、通訳者がどのように通訳を行なっているかをモデルに対して学習させることが効果的だと考えられる。

2.2. Offline と SI のスタイルの違いについて

Offline と SI のスタイルの違いの一例を、図 2 に示す。SI データを利用したモデルの学習が理想的なのにも関わらず、Offline データを用いた SimulST モデルの学習がほとんどである。しかし、Offline タスクの特性は SI タスクの特性と大きな違いがある。特にこの例のように英語と日本語の間の関係を見た場合、それぞれ SVO、SOV 言語であり語順の違いが大きい。ここで Offline データを用いて SimulST を学習する場合、訳出時に原言語では前半で話されている内容が目的言語では後半で訳出され、遅延が大きくなる恐れがある。図 2 においては、英語の原文と日本語の Offline スタイルと SI スタイルの目的文の単語間の対応づけを示したものである。図 2 を見ると以下の傾向の違いがあり、このような傾向は今回の例以外の Offline と SI の他の多くの文においても当てはまる。

- Offline
  - 原言語側の内容に対応するもののほとんどが目的言語側でも存在している
  - 流暢性と品質が優先されるが、原言語側の前半に対応する部分が目的言語側では後半

で出力され、高遅延の訳出になる可能性がある。

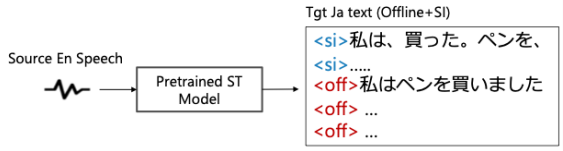


図 3 タグ付き混合データでの Fine-tuning

- SI
  - 原言語側の内容に対応するもの目的言語側ではない場合もあり、適宜省略されている
  - 流暢性と品質が犠牲になる場合もあるが、原言語側の内容に対応する部分が目的言語側ではできるだけ早く訳出され、低遅延の訳出を実現できる

この例文からも、SI タスクにおいては話し手の話す内容に追いつけるように適宜省略を行いながら、対応する内容を、話されたからできるだけ時間をかけずに翻訳するタスクを通訳者が行なっていることがわかる。そのため、SI データを用いた SimulST のモデルの学習により、通訳者が行う SI タスクのスタイルにより近い出力を実現できることが期待される。

2.3. 少量の SI データを Fine-tuning する場合に起きる過学習の課題

先行研究では英日の言語対での SI データが作成されてきたものの [5-8]、Offline データと比較して極めて少量であり、モデルの学習に利用することが難しい背景があった。SimulST モデルの学習の際には、大規模データで事前学習された事前学習済みモデルに対して Fine-tuning する学習方法がまず考えられる。しかしながら、多量の Offline データで学習された事前学習済みモデルに対して、少量の SI データを直接 Fine-tuning させると、過学習の問題が起きやすいという課題がある。過学習が起きると、例えば、SI データの出力長が Offline データと比較して短い上に、明示的な省略の学習ができず、省略するべき部分とそうでない部分の違いを考慮せず過剰に不適切な省略をしてしまうモデルが作成される恐れがある。この課題を克服するために、本研究では新たな学習枠組みを提案する。

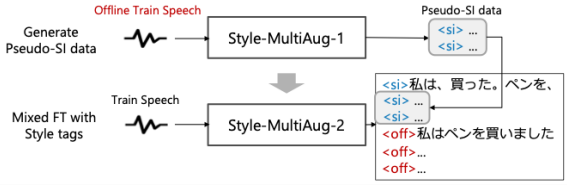


図 4 Multi-stage Fine-tuning

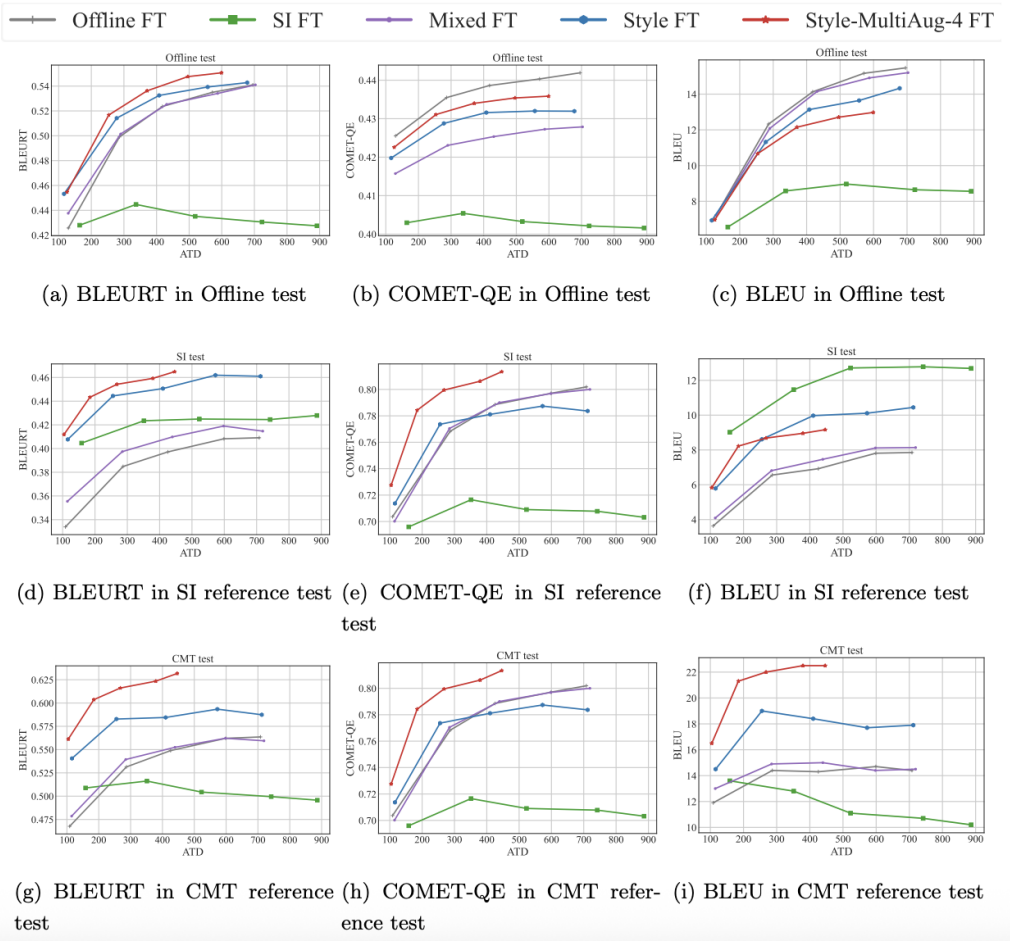
2.4. 提案手法：スタイルタグ付き混合学習

本研究では、該当するテキスト出力が SI スタイルか Offline スタイルのどちらなのか、テキストの先頭に prefix としてスタイルタグを明示的に付与した混合データを用意し、単一の SimulST モデルを学習する手法を提案した。この手法は先行研究[9]における少量の in-domain データと多量に用意可能な out-domain データを用いた Fine-tuning の手法から着想を得ている[9]。スタイルタグの付与により、それぞれのスタイルの出力をモデルが選択して生成するようデコード時にモデルに対して指示を与えることができる。

学習時には、図 3 に示すように、学習データにスタイルタグを付与した混合データを用いて、事前学習済みモデルの Fine-tuning を行う。この際、Offline データには <off>、SI データには <si> というタグをそれぞれのテキストの文頭に付与している。出力時は希望するスタイルのタグを与えて forced decoding を行うことで、そのスタイルの出力を得ることができ、この仕組みによりモデルは入力に応じて出力スタイルを切り替えることが可能となる。これにより、流暢性と品質を重視する Offline タスクと、品質を保ちつつ同時性を重視する SI タスクを一つの SimulST モデルで実現できる。

2.5. 提案手法：タグ付き混合学習とマルチステージ自己学習

また本研究では、メインの提案手法であるスタイルタグ付き混合学習に加えて、手法により一度作成されたモデルを用いて生成された擬似 SI データをさらに次の学習に用いる自己教師あり学習を多段階に行う手法を考えた。これにより、SI データが少量である課題のさらなる軽減を期待できる。図 4 にマルチステージ



(a) BLEURT in Offline test (b) COMET-QE in Offline test (c) BLEU in Offline test

(d) BLEURT in SI reference test (e) COMET-QE in SI reference test (f) BLEU in SI reference test

(g) BLEURT in CMT reference test (h) COMET-QE in CMT reference test (i) BLEU in CMT reference test

図 5 ベースラインおよび提案手法の性能と遅延の関係

自己学習の概要を示す。まず、以上手法により提案モデルを学習する。その後、このモデルに Offline データの原言語音声を入力し、推論時に SI タグ<si>を用いて出力させる。生成された翻訳を擬似 SI データとして再学習に利用する。この過程を多段階的に繰り返すことで、モデルはオリジナルの少量の SI データのみでなく、擬似 SI データも学習の過程で生成しつつ、学習に用いることで、データ少量の問題を軽減しつつ、効率的な SimulST モデルの学習を期待できる。

2.6. 実験設定

本研究では、原言語音声・原言語テキスト・目的言語テキストから構成される ST データを利用し、音声からテキストへの ST モデルを作成する。Offline データは MuST-C v2 [4]、SI データは NAIST-SIC-Aligned-ST [10]を用い、両方とも TED talks をベースにした英日対訳コーパスである。NAIST-SIC-Aligned-ST は NAIST-SIC-Aligned を基としており、原言語のトークと目的言語トークのそれぞれの書き起こし間の文アライメントをとることで構築されている。また、CMT test [11]データについても、原言語に合わせて目的言語をチャンク単位で訳出した、省略の影響を含まない場合の理想的な同時通訳の出力方式を表した順送り訳データとして評価に用いた。

基盤モデルには Fukuda らの提案した base モデルを採用した [12]。これは、事前学習済みの HuBERT [13]をエンコーダ、mBART [14]をデコーダとして初期化し、複数の Offline ST データで Fine-tuning を行ったモデルである。

Offline データおよび SI データをそれぞれで Fine-tuning したものをベースライン（Offline FT, SI FT）とする。次に、提案手法（Style FT）として、Style タグを付与した混合データによる学習を行い、さらなる提

Example 1		Ade.	Flu.
Source	It's probably the smallest of the 21 apps that the fellows wrote last year.	-	-
SI FT (Baseline)	一番小さいアプリです。*(Smallest application.)*	3	3
Style FT (Proposed)	恐らく 21 のアプリの中で、一番小さいものだと思います。*(It is probably the smallest of the 21 applications.)*	4	4
Style-MultiAug-4 FT (Proposed)	これは、おそらく、21 のアプリの中で、最も小さいものです。昨年、フェローが書いたものです。*(This is probably the smallest of the 21 applications. It was written by a fellow last year.)*	5	4
SI reference test	昨年作ってくれたもので。*(It was made last year.)*	-	-
CMT reference test	おそらくそれは最小です、21 のアプリの中で、昨年フェローが書いたものの中で。*(It is probably the smallest, out of the 21 apps the fellows wrote last year.)*	-	-
Example 2			
Source	It was running into bankruptcy last fall, because they were hacked into.	-	-
SI FT (Baseline)	破産したんです。この秋に破産したんです。*(Bankruptcy. I went bankrupt this autumn.)*	3	4
Style FT (Proposed)	これは、去年の秋に、破産したものです。なぜなら、彼らは、ハッキングされたからです。*(These are the ones that, last autumn, went bankrupt. Because they were hacked.)*	5	5
Style-MultiAug-4 FT (Proposed)	それは、昨年、破産につながったものです。なぜなら、彼らは、不正に侵入されたからです。*(It is what led to their bankruptcy last year. Because they were illegally infiltrated.)*	5	5
SI reference test	破産をしたのは、去年の秋なんです。ハッキングをされたからです、*(It was last autumn that I went bankrupt. It was because we were hacked,)*	-	-
CMT reference test	それは、昨秋、破産寸前でした、ハッキングされたためです。*(It was on the verge of bankruptcy last autumn, due to being hacked.)*	-	-

図 6 ベースラインと提案における出力の例

案手法として（Style-MultiAug-*N*FT）として、Style タグ付き混合データに自己教師あり学習を適用し、Style-MultiAug-1 から Style-MultiAug-4 まで段階的に学習を進める多段階学習を設計した。

SimulST の部分出力のデコードの際には Local Agreement (LA) [15] の出力方式を採用した。これは出力を生成するときに、現在のステップでの出力候補文と一つ前のステップの出力候補文の間の共通部分を実際の部分訳出として確定するというものである。入力音声は 200, 400, 600, 800, 1000ms のセグメントに分割し、部分入力に基づく部分出力を生成する形で遅延を制御した。評価では、翻訳品質を BLEU [16]、BLEURT [17]、COMET-QE [18] により測定した。BLEU は参照訳との N-gram 表層一致度による自動評価、BLEURT, COMET-QE は意味的品質の自動評価の手法であり、BLEURT では目的言語の参照訳、COMET-QE では原言語の書き起こし文のテキストを用いて、文の類似度をベースに評価している。また、評価の遅延指標として Average Token Delay (ATD) [19]を、部分出力の訳出開始と終了のタイミングを考慮した遅延評価として採用し、単位は ms とする。

2.7. 実験結果

ベースラインおよび提案手法の性能と遅延の関係を図 5 に示す。結果として、BLEURT および COMET-QE の両指標において提案手法が一貫して性能を向上させることが確認された。特に、自己教師あり学習を取り入れた多段階学習 Style-MultiAug-4 FT は、最も大きな改善を示した。一方、BLEU の評価結果では、SI FT モデルが最も高いスコアとなった。しかしこれは、SI 評価データの参照訳が本来訳出すべき情報を大幅に省略している上で、BLEU で計算されるペナルティが足りないことによって過剰に評価をしていることが原因だと考えられる。図 6 に示すように、SI FT モデルは出力文が短く、SI の参照訳と比較して長さが近く、BLEU が高い傾向にある。ただし、この短縮は「they were hacked into」などの重要な情報を欠落させる場合があり、過剰な省略につながっている。一方で、提案手法モデルは出力がやや冗長になり、「これは」、「なぜなら」といった接続語や指示語などを多めに含む傾向が見られた。しかし、これらは省略された場合も翻訳の意味内容に大きく影響せず、本研究では原言語に含まれる重要な情報を目的言語側でも保持して翻訳する

上で提案手法が有益であることがわかった。特に、自己教師あり学習を段階的に導入した場合には、品質を保ちつつ遅延の低い翻訳が可能だとわかった。また、参照訳の省略傾向が強い SI データに対しては BLEU では過剰評価してしまう傾向があることが SI データと CMT データの比較からわかり、SI タスクの評価においては BLEU を指標とする評価は適切ではない場合があり、COMET-QE のように原言語テキストを基準に含む指標が有効であることが分かった。

3. まとめと今後の展望

本章では、SimulST におけるデータミスマッチの問題に対処するために、SI データを活用した学習手法を提案した。スタイルタグ付き混合学習とマルチステージ自己学習を組み合わせることで、低遅延かつ高品質な翻訳を実現し、実環境で利用可能な SimulST の実現に向けた重要な一歩を示した。本博士論文は、自発発話への頑健性と SimulST の低遅延性という二つの課題に対し、音声の曖昧性を考慮した ST、非流暢性を明示的に考慮した ST、SI データを活用した SimulST という三点の解決策を統合的に提示した。特に SimulST において、スタイルタグと自己学習を組み合わせることで、Offline データでのタスクミスマッチを克服し、語順差の大きい言語ペアにおいても低遅延かつ自然な翻訳を実現した点は大きな貢献である。今後は、大規模な SI データの収集、多言語ペアへの拡張、さらに大規模言語モデルとの統合といった方向性が期待される。

参考文献

[1] Yuka Ko. Optimizing Speech Translation for Low Latency and High Robustness.博士論文, NAIST, 2025.

[2] Ma et al. STACL: Simultaneous Translation with Implicit Anticipation and Controllable Latency. In Proc. ACL, 2019.

[3] Ma et al. Monotonic Multihead Attention. In Proc. ICLR, 2020.

[4] Di Gangi et al. MuST-C: A Multilingual Speech Translation Corpus. In Proc. NAACL-HLT, 2019.

[5] Toyama et al. CIAIR Simultaneous Interpretation Corpus. In Proc. Oriental COCOSDA, 2004.

[6] Shimizu et al. Constructing a Speech Translation System Using Simultaneous Interpretation Data. In Proc. IWSLT, 2013.

[7] 松下ら. 英日・日英通訳データベース. Invitation to Interpreting and Translation Studies, 22:87–94, 2020.

[8] Doi et al. Large-scale English–Japanese Simultaneous Interpretation Corpus: Construction and Analyses with Sentence-aligned Data. In Proc. IWSLT, 2017.

[9] Chu et al. An Empirical Comparison of Domain Adaptation Methods for Neural Machine Translation. In Proc. ACL, pages 385–391.

[10] Zhao et al. NAIST-SIC-aligned: An aligned English–Japanese simultaneous interpretation corpus. In Proc. LREC-COLING 2024, pp. 12046–12052, 2024.

[11] 福田ら. 順送り訳データに基づく英日同時機械翻訳の評価. In IPSJ-SIGNL, 2024-NL-259(14):1–6, 2024.

[12] Ko et al. Tagged End-to-End Simultaneous Speech Translation Training Using Simultaneous Interpretation Data. In Proc. IWSLT 2023, pp. 363–375, 2023.

[13] Hsu et al. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. IEEE/ACM TASLP, 30:162–175, 2022.

[14] Liu et al. Multilingual Denoising Pre-training for Neural Machine Translation. TACL, 8:726–742, 2020.

[15] Liu et al. Low-Latency Sequence-to-Sequence Speech Recognition and Translation by Partial Hypothesis Selection. In Proc. Interspeech 2020, pp. 3620–3624, 2020.

[16] Papineni et al. BLEU: A Method for Automatic Evaluation of Machine Translation. In Proc. ACL 2002, pp. 311–318.

[17] Thibault Sellam et al. BLEURT: Learning Robust Metrics for Text Generation. In Proc. ACL 2020, pp. 7881–7892.

[18] Rei et al. COMET: A Neural Framework for MT Evaluation. In Proc. EMNLP 2020, pp. 2685–2702.

[19] Kano et al. Average Token Delay: A Duration-aware Latency Metric for Simultaneous Translation. Journal of Natural Language Processing, 31(3):1049–1075, 2024.

第 2 回 AAMT 若手翻訳研究会最優秀賞受賞記念

What Language Do Japanese-specialized Large Language Models Think in?

Chengzhi Zhong[1], Fei Cheng[1], Qianying Liu[2], Junfeng Jiang[3], Zhen Wan[1],
Chenhui Chu[1], Yugo Murawaki[1], Sadao Kurohashi[1,2]
[1]Kyoto University, 2National Institute of Informatics, [3]The University of Tokyo

1. Introduction

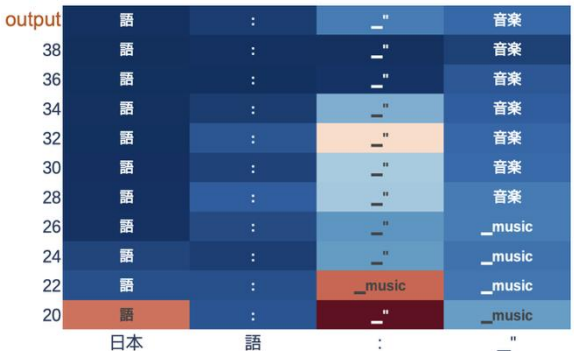
Recent studies have shown that English-centric large language models (LLMs) display distinct patterns in their intermediate layers, where the language distribution is heavily skewed towards English when generating underrepresented languages [1].

This raises our interest in investigating whether LLMs utilize the dominant non-English languages from their training corpora in their intermediate layers during generation. We examine three typical categories of models that are used to process Japanese: Llama2 [2], an **English-centric model**; along with two Japanese-specialized models Swallow [3], an English-centric model with **continued pre-training (CPT)** in Japanese; and LLM-jp [4], a model pre-trained on **balanced corpora of English and Japanese**. More details of these models are shown in Table 1.

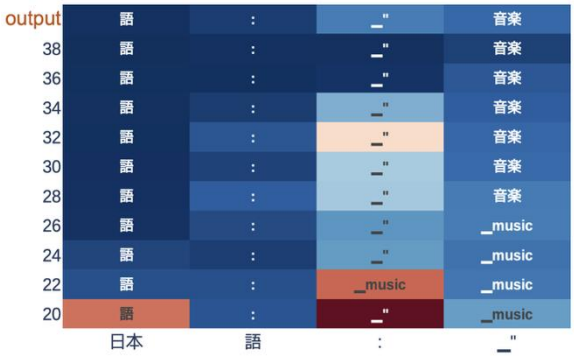
To investigate how the LLMs behave in the intermediate layers, we employ the logit lens method [5], which unembeds each layer's latent representation into the vocabulary space. We verify the latent languages of the three types of models when processing Japanese: While Llama2 uses English as its latent language [1], in contrast, the Japanese CPT model Swallow utilizes both English and Japanese within its intermediate layers, as shown in Figure 1 (a). Meanwhile, Figure 1 (b) shows LLM-jp primarily utilizes Japanese as the latent language in this case.

To further find out the models' latent language when generating languages other than the dominant Japanese and English. We introduce a new setting in which non-Japanese and non-English languages are used as input and target languages to explore the behaviors of the intermediate layers.

Our experiments show that in intermediate layers of the models, the latent language of Japanese-specialized models



(a) Japanese CPT: Swallow



(b) Balanced English and Japanese: LLM-jp

Figure 1: Logit lens results of Japanese-specialized models, (a) Swallow, (b) LLM-jp. The input prompt is “Français: "musique" - 日本語: "”, a French-to-Japanese translation task with the answer “音楽” (music). The figure shows the highest probability token from the intermediate layers, starting from layer 20.

is a distribution over English and Japanese, with the probabilities of these distributions depending on their similarity to both input and target language. In the final layers, the internal predictions transform into the corresponding target language.

Table 1: Categorization of multilingual models based on language proportion and training strategy.

Model Category	Model	Proportion in pre-training data			Token	From scrath
		En	Ja	Other		
English-centric	Llama 2	89.7%	0.1%	10.2%	2,000B	Yes
Japanese CPT	Swallow	10.0%	90.0%	0.0%	100B	Llama-2 based
Balanced English and Japanese	LLM-jp	50.0%	50.0%	0.0%	300B	Yes

In summary, we confirm that Japanese-specialized models Swallow and LLM-jp exhibit two latent languages, English and Japanese. The utilization of these latent languages epends on their similarity to the input and target languages, reflecting a dynamic adjustment in internal language processing.

2. Related Work

2.1. Multilingual Large Language Models

Current frontier large language models, such as GPT-4 [6], Gemini [7], and Llama-2 [2], are primarily trained with English-centric corpora, with other languages constituting only a small portion of the training data. Researchers have sought to enhance these models' multilingual capabilities through various methods. One approach involves continued pre-training with second-language data [8][9][10][11][12], as demonstrated by models like Swallow [3] based on Llama-2. While these approaches have proven effective, ongoing research aims to discover more efficient techniques to further improve the multilingual capabilities of large language models.

2.2. Mechanistic Interpretability

Mechanistic interpretability is the study of understanding how machine learning models work by analyzing their internal components and processes to elucidate the mechanisms that give rise to their behavior and predictions. It encompasses research lines like superposition [13], sparse autoencoders [14], circuit analysis [15] and so on. Within these studies, logits lens [5] and tuned lens [16] focus on decoding the probability distribution over the vocabulary from intermediate vectors of the model, aiding in the comprehension of how the model generates text in the target language.

Previous study [1] showed that Llama-2 models have an abstract “concept space” that lies closer to English than to other languages. When Llama-2 models perform tasks such as translation between non-English languages, the probabilities in the intermediate layers initially focus on the English version of the answer and gradually shift to the target language.

In this work, we expand previous work and utilize these tools to study the distribution of latent languages in different categories of Japanese-specialized LLMs and examined how the probability of internal latent languages is associated with the target language.

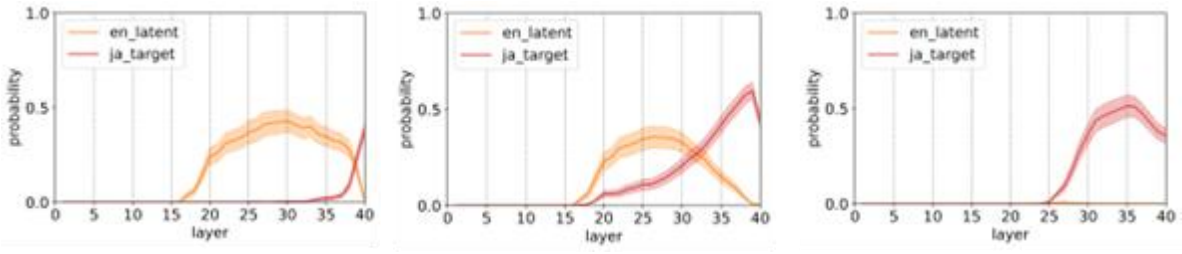
3. Method

3.1. Logit Lens

In the last layer, LLMs use an unembedding matrix to project the hidden vectors onto the vocabulary dimensions.

Then, a softmax function is applied to determine the output token. This process is called unembedding.

By applying the same unembedding operation to the hidden vectors passed between the intermediate layers, we can obtain tokens generated by intermediate layers. Logit lens is a tool designed to achieve this purpose. Therefore, we leverage logit lens to calculate the probability for the model's intermediate layers to generate a specific token sequence.



Llama2-13b Swallow-13b LLM-jp-v2.0

Figure 2: Comparison of English-centric and Japanese-specialized models when processing Japanese Cloze. X-axes denote layer's index of the model, and y-axes denote probability of answer in each language. Translucent area show 95% Gaussian confidence intervals.

3.2. Task Design

Dataset Construction. We first collect parallel words in four languages—English, French, Japanese, and Chinese. To obtain word pairs with different spellings but identical meanings, we construct this dataset based on part of the *Database of Japanese Kanji Vocabulary in Contrast to Chinese* (JKVC) [17]. Then, we use GPT-4 to do translation and obtain the corresponding English and French words or phrases, and then manually review and correct errors. The total size is 166. Based on the parallel words and following previous studies, we demonstrate the following prompts for two tasks, and the corresponding answers for examples will be the same Japanese word “原則” (principle). Models are asked to predict the answer, and we calculate the probability of the answer in the language we want to monitor. We use 4-shot for translation task and 2-shot for cloze task.

Translation task:

Français: "principe" - 日本語:"

Cloze task:

"__"は、基本的なルールや信念です。答え: "

4. Results

4.1. Analysis on Processing Dominant Language – Japanese

To investigate which latent language is used when processing Japanese, we conduct experiments to compare the latent language behaviors of three models when processing cloze task with Japanese set as the target language.

The average result of cloze task is shown in Figure 2. Llama2, which is an English-dominant model, exhibits using English as latent language in its intermediate layers. In contrast, Swallow, which underwent CPT in Japanese, demonstrates a noticeable probability of Japanese in its intermediate layers. For LLM-jp, English probabilities are nearly absent in the intermediate layers. This indicates that these Japanese-specialized models lean to utilize Japanese more as the latent language when processing Japanese.

4.2. Analysis on non-Dominant Languages

We further investigate which latent language the models use when generating non-dominant languages, such as French and Chinese, compared to dominant languages. For this part, we test the models on translation tasks between different languages.

The average result is shown in Figure 3, the source language is always English. When the target language is also English, it becomes a repetition task. Following a left-to-right order, we gradually change the target language. It is observed that for both Swallow and LLM-jp, as the target language gets closer to Japanese, the probability of Japanese in the intermediate layers increases while that of English decreases. Additionally, for Swallow, English and Japanese are consistently intermixed in the intermediate layers, whereas for LLM-jp, the usage of English and Japanese in the intermediate layers is more isolated.

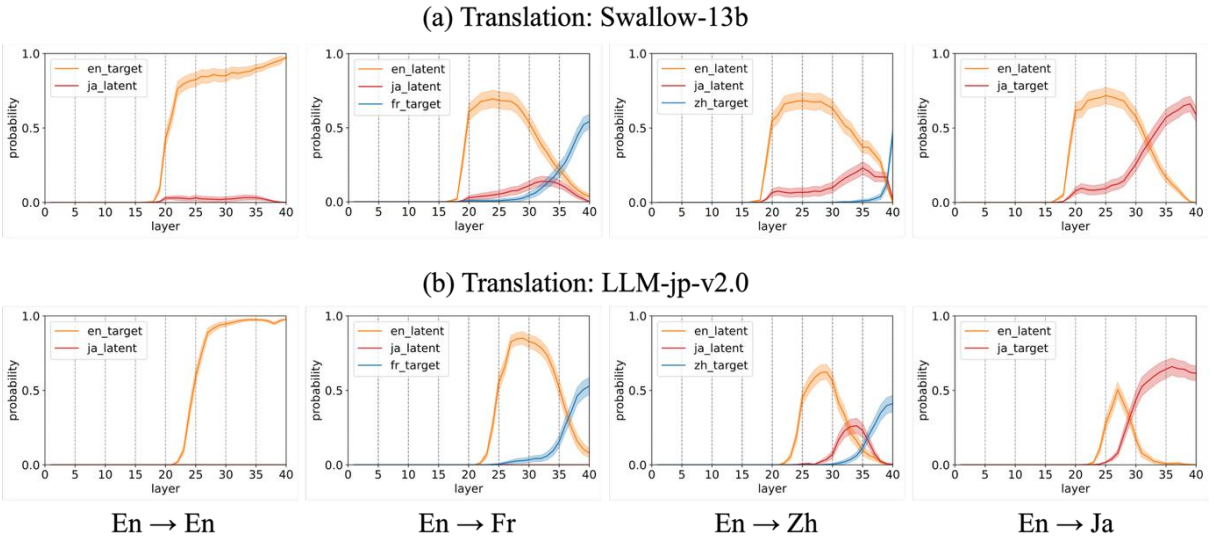


Figure 3: Translation task results of two models with a fixed target language and varying source languages. (a) results for Swallow-13b, (b) results for llm-jp-v2.0. X-axes denote layer's index of the model, and y-axes denote probability of answer in each language. Translucent area show 95% Gaussian confidence intervals.

We also investigate how the source language affects the probability distribution of latent languages. In this case, we set the target language to Japanese. We gradually change the source language to increase its similarity to Japanese. The results are similar that the probability of Japanese in the intermediate layers increases while that of English decreases. In the selection of latent languages in the intermediate layers, the source language has a similar influence to the target language.

The results indicate that the activation of latent languages in LLMs depends on their similarity to the input and target languages.

4.3. How Is Culture Conflict QA Solved?

Because the models 'think' in latent languages, whether this affects the model's reasoning in QA tasks is a question worth discussing. Because some questions can have different answers in different cultural contexts across languages. Thus, we conduct a case study on this topic and use the logit lens to observe the intermediate layers of the models.

As shown in Figure 4, we ask the models about the start date of the school year in Japan with Japanese prompt. In Japan, the new school term begins in April. Llama-2's English-dominant intermediate layers prefer the answer "September/nine," which is the typical start date for American schools. The correct answer for Japan only appears in the latter layers where the probability is concentrated on the target language. In Swallow, the wrong answer "九" (nine) only appear once in layer 36. In contrast, the bilingual-centric LLM-jp does not exhibit this issue. You can see in the early layers that other numbers like "八" and 1 appear. But it is likely due to the chaotic state in the early layers before the answer is determined. This indicates that, for such questions, the knowledge in the primary language context significantly influences the model's predictions. This provides an internal perspective on why operations like knowledge editing should focus on the model's primary language.

5. Conclusion

In this study, we demonstrate that the latent language of LLMs is majorly determined by the language of its training corpora. We confirm that Japanese-specialized Swallow and LLM-jp both utilize Japanese as their latent language when processing Japanese input.

Given that Swallow and LLM-jp exhibit the use of two internal latent languages, the degree to which each latent language is utilized depends on its similarity to the input and target languages. When the input language is more similar to

(a) English-centric: Llama-2-13b (b) Japanese CPT: Swallow-13b (c) Balanced English and Japanese: LLM-jp-v2.0

Figure 4: Results of culture conflict question. We use one-shot format prompts. The question is:「日本の学校新学期が始まる月は：__月、答え：」(The month when the new school term starts in Japan is: _ month, answer:). The correct answer is 「四」(April). The colors in the figures represent entropy: blue indicates the probability is concentrated on the top tokens, while red means it is dispersed across the vocabulary.

Japanese, the proportion of Japanese in the intermediate layers increases, and the same applies to the target language. Additionally, For Swallow, the internal latent language distribution consistently includes both English and Japanese, with English being more dominant. In contrast, LLM-jp tends to favor a single language.

In future research, we aim to extend our investigation to models with other specific dominant languages, such as Chinese, French, and Arabic, to further explore the behavior and mechanisms of non-English-centric LLMs.

Acknowledgment

This work was supported by the “R&D Hub Aimed at Ensuring Transparency and Reliability of Generative AI Models” project of the Ministry of Education, Culture, Sports, Science and Technology.

Reference

[1] Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. Do llamas work in English? on the latent language of multilingual transformers. arXiv preprint arXiv:2402.10588, 2024.

[2] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Niko- lay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open founda- tion and fine-tuned chat models. arXiv preprint arXiv:2307.09288, 2023.

[3] Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. Continual pre-training for cross-lingual LLM adaptation: Enhancing Japanese language capabilities. arXiv preprint arXiv:2404.17790, 2024.

[4] Akiko Aizawa, Eiji Aramaki, Bowen Chen, Fei Cheng, Hiroyuki Deguchi, Rintaro Enomoto, Kazuki Fujii, Kensuke Fukumoto, Takuya Fukushima, Namgi Han, et al. LLM-jp: A cross-organizational project for the research and development of fully open Japanese LLMs. arXiv preprint arXiv:2407.03963, 2024.

[5] Nostalgebraist. Interpreting gpt: The logit lens. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens, 2020. Accessed: 2024-07-28.

[6] 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.

[7] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805, 2023.

[8] Fan-Keng Sun, Cheng-Hao Ho, and Hung-Yi Lee. Lamol: Language modeling for lifelong language learning. In International Conference on Learning Representations, 2020.

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. Advances in neural information processing systems, Vol. 33, pp. 1877–1901, 2020.

[10] Zoltan Csaki, Bo Li, Jonathan Li, Qiantong Xu, Pian Pawakapan, Leon Zhang, Yun Du, Hengyu Zhao, Changran Hu, and Urmish Thakker. Sambalingo: Teaching large language models new languages, 2024.

[11] Yiming Cui, Ziqing Yang, and Xin Yao. Efficient and effective text encoding for Chinese llama and alpaca. arXiv preprint arXiv:2304.08177, 2023.

[12] Julie Hunter, J.r.me Louradour, Virgile Rennard, Isma.l Harrando, Guokan Shang, and Jean-Pierre Lorr. The claire French dialogue dataset. arXiv preprint arXiv:2311.16840, 2023.

[13] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. arXiv preprint arXiv:2209.10652, 2022.

[14] Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In The Twelfth International Conference on Learning Representations, 2023.

[15] Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. In The Eleventh International Conference on Learning Representations, 2022.

[16] Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. arXiv preprint arXiv:2303.08112, 2023.

[17] 松下達彦, 陳夢夏, 王雪竹, 陳林柯. 日中対照漢字語データベースの開発と応用. 日本語教育, Vol.177, pp. 62–76, 2020.

第 2 回 AAMT 若手翻訳研究会優秀賞受賞記念

AoGu:A Japanese-English literary parallel corpus from Aozora Bunko and Project Gutenberg

Guanyu Ouyang[1], Xiaotian Wang[1], Takehito Utsuro[1], Masaaki Nagata[2]

[1]Deg. Prog. Sys.&Inf. Eng., Grad. Sch. Sci.&Tech., University of Tsukuba

[2]NTT Communication Science Laboratories, NTT Corporation

Abstract

This paper introduces a Japanese-English parallel corpus composed of literary works, constructed mainly using bilingual texts from Aozora Bunko and Project Gutenberg. Existing Japanese-English parallel datasets, such as JParaCrawl, JaParaPat, and ASPEC [1, 2, 3], offering coverage of common, patent, and academic domains, they lack resources specifically designed to address discourse-level phenomena and context-aware translation challenges which are existed in literary translation task. To bridge this gap, we build upon the "English-Japanese Translation Alignment Data"[1] developed over a decade ago, updating and expanding it to better support research in discourse-level literary translation and document-level context modeling. Baseline experiments with transformer models on the constructed dataset demonstrate limited performance, highlighting the inherent challenges of literary translation and underscoring the need for more advanced methodologies and resources to enhance translation quality for literary texts.

1. Introduction

Neural Machine Translation (NMT) has advanced significantly in recent years, driven by innovations in neural architectures and the availability of large-scale parallel corpora. While these developments have greatly improved general translation tasks, literary translation presents unique challenges. It demands capturing nuanced semantic meanings and addressing complex discourse-level phenomena, such as pronoun resolution, inter-sentential consistency, and topic coherence [4, 5, 6, 7]. Traditional MT models often struggle with these aspects, resulting in translations that lack stylistic fidelity, contextual awareness, and narrative coherence. To address these issues, researchers have increasingly turned to context-aware and document-level translation approaches that incorporate broader contextual information into the translation process [8, 5].

Lin et al. [9] noted that the poor performance of context-aware MT models often stems not from their inability to handle long-distance dependencies but from the sparsity of discourse-level phenomena in existing datasets. This underscores the critical need for datasets that include such complex linguistic features, alongside advancements in translation models. Meanwhile, recent studies [8, 5] have highlighted literary translation as an ideal testbed for advancing context-aware MT, given the inherent complexity and abundance of discourse-level phenomena in literary texts.

However, resources for Japanese-English literary translation remain scarce. The only existing dataset, the "English-Japanese Translation Alignment Data" [10], was developed over a decade ago and lacks the scale and depth required for modern research. To address this gap, this study builds upon and significantly expands the existing dataset, providing a more comprehensive resource for Japanese-English literary translation. The updated dataset aims to better support research into context-aware and document-level translation methods for Japanese-English language pair.

[1] https://att-astrec.nict.go.jp/member/mutiyama/align/index.html

青空文庫およびプロジェクト・グーテンベルクを対象とした日英文学対訳コーパス AoGu

欧陽冠宇，王小天，宇津呂武仁，永田昌明

筑波大学大学院 システム情報工学研究群，NTT コミュニケーション科学基礎研究所

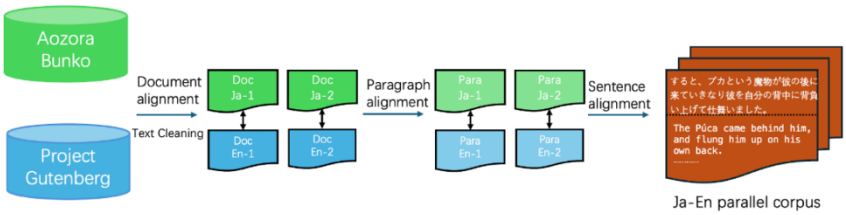


Figure 6 Pipeline of constructing the corpus

2. Related Works

Jin et al. [9] developed a paragraph-aligned Chinese-English dataset containing 10,545 parallel paragraphs extracted from six public-domain novels. This dataset aims to promote research into paragraph-level context-aware MT.

Thai et al. [5] introduced Par3, a multilingual dataset of 121,385 paragraphs from public-domain novels, . Despite its broad scope, the Japanese-English portion remains small, with only 1,857 paragraphs with averaging 4.4 sentences per paragraph(~8,170 sentences).

Jin et al. [11] constructed a large Chinese-English dataset with 5,373 paragraphs, consisting of 548.5K English and 700.9K Chinese sentences. They proposed the challenging chapter-to-chapter (Ch2Ch) translation setting, which showcases the importance of datasets reflecting complex discourse phenomena for literary texts.

Jiang et al. [12] extended the existing BWB [13] corpus with 15,095 discourse-level annotations across 80 documents (~150K words) to better explore the literary MT.

3. Dataset

3.1. Aozora Bunko

Founded in 1997, Aozora Bunko[2]is a digital library providing access to a vast array of public domain works, with a current collection exceeding 17000 items. Moreover, literary works dominate the collection, accounting for approximately 72.4% of the total, with 15,696 titles categorized under this genre alone.

3.2. Project Gutenberg

Project Gutenberg[3], established in 1971 by Michael S. Hart, is the first large-scale digital library dedicated to providing free access to public domain works. It offers over 60,000 texts across genres such as literature, philosophy, history, and science. A notable feature is its collection of professionally translated texts, which ensures high-quality translations for research and linguistic analysis.

3.3. Dataset construction

The main process of dataset construction, as shown in Figure 1, consists of four key steps: document alignment, text preprocessing, paragraph alignment, and sentence alignment.

3.3.1. Document alignment

A random inspection of works from Aozora Bunko and Project Gutenberg (English works) revealed notable differences in their textual characteristics. Most works in Aozora Bunko are partial chapters of novels, individual pieces from collections, or excerpts chosen based on the translator's preferences, rather than complete works. In contrast, most works in Project Gutenberg are complete novels or fully compiled series. This highlights that potential parallel document pairs often differ significantly in content, with an single Aozora Bunko work typically aligning to only a small portion of a single Project Gutenberg work. Based on this observation, rather than relying on traditional semantic text similarity methods for mining parallel document pairs, we leveraged the capabilities of pre-trained large-scale language models, specifically GPT-4o[4] and Claude-3.5-Sonnet[5], to assist in document alignment.

We adopt a 2-stage approach:

1. : For each work in Aozora Bunko, we extract the first 3–5 lines of the text, which typically include the title of the work, the original author's name, and the translator's name. We define a pre-trained model as a retrieve-agent,

[2] https://www.aozora.gr.jp/
[3] https://www.gutenberg.org/
[4] https://openai.com/index/gpt-4o-system-card/
[5] https://www.anthropic.com/news/claude-3-5-sonnet

Using a predefined prompt, we aim for the retrieve-agent to provide the English title of the chapter, the potential associated work title, and the original author's name in English. The details of the prompt are shown in Table 5. Then we implemented an automated script to perform a global character-level match across all metadata of English works in Project Gutenberg using the retrieval information provided by the retrieve-agent. For cases where the retrieve-agent returns "No match" or there are no matching results in Project Gutenberg, we defined a RAG-agent, we first eliminates Japanese works for which they have matched English works. For the remaining Japanese works, we also request retrieval information from the retrieve-agent. If no matches are found, the RAG-agent extracts the first three and last three lines of the text body of Japanese work and sends an updated query to the retrieve-agent. The RAG-agent works to a maximum of three iterations for each Japanese work. The implementation involves RAG-agent module are based on the multi-agent open-source framework AutoGen [14].

2. : We manually reviewed each parallel document obtained from Stage 1, labeled specific chapters in the English works that correspond to the Japanese works, and removed all non-parallel pairs as well as non-English documents from Project Gutenberg. As a result, we obtained a total of 632 parallel document pairs.

3.3.2. Text cleaning

For the Japanese works, we removed the header descriptions and symbol explanations, eliminated phonetic annotations (such as kana readings and kanji readings), deleted input annotations and special character marks, and removed copyright information at the bottom. Additionally, we replaced the iteration mark "／＼" with the vertical kana repeat mark (U+3031) and replaced "／″ ＼" with the vertical kana repeat with voiced sound mark(U+3032).

For the English works, we removed all illustration tags and all annotation information.

3.3.3. Paragraph alignment

Using the labeling information from Stage 2 of document alignment, we extracted paragraphs from the English documents. The final parallel paragraphs consist of the original documents of the Japanese works and the corresponding chapters from the English documents.

3.3.4. Sentence alignment

In the presence of irregular line breaks within the text, including intra-sentence line breaks, we merged all lines within each paragraph for both English and Japanese works. Subsequently, we applied the sat-12l-sm model [15] from wtpsplit [16] to perform sentence segmentation on the merged paragraphs, setting a threshold of 0.01 to achieve finer-grained sentence segmentation. Because we aim to use Vecalign [17] to achieve a more reasonable granularity of parallel sentences.

For all segmented sentences, Vecalign was utilized to perform sentence alignment across all parallel paragraphs. The parameters were configured with an overlap size of 12 and a maximum allowable number of merged sub-sentences set to 12. The embedding models employed included the LaBSE model [18] and the LASER2 model [19].

Table 1 Statistics of AoGu and Utiyama's dataset. #subword refers to the total number of subwords, #sentence refers to the total number of sentence pairs, #doc refers to the total number of document pairs

Embedding Model	#subword (Japanese)	#subword (English)	#sent	#doc	#subword/sent (Japanese)	#subword/sent (English)	#sent/doc
LaBSE	9.73M	7.37M	292,298	513	33.3	25.2	569.8
LASER2	9.72M	7.16M	311,265	513	31.2	23.0	606.8
Utiyama's dataset	2.44M	1.72M	109,431	160	22.3	15.8	683.9

3.4. Dataset statistics

We completed sentence alignment for 513 out of the 632 parallel documents. For sentence embedding, we employed both the LaBSE and LASER2 models. Table 1 presents detailed statistics of the sentence-level datasets initially constructed using these two embedding models. To compute the number of subwords, the tokenizer from the LaBSE model was utilized.

In 2003, Masao Utiyama et al. developed a Japanese-English parallel corpus[6], aligned at the sentence level,

[6] https://att-astrec.nict.go.jp/member/mutiyama/align

utilizing resources from Aozora Bunko, Project Gutenberg, and Project Sugita Genpaku, et al. This corpus is primarily composed of literary works and poetry, encompassing a total of 160 documents in both Japanese and English. AoGu was built upon this foundation and further updated and expanded. To compare the specific differences, The rows of Utiyama's dataset in Table 1 presents the statistical information of the dataset developed by Masao Utiyama et al.

4. Baseline Experiment and Case analysis

We sampled the two datasets obtained using LaBSE and LASER2 with the LaBSE model, setting up two sampling groups with thresholds of 0.4 and 0.6. Four 6-layer transformer baseline models were trained on the sentence-level dataset using Fairseq [20]. The specific parameter settings are as follows: the Adam optimizer was used, with a label smoothing value of 0.1, a dropout rate of 0.3, an initial learning rate of 4e-4, 3000 warm-up update steps, a maximum of 6144 tokens per batch, an update frequency of 4, and a total of 50 epochs.For evaluation, the BLEU [21] and COMET [22] metrics were adopted, with a beam search size of 4. The COMET model used is wmt22-comet-da [23]. The specific results are shown in Table 2. All experiments are conducted on two A6000 GPUs.

Table 2 The baseline of the sentence-level dataset for 4 different configurations

Method	Dataset Size			Metrics	
	Train	Valid	Test	COMET	BLEU
Vecalign (LaBSE) + LaBSE sampling (>0.4)	260,802	13,041	13,041	0.683	8.08
Vecalign (LaBSE) + LaBSE sampling (>0.6)	201,083	10,055	10,055	0.688	8.18
Vecalign (LASER2) + LaBSE sampling (>0.4)	272,812	13,640	13,640	0.680	11.83
Vecalign (LASER2) + LaBSE sampling (>0.6)	224,702	11,235	11,235	0.685	11.64

From Table 2, it can be observed that the BLEU scores for the four baseline settings are relatively low, while the COMET scores are comparatively higher. Table 4 illustrates four translation cases under one specific setting (Vecalign (LASER2) + LaBSE sampling with similarity > 0.4), where the model's understanding of complex semantics is constrained to the sentence level.

For case 1, the source sentence reflects the speaker's perspective (Gryde speaking), whereas the reference adopts the listener's perspective (people listening). The model maintained the source's perspective. Additionally, "夢中になって" can be ambiguous, describing either the speaker's state (chosen by the model) or the listener's state (chosen by the reference).

In case 2, the source text uses "手 紙" (letter) as the pronoun, and the reference preserves "letter" in the same role. However, the model replaces it with "he," altering the original perspective. This demonstrates the model's insufficient understanding of contextual coherence.

For case 3, the model failed to handle pronouns correctly, and compared to the model's direct translation "put his foot to my house twice," the reference translation leans more toward a free translation: "you would never have put another foot." Additionally, the reference tends to use the free translation rather than direct translation: "そいつ ぁ 間違え っこ なし だ 。" -> "you may lay to that ."

For case 4, the reference translation's sentence structures are more diverse, reflecting the characteristics of literary texts, whereas the model's translation tends to adhere closely to the sentence structure of the source text.

Table 3 The baseline settings tested on out-domain ASPEC test set

Method	Dataset Size	Metrics	
	Test	COMET	BLEU
Vecalign (LaBSE) + LaBSE sampling (>0.4)	1,808	0.534	2.4
Vecalign (LaBSE) + LaBSE sampling (>0.6)	1,808	0.518	2.8
Vecalign (LASER2) + LaBSE sampling (>0.4)	1,808	0.539	2.24
Vecalign (LASER2) + LaBSE sampling (>0.6)	1,808	0.529	2.21

These cases reveal that the baseline model trained at the sentence level exhibits limited capabilities in pronoun resolution, modeling complex semantic relationships, and capturing the stylistic and contextual nuances of literary texts. These limitations highlight the need for more advanced approaches, such as paragraph-level or context-aware

training, to improve the model's performance in literary translation tasks.

We also conducted testing on the out-domain ASPEC dataset, and the results are shown in Table 3. The results indicate that the model trained on literary sentence-level data has significantly limited generalization ability, highlighting the substantial differences in characteristics between literary and non-literary texts.

5. Conclusion

This paper introduces a parallel Japanese-English literary corpus, detailing its development process and statistical information. The baseline experimental results demonstrate that literary machine translation tasks impose higher demands on translation models in terms of context awareness, complex semantic relationship modeling, and contextual coherence.

Table 4 Cases for Vecalign (LASER2) + LaBSE sampling with similarity >0.4 settings

#	Metrics	Source	Hypothesis	Reference
1	BLEU = 41.80 COMET = 0.674	二十分間グライドは夢中になって喋った。	"II For twenty minutes Gryde was talking wildly."	"For twenty minutes Gryde was followed with rapt attention."
2	BLEU = 7.24 COMET = 0.674	ここまでは手紙はすこぶる落着いて書いてあったが、ここでペンが急に走り書きになって、筆者の感情が抑え切れなくなっていた。「	Up to this he had written a very quiet note , but here he scribbled a note , and the writer 's feelings relaxed .	So far the letter had run composedly enough, but here with a sudden splutter of the pen, the writer's emotion had broken loose
3	BLEU = 4.72 COMET = 0.681	「もしあんなような奴とつきあってたんなら、二度と己の家へ足を入れさすんじゃなかったぞ。そいつぁ間違えっこなしだ。	"If he had met such a fellow, he wouldn't have put his foot to my house twice, he would have been mistaken."	" If you had been mixed up with the like of that , you would never have put another foot in my house , you may lay to that .
4	BLEU = 9.85 COMET = 0.790	彼の考えそのものが間違いなのか、それとも彼は今、謎の核心へと導かれているのだろうか。」私はひとり考えた。	Was his thoughts doubtless mistaken, or he now led to the point of the mystery?" I thought.	"Either his whole theory is incorrect," I thought to myself, "or else he will be led now to the heart of the mystery."

Table 5 The prompt for retrieve-agent

You are now a distinguished scholar of world literature, with a particular expertise in both Japanese and English literature.

Task:

I will provide you with the name of an author in Japanese and the title of their work in Japanese. Your task is to:

1. Identify the English name of the author.
2. Provide the corresponding English title for the work.
3. If the provided title represents a chapter or section of a larger work, also provide the title of the larger work to which it belongs.
4. If there is no match for one work, please just return "No match".
5. If you are not confident with the result, please list all possible result in each "Author", "Chapter Title" and "Parent Work Title" section.
6. You are also supported by a RAG-agent, in the case I sent the extra content of works, please using this information to further identify.

Guidelines:

Carefully analyze each input to determine whether the given title is a standalone work or part of a larger collection.

Provide accurate and internationally recognized English titles wherever possible.

Always follow the format demonstrated in the example below.

Example:

Q:

アーヴィング ワシントン

ウェストミンスター寺院

A:

Author: Irving, Washington

Chapter Title: Westminster Abbey

Parent Work Title: *The Sketch Book of Geoffrey Crayon, Gent.*

References

[1] M. Morishita, K. Chousa, J. Suzuki, and M. Nagata. JParaCrawl v3.0: A large-scale English-Japanese parallel corpus. In Proc. 13th LREC, pp. 6704–6710, 2022.

[2] M. Nagata, M. Morishita, K. Chousa, and N. Yasuda. JaParaPat: A large-scale Japanese-English parallel patent application corpus. In Proc. LREC-COLING, pp. 9452–9462, May 2024.

[3] T. Nakazawa, M. Yaguchi, K. Uchimoto, M. Utiyama, E. Sumita, S. Kurohashi, and H. Isahara. ASPEC: Asian scientific paper excerpt corpus. In Proc. 10th LREC, pp. 2204–2208, 2016.

[4] E. Matusov. The challenges of using neural machine translation for literature. In Proc. the Qualities of Literary Machine Translation, pp. 10–19, 2019.

[5] K. Thai, M. Karpinska, K. Krishna, B. Ray, M. Inghilleri, J. Wieting, and M. Iyyer. Exploring document-level literary machine translation with parallel paragraphs from world literature. In Proc. EMNLP, pp. 9882–9902, 2022.

[6] M. Fonteyne, A. Tezcan, and L. Macken. Literary machine translation under the magnifying glass: Assessing the quality of an NMT-translated detective novel on document level. In Proc. 12th LREC, 2020.

[7] Y. Liu, Y. Yao, R. Zhan, Y. Lin, and D. Wong. NovelTrans: System for WMT24 discourse-level literary translation. In Proc. 9th WMT, pp. 980–986, 2024.

[8] K. Marzena and I. Mohit. Large language models effectively leverage document-level context for literary translation, but critical errors persist. In Proc. 8th WMT, pp. 419–451, 2023.

[9] J. Lin, J. He, J. May, and X. Ma. Challenges in context-aware neural machine translation. In Proc. EMNLP, p. 15246–15263, 2023.

[10] Utiyama M. and Takahashi M. English-Japanese translation alignment data., 2003.

[11] L. Jin, Li A., and X. Ma. Towards chapter-to-chapter context-aware literary translation via large language models, 2024.

[12] Y. Jiang, T. Liu, S. Ma, D. Zhang, M. Sachan, and R. Cotterell. Discourse-centric evaluation of document-level machine translation with a new densely annotated parallel corpus of novels. In Proc. 61st ACL, pp. 7853–7872, 2023.

[13] Y. Jiang, T. Liu, S. Ma, D. Zhang, J. Yang, H. Huang, R. Sennrich, R. Cotterell, M. Sachan, and M. Zhou. BlonDe: An automatic evaluation metric for document-level machine translation. In Proc. NAACL, pp. 1550–1565, 2022.

[14] Q. Wu, G. Bansal, J. Zhang, Y. Wu, B. Li, E. Zhu, L. Jiang, X. Zhang, S. Zhang, J. Liu, A. Awadallah, R. White, D. Burger, and C. Wang. Autogen: Enabling next-gen LLM applications via multi-agent conversations. In Proc. 1st COLM, 2024.

[15] M. Frohmann, I. Sterner, I. Vulić, B. Minixhofer, and M. Schedl. Segment any text: A universal approach for robust, efficient and adaptable sentence segmentation. In Proc. EMNLP, pp. 11908–11941, 2024.

[16] B. Minixhofer, J. Pfeiffer, and I. Vulić. Where's the point? self-supervised multilingual punctuation-agnostic sentence segmentation. In Proc. 61st ACL, pp. 7215–7235, 2023.

[17] B. Thompson and P. Koehn. Vecalign: Improved sentence alignment in linear time and space. In Proc. EMNLP and 9th IJCNLP, pp. 1342–1348, 2019.

[18] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang. Language-agnostic BERT sentence embedding. In Proc. 60th ACL, pp. 878–891, 2022.

[19] K. Heffernan, O. Çelebi, and H. Schwenk. Bitext mining using distilled sentence representations for low-resource languages. In Findings of EMNLP, pp. 2101–2112, 2022.

[20] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli. fairseq: A fast, extensible toolkit for sequence modeling. In Proc. NAACL, pp. 48–53, 2019.

[21] K. Papineni, S. Roukos, T. Ward, and W. Zhu. Bleu: a method for automatic evaluation of machine translation. In Proc. 40th ACL, pp. 311–318, 2002.

[22] R. Rei, C. Stewart, A. Farinha, and A. Lavie. COMET: A neural framework for MT evaluation. In Proc. EMNLP, pp. 2685–2702, 2020.

[23] R. Rei, J. C. de Souza, D. Alves, C. Zerva, A. Farinha, T. Glushkova, A. Lavie, L. Coheur, and A. Martins. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In Proc. 7th WMT, pp. 578–585, 2022.

第 2 回 AAMT 若手翻訳研究会優秀賞受賞記念

複数の LLM を活用した機械翻訳のための協力デコーディング

白井尚登 衣川和尭 伊藤均 美野秀弥 河合吉彦
NHK 放送技術研究所

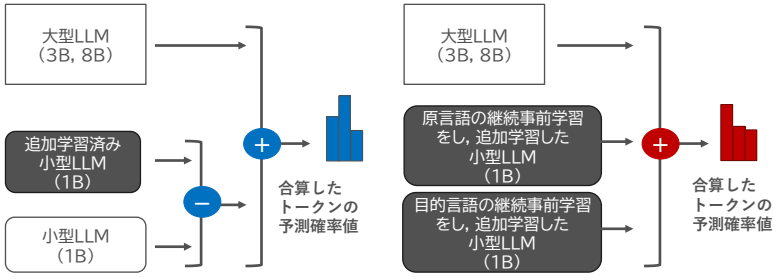


図 1. Proxy-tuning (左)と Collab-MT(右)の概略図.

1. はじめに

大規模言語モデル（Large Language Model, LLM）を機械翻訳分野へ応用する動きが進み，公開された LLM を継続して事前学習する継続事前学習と対訳データの追加学習とを組み合わせた ALMA [1]や，翻訳のワークフローを模して学習を行った Tower [2]など多くの翻訳特化型 LLM が登場している．

翻訳タスクをはじめ，LLM の文章生成タスクへの有効性が示される一方，モデルの大規模化に伴い，学習や推論に必要な計算コストの増加が課題となっている[3]．この課題に対応するため，モデルのパラメータサイズが大きな LLM（大型 LLM）の追加学習ではなく，より小さな LLM（小型 LLM）のみを追加学習し，大型 LLM と小型 LLM の両方を用いて次のトークンを予測することで性能向上を目指す手法（協力デコーディング）が提案されている．例えば，Proxy-tuning [4]とよばれる手法では，小型 LLM の学習前後に出力する予測確率値の差分を大型 LLM に加算することで，質問応答やコード生成のタスクにおいて，直接大型 LLM を追加学習したモデルに匹敵する性能を示した．しかしながら，Proxy-tuning は翻訳タスクには適用しておらず，小型 LLM の学習情報が翻訳性能の向上に寄与するかは未解明である．

本研究では，機械翻訳への応用を目的とした新たな協力デコーディング手法 Collaborative Decoding for Machine Translation (Collab-MT) を提案する．Collab-MT は，小型 LLM の予測確率値を大型 LLM に加算することで，翻訳性能を向上させる手法である．原言語と目的言語に特化した小型 LLM を用いることで，言語間の知識を効果的に活用し，汎用的な大型 LLM の性能を向上させることを目指す．

実験の結果，大型 LLM の追加学習モデルの性能には及ばなかったものの，既存手法 Proxy-tuning よりも BLEU スコア [5]において最大 12.73 ポイント上回った．

2. 提案手法

2.1. 既存手法: Proxy-tuning

従来の協力デコーディング手法 Proxy-tuning（図 1）は，小型 LLM を特定のタスクに合わせて追加学習し，学習前後の予測確率の差分を利用して，大型 LLM の予測を補正することで翻訳性能を向上させる．これにより，大型 LLM を直接追加の学習することなく，特定タスクの性能を向上させることができる．この Proxy-tuning はこれまで機械翻訳のタスクには適用されておらず，機械翻訳のタスクへの効果は明らかではない．

Collaborative Decoding for Machine Translation Using Multiple Large Language Models
Naoto Shirai, Kazutaka Kinugawa, Hitoshi Ito, Hideya Mino and Yoshihiko Kawai
NHK Science & Technology Research Laboratories

表 1. 各手法の翻訳結果. 太字は一番良い結果. 1B, 3B の Collab-MT とは, 小型 LLM が 1B, 大型 LLM が 3B の設定を指す.

モデル	手法	日→英	英→日	独→英	英→独
		BLEU / BERTScore	BLEU / BERTScore	BLEU / BERTScore	BLEU / BERTScore
1B	Zero-shot	15.80 / 0.92	3.39 / 0.78	31.91 / 0.94	23.19 / 0.84
	Fine-tuning	20.54 / 0.93	7.95 / 0.85	31.37 / 0.94	23.67 / 0.85
3B	Zero-shot	24.40 / 0.94	4.67 / 0.83	39.64 / 0.95	31.69 / 0.87
	Fine-tuning	30.08 / **0.95**	8.76 / 0.86	39.53 / **0.96**	31.17 / 0.87
1B, 3B	Proxy-tuning	20.26 / 0.93	5.35 / 0.81	20.15 / 0.93	17.34 / 0.80
	Collab-MT	26.73 / 0.94	7.41 / 0.85	38.02 / 0.95	30.30 / 0.87
8B	Zero-shot	28.87 / **0.95**	7.76 / 0.84	**44.75 / 0.96**	**37.90 / 0.89**
	Fine-tuning	**33.07 / 0.95**	**9.15 / 0.87**	43.18 / **0.96**	35.52 / 0.88
1B, 8B	Proxy-tuning	21.56 / 0.93	6.03 / 0.83	26.45 / 0.94	21.15 / 0.82
	Collab-MT	28.09 / 0.94	7.96 / 0.85	39.18 / 0.95	31.78 / 0.87

2.2. 提案手法: Collab-MT

本研究では，機械翻訳に特化した新たな協力デコーディング手法 Collab-MT を提案する（図 1）. Collab-MT では，2 つの小型 LLM を用意し，それぞれに原言語，あるいは，目的言語に応じた継続事前学習と，翻訳ペアに対して追加学習を行う．そして，これらの小型 LLM の予測確率値を，追加学習していない大型 LLM の予測確率値に加算する．出力結果を加算することにより，協力して文章を生成し，翻訳性能の向上を試みる.

3. 実験設定

本研究では, 提案手法 Collab-MT の有効性を検証するため, 日英および独英の双方向翻訳タスクにおいて, 既存手法との比較実験を実施した.

3.1. データセット

データセットとして，日英・英日翻訳には ALT[1]，独英・英独翻訳には WMT19[2]を使用した．ALT データセットは，継続事前学習に 1000 文，追加学習に 18,083 文，テストに 1,017 文，WMT19 データセットは，継続事前学習に 2,000 文，追加学習に 18,000 文，テストに 2,998 文を使用した. いずれも既存の対訳データを活用し，少量の単言語データによる学習を行った.

3.2. モデル

Llama 3 [6] シリーズのモデルを使用し，大型 LLM には Llama-3.2-3B[3]/ Llama-3.1-8B[4]の Instruct モデルを使用した．小型 LLM には，Llama-3.2-1B[5]を使用した.

3.3. 比較手法

本実験では，追加学習をしない Zero-shot モデル（Zero-shot）と追加学習を行ったモデル（Fine-tuning）をベースラインとし，既存の協力デコーディング手法 Proxy-tuning[6]と提案手法 Collab-MT とを比較した.

また, 翻訳性能の評価には, 正解データと LLM の翻訳結果の表層的な一致度を測る BLEU[7]と意味的な類似度を測る BERTScore [7] を用いた.

4. 実験結果

日英および独英の双方向の翻訳タスクの結果を表 1 に示す. 提案手法 Collab-MT は，すべての翻訳タスクにおいて既存手法である Proxy-tuning を上回る性能を示した. 特に，8B モデルを使用した独英翻訳においては提案手法が既存手法に対して BLEU スコアを 12.73 ポイント改善した．また，BERTScore においても，提案手法は既存手法を上回った.

一方で，Zero-shot の大型 LLM との比較では，日英・英日翻訳では同等の性能を示したが, 8B モデルを使用した英独翻訳では BLEU スコアが 6.12 ポイント下回

[1] https://huggingface.co/datasets/mutiyama/alt
[2] https://huggingface.co/datasets/wmt/wmt19
[3] https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct
[4] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
[5] https://huggingface.co/meta-llama/Llama-3.2-1B
[6] https://github.com/alisawuffles/proxy-tuning
[7] https://github.com/mjpost/sacrebleu

る結果となった．また，追加学習を施したモデル（Fine-tuning）はすべてのタスクにおいて提案手法を上回ったものの，独英・英独翻訳では Zero-shot の BLEU スコアを下回る結果となった．この結果は，本提案手法のみでは大型 LLM の性能を十分に活かせない場合があることを示している．提案手法は小型 LLM と大型 LLM の予測確率値を一律に加算しているが，翻訳対象によって加算する割合や加算するかどうかを判定するなどの手法を検討する必要がある．

5. おわりに

本研究では，機械翻訳における協力デコーディングの新たな手法として Collab-MT を提案し，既存手法との比較を通じてその有効性を検証した．提案手法は，出力の差分ではなく小型 LLM の出力そのものを活用することで，翻訳性能の向上を実現した．実験結果からは，一部の翻訳タスクでは既存手法を上回る性能を示す一方で，学習設定によっては大型 LLM の性能を十分に引き出せない場合があることも確認された．

今後は，より効果的な継続事前学習や追加学習の設計，および小型 LLM の選定・組み合わせを選定することで，協力デコーディングの性能向上に取り組む．なお，本稿は言語処理学会第 31 回年次大会の発表論文に基づいて作成した [8].

6. 謝辞

本研究成果は，国立研究開発法人情報通信研究機構の委託研究（課題 225）により得られたものです．

7. 参考文献

[1] Xu, Haoran, et al. "A paradigm shift in machine translation: Boosting translation performance of large language models." In *International Conference on Learning Representations (ICLR),* 2024.

[2] Alves, Duarte M., et al. "Tower: An open multilingual large language model for translation-related tasks." *arXiv preprint arXiv:2402.17733*, 2024.

[3] Lu, Jinliang, et al. "Merge, ensemble, and cooperate! a survey on collaborative strategies in the era of large language models." *arXiv preprint arXiv:2407.06089*, 2024.

[4] Liu, Alisa, et al. "Tuning language models by proxy." *arXiv preprint arXiv:2401.08565*, 2024.

[5] Papineni, et al. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

[6] Grattafiori, Aaron, et al. "The llama 3 herd of models." *arXiv preprint arXiv:2407.21783*, 2024.

[7] Zhang, Tianyi, et al. "Bertscore: Evaluating text generation with bert." *arXiv preprint arXiv:1904.09675*, 2019.

[8] 白井尚登 他．複数の LLM を活用した機械翻訳のための協力デコーディング．言語処理学会 第 31 回年次大会 発表論文集, 2025.

第 2 回 AAMT 若手翻訳研究会優秀賞受賞記念

多言語文符号化器からの言語非依存な文埋め込みの抽出

福島 啓太
愛媛大学

1. 概要

本研究では、多言語文符号化器から得られる文埋め込みを、言語に依存しない情報である意味表現と、言語固有の情報を持つ言語表現に分離し、言語横断の文間類似度推定の性能を改善する。これにより、言語間の埋め込みの差異を抑制し、文の類似度をより正確に推定できることを確認した。

2. はじめに

近年、文符号化器を用いて文章を文埋め込みに変換し、そのベクトル類似度を自然言語処理タスクに応用する手法が主流となっている[1,2,3]。これを多言語に拡張した多言語文符号化器も盛んに開発されている[4,5,6]が、異言語間の文埋め込みの類似度推定には課題が残されている。その要因は、多言語文符号化器の文埋め込みが文章の意味そのものよりも、言語的特徴に強く影響されてしまう点にある[7]。その結果、同一の意味を持つ文であっても、言語が異なると文埋め込みが乖離し、意味的類似度の正確な推定が困難となる。この問題に対し、先行研究では、文埋め込みから、言語に共通する意味の情報と言語ごとの特徴をそれぞれ別のモデルで抽出・分離するアプローチが取られてきた[7,8]。しかし、この方法では分離の過程で元の文埋め込みが持つ情報の一部が失われ、自然言語処理タスクの精度が低下するという欠点があった。

そこで本研究では、意味の情報を抽出するモデルのみを訓練し、言語表現を元の文埋め込みと意味表現との差分として扱う手法を提案する。この構造により、情報を失うことなく、より正確な意味表現の獲得が見込める。機械翻訳の品質推定タスク[9]で性能を評価した結果、提案手法は従来手法を上回る結果を達成した。

3. 提案手法

先行研究の DREAM[7]と MEAT[8]では、言語に依存しない意味表現と言語固有の情報をもつ言語表現の分離のために、2 つの多層パーセプトロン (MLP)で構成されていたのに対し、本研究では 1 つの MLP で構成する。多言語文符号化器から得られる文埋め込み eを MLP に通し、意味表現 $\hat{e}_M$を抽出する。文埋め込み e と抽出した意味表現 $\hat{e}_M$の減算結果を言語表現 $\hat{e}_L$とすることで、意味と言語情報の分離を目指す。以下に意味表現と言語表現の獲得方法を示す。

$$\hat{e}_M = MLP(e) \quad (1)$$

$$\hat{e}_L = e - \hat{e}_M \quad (2)$$

提案手法では、意味表現抽出器 MLP を以下の 4 つの損失関数に基づくマルチタスク学習によって訓練する。

$$L = L_S + L_M + L_L + L_C \quad (3)$$

文表現分離損失 L_S

本研究では、文埋め込み eから意味表現 $\hat{e}_M$を抽出し、文埋込みと意味表現の差分を言語表現 $\hat{e}_L$として獲得する。この文表現分離損失は意味表現と言語表現を分離させることを促す。原言語文、目的言語文側の両方に以下に損失をとる。

$$L_S = \max(0, \cos(\hat{e}_M, \hat{e}_L)). \quad (4)$$

意味表現損失 L_M

抽出した対訳の意味表現同士は類似させ、対訳でない意味表現同士は類似させない。原言語文 sと目的言語文 tは意味的に等価であるため、原言語文の意味表現 $\hat{s}_M$と目的言語文の意味表現 $\hat{t}_M$を近づけることで MLPを訓練する。また、バッチ内から、ランダムに同言語の文 s', t'を取得し、意味表現 $\hat{s'}_M, \hat{t'}_M$を抽出する。

Extraction of Language-Independent Sentence Embeddings from Multilingual Sentence Encoders
Keita Fukushima
Ehime University

表 1. WMT20 品質推定タスクにおける人で評価とのピアソン相関係数

多言語文符号化器	意味表現	en-de	en-zh	ro-en	et-en	ne-en	si-en	Avg.
mE5-base	Baseline	0.003	0.074	0.674	0.443	0.486	0.463	0.357
	DREAM	0.120	0.213	0.738	0.499	0.527	0.515	0.435
	MEAT	0.119	0.209	0.735	0.500	0.533	0.514	0.435
	提案手法	0.116	0.190	0.741	0.513	0.543	0.525	0.438
mE5-large	Baseline	0.020	0.100	0.734	0.556	0.538	0.493	0.407
	DREAM	0.172	0.257	0.783	0.629	0.584	0.541	0.494
	MEAT	0.117	0.186	0.751	0.610	0.541	0.499	0.451
	提案手法	0.175	0.249	0.782	0.636	0.591	0.544	0.496
mE5-large-instruct	Baseline	0.143	0.203	0.767	0.590	0.549	0.422	0.446
	DREAM	0.212	0.290	0.765	0.595	0.585	0.499	0.491
	MEAT	0.215	0.283	0.757	0.607	0.563	0.476	0.484
	提案手法	0.215	0.284	0.762	0.611	0.598	0.515	0.498

ランダムに得た同言語の意味表現は意味的に等価でないため、それぞれ$\hat{s}_M, \hat{t}_M$と遠ざける。以下に意味表現損失を定義する。

$$L_M = 2(1-\cos(\hat{s}_M, \hat{t}_M)) + \max\left(0, \cos\left(\hat{s}_M, \hat{s'}_M\right)\right) + \max\left(0, \cos\left(\hat{t}_M, \hat{t'}_M\right)\right) \quad (5)$$

ここで、類似させる重みと類似させない重みを一致させるために値が 2 倍になっている点に注意されたい。

言語表現損失 L_L

多言語文符号化器から得られる文埋め込みと抽出した意味表現の差分として得た言語表現は、同一言語同士、つまり、s, s'間およびt, t'間を類似させる。以下に言語表現損失を定義する。

$$L_L = \left(1-\cos\left(\hat{s}_L, \widehat{s'}_L\right)\right) + \left(1-\cos\left(\hat{t}_L, \widehat{t''}_L\right)\right) \quad (6)$$

交差復元損失 L_C

本研究では、文埋め込みを意味と言語に分離することが目的としている。既存研究である DREAM [7]と MEAT [8]は、分離させた意味と言語から文埋め込みを復元するという損失関数を実装しているが、提案手法では、構造的に復元損失を取っても誤差が生じない。そこで、対訳コーパスの意味表現同士を置換したり、同言語の言語表現同士を置換したりすることで、損失を取る。以下に交差復元損失を定義する。

$$L_C = \left(1-\cos\left(\hat{s}, \hat{s}_M + \widehat{s'}_L\right)\right) + \left(1-\cos\left(\hat{t}, \hat{t}_M + \widehat{t'}_L\right)\right) + \left(1-\cos\left(\hat{s}, \hat{t}_M + \hat{s}_L\right)\right) + \left(1-\cos\left(\hat{t}, \hat{s}_M + \hat{t}_L\right)\right) \quad (7)$$

第 1 項と第 2 項は同一言語同士の言語表現を互いに置換できることを、第 3 項と第 4 項は対訳文対同士の意味表現同士を置換できることを表している。

4. 評価実験

WMT20 における機械翻訳の品質推定 (QE：Quality Estimation) タスクで提案手法を評価する。QE は機械翻訳の出力文と、原言語文を入力として、出力文の翻訳品質を推定するタスクである。本研究では、出力文と原言語文の類似度を推定値とする。公式の評価方法に従い，モデルが推定した翻訳品質推定値と人手評価値とのピアソン相関によって性能を評価する。

4.1. 実験設定

データセット WMT20 の QE タスクには、6 つの言語対が含まれている。6 つの言語対の内訳は，英語からドイツ語 (en-de) および英語から中国語 (en-zh) の多資源言語対、ルーマニア語から英語 (ro-en) およびエストニア語から英語 (et-en) の中資源言語対、ネパール語から英語 (ne-en) およびシンハラ語から英語 (si-en) の少資源言語対である。各言語対において、

1,000 文対の原言語文および機械翻訳の出力文と、人手評価値の組が提供されている。評価用の機械翻訳器は fairseq ツールキット[10]を用いて訓練された Transformer モデル[11]である。訓練には、先行研究[8]と同量のコーパスを使用し、多資源言語対は 100 万文対、中資源言語対は 20 万文対、少資源言語対は 5 万文対用いた。

モデル 本研究では、MLP に 1 層のフィードフォワードニューラルネットワークを用いた。多言語文符号化器は mE5[12]を用いた。用いる文埋め込みは最終層の平均プーリングを用いた。提案手法ではバッチサイズを 512、最適化手法を Adam[13]とし、学習率は先行研究[8]に則り、10^{-4} として HuggingFace Transformers[14]を用いて訓練した。検証データは訓練用データから無作為に 10%ずつ抽出し、検証データにおける式 (3)の損失が 5 エポック改善しない場合に訓練を終了した。なお、訓練するのは MLP のみであり、多言語文符号化器は訓練しない。

4.2. 実験結果

表 1 に QE タスクの実験結果を示す。表 1 の各段は上から、元の文埋め込みによる QE タスク、既存手法[7, 8]、提案手法の結果を示したものである。モデル全体の性能の平均値に注目すると、提案手法は既存手法を上回り、その有効性を示す結果となった。特に、中資源言語対の et-en、及び少資源言語対の ne-en、si-en においても一貫して性能が向上した。

5. おわりに

本研究では、多言語文符号化器から得られる文埋め込みを言語に依存しない情報を持つ意味表現と、言語固有の情報を持つ言語表現に分離する手法を提案した。抽出した意味表現を機械翻訳の品質推定タスクに用いることで、提案手法による QE 性能の向上を確認できた。提案手法は QE 性能を向上させただけでなく、既存手法である DREAM、MEAT と比較しても有効性を確認できた。提案手法は、言語表現を獲得する際に、元の文埋め込みと抽出した意味表現の差分によって獲得する。この方法により、抽出時の情報欠落を防ぎ、QE タスクの性能改善に寄与することを確認した。

参考文献

[1] Cer, D., et al.: SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation, In SemEval, pp. 1–14 (2017).

[2] Reimers, N., Gurevych, I.: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, In EMNLP, pp. 3982–3992 (2019).

[3] Wang, L., et al.: Text Embeddings by Weakly-Supervised Contrastive Pre-training, arXiv:2212.03533 (2022).

[4] Reimers, N., Gurevych, I.: Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation, In EMNLP, pp. 4512–4525 (2020).

[5] Feng, F., et al.: Language-agnostic BERT Sentence Embedding, In ACL, pp. 878–891 (2022).

[6] Wang, L., et al.: Multilingual E5 Text Embeddings: A Technical Report, arXiv:2402.05672 (2024).

[7] Tiyajamorn, N., et al.: Language-agnostic Representation from Multilingual Sentence Encoders for Cross-lingual Similarity Estimation, In EMNLP, pp. 7764–7774 (2021).

[8] Kuroda, Y., et al.: Adversarial Training on Disentangling Meaning and Language Representations for Unsupervised Quality Estimation, In COLING, pp. 5240–5245 (2022).

[9] Specia, L., et al.: Findings of the WMT 2020 Shared Task on Quality Estimation, In WMT, pp. 743–764 (2020).

[10] Ott, M., et al.: fairseq: A Fast, Extensible Toolkit for Sequence Modeling, in NAACL, pp. 48–53 (2019)

[11] Vaswani, A., et al.: Attention is All you Need, in NIPS, pp. 5998–6008 (2017)

[12] Wang, L., et al.: Multilingual E5 Text Embeddings: A Technical Report, arXiv:2402.05672 (2024)

[13] Kingma, D. P.Ba, J. L.: Adam: A Method for Stochastic Optimization, in ICLR (2015)

[14] Wolf, T., et al.: Transformers: State-of-the-art Natural Language Processing, in EMNLP, pp. 38–45 (2020)

第 2 回 AAMT 若手翻訳研究会スポンサー賞受賞記念

事前訓練済みモデル・LLM を用いた特許翻訳の二段階自動後編集

武馬光星 [1], 西村 柾人 [1], 宇津呂武仁 [1], 永田昌明 [2]
[1] 筑波大学大学院 システム情報工学研究群 [2]NTT コミュニケーション科学基礎研究所

1. はじめに

大規模言語モデル（LLMs）の近年の進展により，LLMRefine [1] に代表される多段階推論手法が提案され，反復的な分析と修正を通じて翻訳出力を段階的に改善する枠組みが注目されている．さらに，Google の最近の研究 [2] が示すように，この段階的推論を翻訳タスクに応用する試みも進められている．しかし，機械翻訳のすべての処理を LLMs に依存する必要は必ずしもなく，例えば翻訳誤り検出タスクは，事前学習済みの多言語エンコーダモデルを活用することで，より高精度かつ低計算コストで実行できる可能性がある．

本研究では，図 1 に示すように，エンコーダモデルによる誤り検出と LLM による翻訳訂正を統合した二段階の翻訳訂正手法を提案する．第 1 段階では，多言語 BERT（mBERT）[3] を用いて各トークンに翻訳誤りラベルを付与する．日英特許翻訳に関しては誤り注釈付きデータセットが存在しないため，対訳特許コーパスのターゲット文に人工的な誤りを挿入し，合成データセットを構築した．これにより mBERT はトークンレベルでの誤り検出を学習できる．第 2 段階では，LLM（GPT-4o）[4] を用い，検出された誤りタグに基づいて翻訳文を訂正する．

本手法は，翻訳誤りが許容されず厳格な後編集が求められる分野の一例として，特許翻訳に適用し評価を行った．人工誤り文，繰り返し誤り文，訳抜け文の 3 種類のデータセットで実験した結果，エンコーダモデルによる誤り検出と LLM による訂正を組み合わせた手法は，BLEU [5] および COMET [6] の両評価指標において LLM 単独の手法を上回った．ただし，訳抜けの完全な解消は依然として課題として残る．以上より，多段階 LLM 推論は強力である一方，コンパクトなエンコーダモデルを選択的に組み合わせることで，機械翻訳における誤り検出・訂正の精度と効率をさらに高められることが示唆された．

2. 関連研究

Wei らは，多言語 BERT（mBERT）を用いた単語レベル品質推定（QE）の教師あり学習を行った [7]．具体的には，原文と翻訳文を連結した入力を回帰モデルに与え，各トークンが BAD と判定される確率を出力するモデルを構築している．本研究は，この手法を発展させ，特許文書を対象とした教師あり学習に適用することで，特許翻訳における単語レベル QE の精度向上を目指す．

単語レベル QE の最先端手法としては，COMET を拡張し単語レベルスコアを付与する xCOMET [8] が挙げられる．xCOMET は XLM-R [9] を基盤としており，複数の WMT ベンチマークにおいて優れた性能を示している．しかし，XLM-R は学習データが豊富な言語と比較して日本語に関するカバレッジが限定的であり，日英翻訳設定における有効性については依然として明らかではない．

一方，近年は機械翻訳の後編集において LLM を活用する研究も進展している．Xu らは LLMRefine [1] を提案し，翻訳結果に対して LLM による誤り検出を行う枠組みを構築した．具体的には，エラーカテゴリとエラースパンのリストを生成するように学習した LLM により誤りを特定し，そのフィードバックを用いて LLM が反復的に修正を行うことで，英独および中英翻訳タスクにおいて翻訳品質の向上を実現してい

Two Step Automatic Post Editing of Patent Translation based on Pre-trained Models and LLMs
Kosei Buma, Masato Nishimura, Takehito Utsuro, Masaaki Nagata
University of Tsukuba, NTT, Inc.

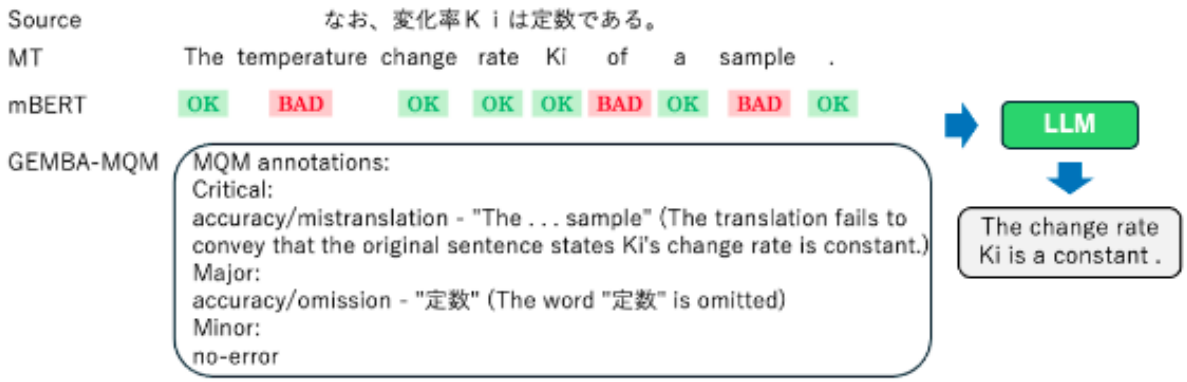


図 1: 誤り検出→ LLM での誤り訂正

る．また，Ki らは外部からのフィードバックを取り込んだ LLM ベースの後編集手法を提案した [10]．この研究では Multidimensional Quality Metric (MQM) [11] に基づくエラーアノテーションを活用し，LLM に外部フィードバックを与えることで翻訳品質を改善した．中英，英独，英露のデータセットを用いた実験において，TER，BLEU，COMET の各指標で性能向上が報告されている．

3. 誤訳検出

3.1. mBERT を用いた誤訳検出

本研究では，多言語事前学習済み言語モデルである mBERT を用いて，トークンレベルでの翻訳品質評価を行う．具体的には，翻訳文中の誤りを検出するために mBERT の学習を行い，各トークンに誤りラベルを付与する．学習データの構築にあたっては，Deguchi らのデータ拡張手法 [12] に基づき，NTCIR-7 [13] および NTCIR-8 [14] の対訳特許文に対して以下の操作を施し，人工的に誤りを付与したデータを生成した．

・削除: トークンを 5%の確率で削除
・挿入: トークンを 10%の確率で挿入
・置換: トークンを 30%の確率で置換

挿入および置換においては，mBERT を用いた MASK 補完手法を採用した．具体的には，対象位置に MASK トークンを配置し，mBERT によって適合するトークンを予測させる．その際，元のトークンとの類似性を低減するため，Transformers の Pipeline が出力する候補のうち，スコアが最も低いトークンを選択した．生成された擬似データにおいて，誤りに相当するトークンには BAD タグを，その他のトークンには OK タグを付与し，教師データを構築した．

最終的に，この手法により作成した特許領域の教師データ 8,000 文を用いて mBERT の学習を行った．

3.2. LLM を用いた誤訳検出

LLM を用いた誤訳検出手法として，Kocmi らが提案した GPT ベースの評価手法である GEMBA-MQM [15] を採用する．この手法は GPT-4 を基盤とし，言語に依存しない固定の 3-shot プロンプトを用いて翻訳エラーの範囲および種類を出力するものである．GEMBA-MQM に基づき，本研究では以下の 2 種類の設定で誤り検出を実施した．

・0-shot: 事前の例示なしで翻訳誤り検出を行う．
・3-shot: 言語に依存しない 3 つの具体例を提供し，翻訳誤り検出を行う．

特に 3-shot 設定は，GPT-4 を用いた誤り検出手法の中で最も高い精度を達成するものとして報告されている [15]．

なお，mBERT および LLM を用いた誤訳検出は，後続の翻訳訂正における前処理として機能し，検出結果を活用することで翻訳訂正の精度向上を目指す．

4. LLM を用いた誤訳訂正

原文と翻訳文を入力とすることで，LLM は翻訳文中の誤りを分析し，訂正文を生成する．具体的には，LLM は検出された誤訳部分を分析し，誤り箇所およびその内容に関する説明を提示した上で，これに基づき訂正文を生成する．この際，誤りに対する説明を明示することで，訂正結果の透明性を高めるとともに，翻訳改善の過程を明確化することが可能となる．

さらに本研究では，前章で述べた誤訳検出結果を基盤として，LLM による翻訳訂正手法を提案する．特に，LLM や mBERT による誤訳検出結果を入力として活用することで，訂正精度の一層の向上を図る．

表 1 人工誤りを対象とした誤り検出評価

(a) 日英翻訳

モデル	項目	Precision	Recall	F1
GPT-4o(0shot)	OK	0.843	0.0774	0.142
	BAD	0.389	0.976	0.556
	TOTAL	F1: 0.298, MCC: 0.111		
GPT-4o(3shot)	OK	0.755	0.268	0.395
	BAD	0.409	0.853	0.553
	TOTAL	F1: 0.454, MCC: 0.141		
mBERT	OK	0.831	0.895	**0.862**
	BAD	0.797	0.696	**0.743**
	TOTAL	F1: **0.817**, MCC: **0.609**		

(b) 英日翻訳

モデル	項目	Precision	Recall	F1
GPT-4o(0shot)	OK	0.804	0.287	0.423
	BAD	0.415	0.879	0.563
	TOTAL	F1: 0.474, MCC: 0.190		
GPT-4o(3shot)	OK	0.709	0.503	0.588
	BAD	0.419	0.634	0.504
	TOTAL	F1: 0.558, MCC: 0.132		
mBERT	OK	0.824	0.925	**0.872**
	BAD	0.831	0.651	**0.730**
	TOTAL	F1: **0.821**, MCC: **0.615**		

5. 評価

5.1. データセット

本論文では，以下の 3 種類のデータに対して評価を行う．

・人工誤り特許データ

・繰り返し誤り特許請求項

・訳抜け誤り特許請求項

人工誤り特許データは，3.1 節で述べた手法に基づき，NTCIR-7 および NTCIR-8 の特許対訳データに人工的に誤りを付与して作成したものである．本研究では，この人工誤り特許データ 200 文を用いて，誤訳に対する検出・訂正能力を評価する．

さらに，繰り返し誤りおよび訳抜け誤りに対する翻訳訂正精度を評価するため，特許請求項を Transformer により日英翻訳したデータから，以下の基準に従い文を抽出した．

・繰り返し誤り特許請求項: Transformer による翻訳文の文長が参照訳文の 2 倍以上である文

・訳抜け誤り特許請求項: Transformer による翻訳文の文長が参照訳文の 0.5 倍以下である文

抽出対象は 2021 年の特許請求項とし，これを用いて訂正精度を評価する．また，対訳データの品質を高めるため，LaBSE [16] による埋め込みを用いて原文と参照文の類似度を計算し，類似度が 0.8 以上 0.98 以下の文を抽出した．最終的に，繰り返し誤り特許データ 211 文および訳抜け誤り特許データ 200 文を評価に用いた．

5.2. 評価手順

5.2.1. 翻訳誤り検出

誤り検出の評価においては，特許文に人工的に誤りを付与した 200 文を対象とし，翻訳誤り箇所に BAD タグを付与したトークン単位の評価を行った．mBERT および LLM に対しては，図 1 に示す形式と同様のタグアノテーションを付与した．

・LLM(GPT-4o) : 0-shot

・LLM(GPT-4o) : 3-shot

・mBERT

評価指標としては，F1 値および Matthews 相関係数 (MCC) を用いた．

5.2.2. 翻訳誤り訂正

誤り訂正の評価では，人工的に誤りを付与した特許文，繰り返し誤りを含む特許請求項，および訳抜け誤りを含む請求項を対象として，LLM による誤り訂正を実施した．訂正には，原文および翻訳文に加えて，

mBERT および LLM による誤り検出結果を入力として与えた．モデルの組み合わせは以下の通りである．

・LLM(GPT-4o) : 検出(誤り説明) + 修正

・LLM (GPT-4o) : 検出(GEMBA:0-shot)

→ LLM (GPT-4o) : 検出(誤り説明) + 修正

・LLM (GPT-4o) : 検出(GEMBA:3-shot)

→ LLM (GPT-4o) : 検出(誤り説明) + 修正

・mBERT : 検出(タグ)

→ LLM (GPT-4o) : 検出(誤り説明) + 修正

翻訳訂正後の文は BLEU [5] および COMET [6] により評価した．BLEU は sacreBLEU [17] を用いて算出し，COMET は wmt22-comet-da モデルを使用した．

5.3. 評価結果

5.3.1. 翻訳誤り検出評価

人工誤りを付与した特許データに対する誤り検出の結果を表 1 に示す．表から，mBERT による翻訳誤り検出が F1 値および MCC の両指標において最も優れた性能を示したことが確認できる．特に，Precision と Recall のバランスの良さが総合性能の高さに寄与していると考えられる．一方，GPT-4o はプロンプト設定によって結果に差が見られ，3-shot 設定では 0-shot 設定より改善が確認されたものの，mBERT には及ばなかった．さらに，GPT-4o の出力を詳細に分析した結果，多くのトークンに BAD タグを過剰に付与する傾向が認められた．このため BAD タグの Recall は高い値を示したが，OK タグの Recall が大幅に低下する傾向が見られた．

これらの結果は，特許文で学習を行った mBERT が特許文に対するトークンレベルでの誤り検出とタグ付けを効果的に処理できることに起因すると推測される．また，作成した擬似データを用いた mBERT の学習手法が有効であることを示唆している．一方，GPT-4o については検出性能の低さやプロンプト設定に依存する性能のばらつきが課題として残り，より高精度な誤訳検出を実現するためには追加的な工夫が必要である．

5.3.2. 翻訳誤り訂正評価

表 2 に示す結果から，mBERT ベースの誤り検出と LLM を組み合わせた提案手法が，BLEU および COMET の両指標において他手法を上回り，最も高い翻訳訂正精度を達成したことが確認された．特に，日英翻訳における人工誤り特許文に対して，他手法と比較して BLEU が統計的に有意に向上した．さらに，表 3 に示す翻訳修正例からも，提案手法が人工誤り文の誤訳を適切に訂正できていることが確認できる．これらの結果は，mBERT によるトークンレベルの誤訳検出が高精度に機能し，その情報を活用することで LLM が適切に翻訳訂正を行えることを示している．

表 2: 翻訳訂正評価 (BLEU/COMET)

	手法	人工誤り (日英)	人工誤り (英日)	繰り返し誤り (日英)	訳抜け誤り (日英)
1	修正なし	31.81/70.58	31.65/75.73	21.33/69.59	19.01/71.52
2	LLM - 検出 (誤り説明) + 修正	47.14/83.87	41.29/90.18	26.34/76.78	65.03/85.37
3	LLM - 検出 (GEMBA:0-shot)⇒ LLM - 検出 (誤り説明) + 修正	43.44/83.29	38.16/89.91	25.37/76.79	63.27/86.67
4	LLM - 検出 (GEMBA:3-shot)⇒ LLM - 検出 (誤り説明) + 修正	47.32/83.97	39.4/90.03	25.8/76.78	**66.08**/88.52
5	mBERT - 検出 (タグ) ⇒ LLM - 検出 (誤り説明) + 修正	**49.37**/84.08	**41.58**/90.11	**27.73**/76.65	56.65/83.71

表 3: 提案手法による人工誤りの訂正例

原文
ステップＳ１１において、プライマリプーリ１１への入力トルクを計算する。
参照訳文
In a step S11 , an input torque to the primary pulley 11 is calculated .
人工誤り文
In a processing stepd , The input torque to be primary pulley 11 is achieved :
提案手法
In step S11, the input torque to primary pulley 11 is calculated.

繰り返し誤り特許請求項においても，提案手法は最も良好な性能を示し，BLEU が統計的に有意に向上した．表 4 の翻訳修正例から，提案手法が繰り返し誤りを適切に訂正できていることが確認された．これらの結果から，繰り返し誤りに対しても提案手法が有効であることが示唆される．

一方，訳抜け誤り特許文においては，最も高い BLEU スコアを示したのは LLM 単独で検出および訂正を行う手法であった．表 2 に示すように，訂正なしの BLEU スコアが 19.01 であったのに対し，提案手法では 56.65 と大幅に向上した．また，表 5 の翻訳修正例からも，提案手法が訳抜け誤りを訂正できていることが確認できる．しかしながら，精度の面では LLM 単独手法に劣る結果となった．この要因として，ターゲット側のタグ情報では訳抜け誤りを十分に表現できず，翻訳訂正に必要な情報を提供できなかった可能性が考えられる．

6. 結論

本論文では，特許文で学習した mBERT により翻訳誤り検出を行い，その結果を LLM に入力して翻訳誤り訂正を行う二段階手法を提案し，評価を行った．実験の結果，BLEU が他手法と比較して統計的に有意に向上し，誤訳や繰り返し誤りを効果的に改善できることを確認した. 特に, mBERT による高精度な誤訳検出が LLM の訂正を支え，誤りトークンの適切な訂正を通じて翻訳精度の向上に寄与した．また，特許文の擬似データを用いた学習により，人手によるアノテーションが存在しない場合でも誤り検出モデルの訓練が可能であることを示した. これらの結果から, エンコーダモデルを高品質な訓練データで学習させることにより，汎用 LLM を上回る精度を達成できる可能性が示唆された．一方，訳抜け誤りの訂正に関しては，LLM が検出と訂正を一貫して行う手法が提案手法を上回り，タグ情報の活用には課題が残された．この結果は，誤

表 4: 提案手法による繰り返し誤りの訂正例

原文
a. 配電ハードウェアの構成部品として、少なくとも 1 つの受動電磁センサをインストールするステップと、b.
参照訳文
a.Installing at least one Passive Electromagnetic Sensor as a component of distribution hardware;b.
MT 文
a. installing at least one passive electromagnetic sensor as a component part of the electrical distribution hardware;b. controlling the at least one passive electromagnetic sensor to emit electromagnetic radiation;c. controlling the at least one passive electromagnetic sensor to emit electromagnetic radiation;d. controlling the at least one passive electromagnetic sensor to emit electromagnetic radiation;e. controlling the at least one passive electromagnetic sensor to emit electromagnetic radiation;f. ... controlling the at least one passive electromagnetic sensor to emit electromagnetic radiation;g
提案手法
a. installing at least one passive electromagnetic sensor as a component part of the electrical distribution hardware; b.

表 5: 提案手法による訳抜け誤りの訂正例

原文
前記 NK 細胞は、血液または細胞株に由来し、好ましくは、細胞株に由来し、より好ましくは、前記細胞株に由来する NK は NK92 細胞株であることを特徴とする 請求項 12 に記載の免疫細胞。
参照訳文
The immune cell of claim 12, wherein the NK cell is derived from blood or a cell line; preferably, from a cell line; and more preferably, the NK cell from a cell line is NK92 cell line.
MT 文
The immune cell of claim 12, wherein the NK cell is derived from blood or a cell line.
提案手法
The immune cell of claim 12, characterized in that the NK cell is derived from blood or a cell line, preferably from a cell line, and more preferably from the aforementioned cell line, specifically the NK92 cell line.

りの種類に応じた訂正手法の選択や，誤り情報の提示方法の最適化が重要であることを示唆している．

参考文献

[1] W. Xu, D. Deutsch, M. Finkelstein, J. Juraska, B. Zhang, Z. Liu, W. Y. Wang, L. Li, and M. Freitag. 2024. LLMRefine: Pinpointing and refining large language models via fine-grained actionable feedback. In Findings of NAACL, pages 1429–1445.

[2] E. Briakou, J. Luo, C. Cherry, and M. Freitag. 2024. Translating step-by-step: Decomposing the translation process for improved translation quality of long-form texts. In Proc. WMT, pages 1301–1317.

[3] J. Devlin, M. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proc. NAACL, pages 4171–4186.

[4] OpenAI. 2024. Gpt-4o system card. https://arxiv.org/abs/2410.21276

[5] K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Proc. 40th ACL, pages 311–318

[6] R. Rei, J. G. C. de Souza, D. Alves, C. Zerva, A. C. Farinha, T. Glushkova, A. Lavie, L. Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In Proc. 7th WMT, pages 578–585.

[7] Y. Wei, T. Utsuro, and M. Nagata. 2022. Extending word-level quality estimation for post-editing assistance. https://arxiv.org/abs/2209.11378.

[8] N. Guerreiro, R. Rei, D. Stigt, L. Coheur, P. Colombo, and A. Martins. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. TACL, pages 979–995.

[9] N. Goyal, J. Du, M. Ott, G. Anantharaman, and A. Conneau. 2021. Larger-scale transformers for multilingual masked language modeling. In Proc. 6th Workshop on Representation Learning for NLP, pages 29–33.

[10] D. Ki and M. Carpuat. 2024. Guiding large language models to post-edit machine translation with error annotations. In Findings of NAACL, pages 4253–4273

[11] A. Burchardt. 2013. Multidimensional quality metrics: a flexible system for assessing translation quality. In Proc. TC, pages 1–7.

[12] H. Deguchi, M. Nagata, and T. Watanabe. 2024. Detector–corrector: Edit-based automatic post editing for human post editing. In Proc. 25th EAMT, pages 191–206.

[13] A. Fujii, M. Utiyama, M. Yamamoto, and T. Utsuro. 2008. Overview of the patent translation task at the NTCIR-7 workshop. In Proc. 7th NTCIR, pages 389–400.

[14] A. Fujii, M. Utiyama, M. Yamamoto, T. Utsuro, T. Ehara, H. Echizen-ya, and S. Shimohata. 2010. Overview of the patent translation task at the NTCIR-8 workshop. In Proc. 8th NTCIR, pages 371–376.

[15] T. Kocmi and C. Federmann. 2023. GEMBA-MQM:Detecting translation quality error spans with GPT-4. In Proc. 8th WMT, pages 768–775

[16] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang. 2022. Language-agnostic BERT sentence embedding. In Proc. 60th ACL, pages 878–891.

[17] M. Post. 2018. A call for clarity in reporting BLEU scores. In Proc. 3th WMT, pages 186–191

事例・研究

Quality Estimation Reranking for Document-Level Translation

Krzysztof Mrozinski, Minji Kang, Ahmed Khota, Vincent Michael Sutanto, Giovanni Gatti De Giacomo
Yaraku, Inc.

Abstract

Quality estimation (QE) reranking is a form of quality-aware decoding which aims to improve machine translation (MT) by scoring and selecting the best candidate from a pool of generated translations. While known to be effective at the sentence level, its application to the increasingly prominent domain of document-level translation remains underexplored. In this work, we evaluate QE reranking performance on document-level (rather than the typical sentence-level) translation, using various learned and large language model (LLM)-based QE metrics. We find that with our best learned metric, **SLIDE**, BLEURT-20 scores improve by **+2.00** with only two candidates, and by **+5.09** with 32, across both decoder-only LLM models and encoder-decoder neural machine translation (NMT) models. Using the best LLM-based metric, **GEMBA-DA**, gains of **+1.63** and **+4.30** are achieved under the same conditions. Although gains shrink with longer inputs, reranking with 32 candidates yields improvements of **+2.34** (SLIDE) and **+1.40** (GEMBA-DA) on our longest documents (512-1024 source tokens). These findings demonstrate the practical value of document-level QE, with minimal runtime overhead given suitable translation models and hardware.

1. Introduction

Machine Translation (MT) evaluation metrics are widely used to assess system performance, having been shown to align strongly with human evaluation [1]. In contrast, the standard decoding strategy of maximising model likelihood (MAP) has been shown to diverge from human evaluation [2] [3]. This motivates **quality-aware decoding** [4], where MT evaluation metrics are directly integrated into the translation process.

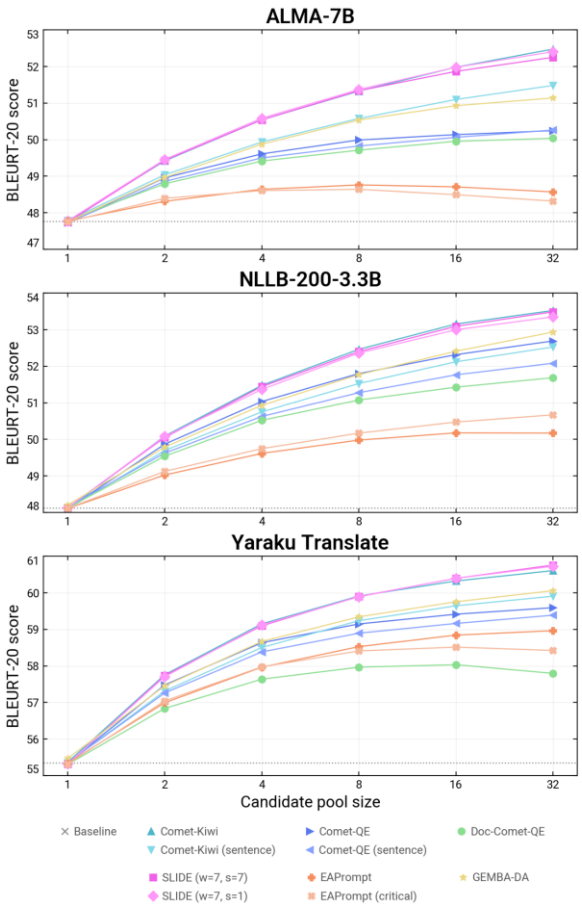


Figure 1: BLEURT-20 scores for QE reranking across different pool sizes, evaluated with all QE metrics and translation models. A pool size of 1 serves as the baseline (no reranking). Scores generally increase with larger pools under most QE metrics, for all translators.

Quality-aware decoding uses an MT evaluation metric to select an optimal translation from a candidate pool. **Minimum Bayes-Risk** (MBR) decoding [5] uses reference-based metrics such as BLEU [6] or COMET [7], comparing translation candidates against each other to select the highest utility candidate. Conversely, reference-free metrics, or

Quality Estimation Reranking for Document-Level Translation
Krzysztof Mrozinski, Minji Kang, Ahmed Khota, Vincent Michael Sutanto, Giovanni Gatti De Giacomo
Yaraku, Inc.

Quality Estimation (QE) metrics, such as Comet-Kiwi [8], may be used for **QE reranking**, where the highest-scoring candidate is chosen as the final output. Although MBR mitigates weaknesses of MAP decoding [9], it requires $O(N^2)$ pairwise comparisons compared to the linear $O(N)$ complexity of QE reranking, and translation quality gains over QE reranking are not definitive [4] [10]. Therefore, we focus our work on QE reranking.

While document-level MT is becoming increasingly prominent [11], to date, document-level QE reranking has received little attention, despite one implementation showing promising results [12]. The leading QE metrics such as Comet-Kiwi being sentence-level raises uncertainty about their suitability for document-level evaluation.

In this work, we investigate the applicability of QE metrics to document-level QE reranking. We evaluate translation quality improvements compared to standard decoding, examining differences across QE metrics, translation models, candidate pool sizes, and document lengths, as well as the associated computational trade-offs.

Our contributions are as follows:

- We demonstrate that QE reranking improves document-level MT quality across multiple QE metrics and translation models.
- We analyse how reranking effectiveness varies with candidate pool size and document length.
- We quantify the computational trade-offs of document-level QE reranking.

2. Method

2.1. Translators

For candidate generation, we evaluate both decoder-only large language models (LLMs) and encoder-decoder neural machine translation (NMT) models. While the use of NMT models is the traditional approach, the broad pretraining of LLMs typically enables them to generate more diverse outputs [10], making them well-suited for QE reranking.

We experiment with ALMA-7B [13], a LLaMA2-7B LLM finetuned for translation, and NLLB-200-3.3B [14], a widely used multilingual NMT model. Although both were trained on sentence-level data, we found them sufficient for document-level translation, given the scarcity of publicly available document-trained models. For ALMA, the LLaMA pretraining and additional monolingual fine-tuning stage may further preserve document-level capabilities. We also evaluate Yaraku Translate, our proprietary NMT model trained directly for English-Japanese document translation.

For decoding, we adopt nucleus sampling ($p = 0.9$) for ALMA [15]. For NLLB, while beam search is the typical choice for NMT models, we opt for epsilon sampling ($\varepsilon = 0.02$), shown to excel for MBR decoding [16]. Temperatures of 0.6 (ALMA) and 0.5 (NLLB) balance candidate pool diversity and quality, and have been effective for QE reranking [10]. For Yaraku Translate, we opt for diverse beam search [17] as an alternative to sampling, to address difficulties in achieving diverse yet high-quality outputs with sampling. We set $G = 16$ groups and $\lambda = 0.5$ diversity strength.

2.2. QE Metrics

2.2.1. Learned QE Metrics

We adopt the COMET model family as baseline QE metrics, namely **COMET-QE** [18] and **Comet-Kiwi** [9]. Although trained for sentence-level evaluation, we test two strategies for adapting them to the document-level.

The first averages sentence-level predictions across a document. Documents are segmented into sentences using Punkt [19] for English and a simple regular expression for Japanese, then aligned by order. When source and target sentence counts differ, the shorter text is padded by duplicating its final sentence, ensuring equal segment counts so that all sentences are scored. While this approach closely aligns with the intended use case, it is flawed in practice since document-level translators rarely preserve one-to-one sentence alignment, which compounds with longer documents. We refer to these metrics as **COMET-QE (sentence)** and **Comet-Kiwi (sentence)**.

The second strategy passes full documents as single segments. Although not the intended use-case, prior work shows this may perform comparably to metrics directly trained for longer-context evaluation [20], likely due to the long-context pretraining of the underlying InfoXLM encoders [21]. This method is constrained by sequence length limits (512 tokens for source + target for Comet-Kiwi, 512 per text for COMET-QE). We refer to these metrics simply as **COMET-QE** and **Comet-Kiwi**.

Doc-COMET-QE [22] extends COMET-QE by concatenating two preceding source and target sentences (where available) to provide additional document-level context. Sentence-level score is calculated only for the current sentence via masking, and document-level score as the average of all sentence-level scores. This is compatible with COMET-QE, which pools token representations and allows for selective masking, but not with Comet-Kiwi, whose representation collapses into a single `[CLS]` token. Although shown to improve accuracy over COMET-QE in isolated evaluations, Doc-COMET-QE inherits the same alignment problems as COMET-QE (sentence), limiting its utility for QE reranking. We use the same sentence alignment and padding approaches as outlined in COMET-QE (sentence).

SLIDE [23] is a document-level QE approach requiring no architectural changes, so we implement it on top of Comet-Kiwi. It segments documents into fixed-sentence-width, strided windows, scoring each window independently and averaging to obtain a document-level score. SLIDE is identical to Comet-Kiwi for any documents shorter than the window length but mitigates sequence length limitations for longer documents. The original work reported optimal performance with $w = 6,\ s = 6$ (window size, stride) in idealised conditions where documents segmented evenly. We instead adopt their proposed weighted partial window approach, which accommodates arbitrary document lengths. Since the best configuration is unclear, we experiment with both $w = 7,\ s = 7$ and $w = 7,\ s = 1$ as they both show good performance while representing two extremes of the method. We use the same padding approach as outlined in COMET-QE (sentence).

2.2.2. LLM-based QE Metrics

Given their strong performance in translation tasks, LLMs are a natural choice for document-level QE. We evaluate two prompting-based methods using Gemma 3 27B [24] as the backbone. Although originally developed for sentence-level QE, we expect them to transfer effectively to documents due to the long-context capabilities of LLMs in MT [25].

GEMBA-DA [26] tasks the LLM with Direct Assessment, assigning a score from 0 to 100 in a zero-shot manner. This method is efficient, requiring minimal token generation. We make minimal modifications to the original prompt to account for Gemma being instruction tuned. To mitigate the unpredictable nature of LLM output, we adopt the failure-recovery strategy from the original work, retrying with gradually higher temperature for up to five attempts, after which the candidate is discarded. This introduces a small chance of no valid candidates kept, so a fallback QE metric may be important.

EAPrompt [27] emulates the MQM human evaluation framework [28]. To compute the score, the LLM identifies major and minor errors, from which a weighted sum is computed via a regular expression. We weight major errors eight times higher than minor errors, shown to be effective for segment-level evaluation. We adopt one-shot prompting with language-pair-specific in-context examples, with minor prompt adjustments to accommodate the output style of Gemma. Both EAPrompt and GEMBA-DA frequently produce tied scores, which we resolve via random selection.

Although EAPrompt has shown state-of-the-art (SOTA) performance, our experiments revealed some limitations. First, unlike GEMBA-DA, it lacks a failure-recovery mechanism: erroneous outputs with no listed errors are indistinguishable from valid assessments of perfect translations. Second, the scoring scheme is overly lenient on critical translation errors—e.g., a nonsensical translation may only receive one major error, while a flawed yet

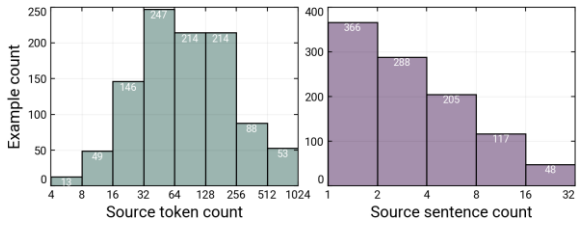


Figure 2: Distribution of source token and source sentence counts across our WMT23 dataset. Average example source text is 4.30 sentences and 138 tokens long.

comprehensible translation has more scope for identifying several errors. This issue is likely amplified at the document level, where long contexts increase the risk of critical errors. To address this, we introduce **EAPrompt-Critical**, which adds a critical error category weighted at 100.

3. Experiments

For our main experiment, we generate a large pool of candidate translations for each source text in our dataset, score them with each QE metric, and then trim the candidate pool to various sizes. From each pool, we select the top-scoring candidate and evaluate it against the reference translation using several reference-based metrics.

3.1. Dataset

We use the WMT23 test set [29] as our source of ground-truth translations, evaluating bidirectionally between English and Japanese. As the dataset is segmented at the document, paragraph, and sentence levels, we merge segments to obtain document-level translations. We augment the data with a balanced mix of full documents and individual paragraphs to better examine the relationship between document length and performance. As the dataset was released after the COMET models we evaluate, there is no risk of overfitting, although some risk remains for Gemma. Dataset length distributions are shown in Figure 2.

3.2. Evaluation Metrics

We evaluate QE reranking performance using reference-based metrics, treating the reranked translation as the hypothesis and the dataset translation as the reference. We consider two families: **neural metrics**, using **BLEURT-20** [30] and **COMET-22** [31], the latter setting the SOTA in WMT 2022 shared task [32]; and **LLM-based evaluation**, using the reference-based prompting framework **GEMBA-DA** [26], which we had found to outperform EAPrompt. For the backbone LLM, we use GPT-4.1-mini [33] for its strong natural language understanding capabilities and to minimise the risk of overfitting with our QE metrics using Gemma. Notably, this is the only evaluation metric directly compatible with document-level translation, as both COMET and BLEURT impose a strict 512 token cap, limiting their reliability on long documents.

We acknowledge the risk of overfitting when using the same metric family for both QE and evaluation, which can lead to evaluation scores diverging from human judgement [4]. Nonetheless, we include COMET-22 as an evaluation metric to enable comparison across a broad range of evaluators and discuss the implications of overfitting in Section 4.

4. Results

4.1. Pool Size

We first examine the effect of candidate pool size on reranking performance. A pool size of one serves as the baseline, equivalent to no QE reranking. As shown in Figure 1, scores generally increase with larger pools, confirming the effectiveness of QE reranking at the document level. Gains are observed for both LLMs and NMT models, with the largest improvements in Yaraku Translate, likely reflecting its document-level training. Performance does not reach a full plateau at pool size 32, suggesting larger pools could yield further gains. While improvements are consistent across all evaluators, the leading QE metric varies, likely due to evaluator-specific biases such as overfitting (e.g., COMET-based QE metrics evaluated with COMET-22), and

QE Metric	Baseline	2	4	8	16	32
ALMA-7B (GEMBA-DA / BLEURT-20 / COMET-22)						
Comet-Kiwi	47.39 / 47.77 / 72.62	51.24 / 49.42 / 74.81	53.77 / 50.54 / 76.04	55.46 / 51.33 / 76.80	56.91 / 51.98 / 77.35	58.15 / 52.48 / 77.71
Comet-QE		49.88 / 48.95 / 74.79	51.32 / 49.61 / 75.91	52.06 / 49.99 / 76.52	52.21 / 50.14 / 76.83	52.47 / 50.25 / 77.05
Comet-Kiwi (sentence)		50.18 / 49.04 / 74.16	52.12 / 49.94 / 75.04	53.57 / 50.58 / 75.64	54.68 / 51.10 / 76.07	55.57 / 51.48 / 76.37
Comet-QE (sentence)		49.39 / 48.86 / 74.45	50.62 / 49.50 / 75.32	51.18 / 49.83 / 75.73	51.31 / 50.06 / 75.85	51.22 / 50.26 / 75.85
Doc-Comet-QE		49.26 / 48.79 / 74.34	50.50 / 49.41 / 75.28	50.92 / 49.71 / 75.74	51.16 / 49.95 / 75.98	51.10 / 50.04 / 76.05
SLIDE (w=7, s=7)		51.24 / 49.44 / 74.74	53.66 / 50.54 / 75.95	55.50 / 51.34 / 76.75	56.77 / 51.87 / 77.29	58.17 / 52.25 / 77.74
SLIDE (w=7, s=1)		51.23 / 49.46 / 74.76	53.84 / 50.59 / 76.03	55.61 / 51.37 / 76.82	56.99 / 51.98 / 77.41	58.32 / 52.40 / 77.81
GEMBA-DA		51.95 / 48.97 / 74.14	55.44 / 49.88 / 75.25	57.96 / 50.54 / 75.87	59.64 / 50.93 / 76.30	60.81 / 51.14 / 76.54
EAPrompt		49.18 / 48.31 / 72.88	50.40 / 48.64 / 72.76	51.15 / 48.76 / 72.40	51.42 / 48.71 / 71.88	51.53 / 48.57 / 71.30
EAPrompt (critical)		49.53 / 48.40 / 73.29	50.80 / 48.61 / 73.29	51.39 / 48.64 / 73.00	51.58 / 48.49 / 72.46	51.54 / 48.32 / 71.88
NLLB-200-3.3B (GEMBA-DA / BLEURT-20 / COMET-22)						
Comet-Kiwi	50.92 / 48.09 / 73.72	55.88 / 50.09 / 75.94	58.94 / 51.48 / 77.26	61.12 / 52.46 / 78.14	62.66 / 53.16 / 78.78	63.75 / 53.53 / 79.25
Comet-QE		54.47 / 49.87 / 75.93	56.53 / 51.03 / 77.19	57.67 / 51.80 / 78.03	58.31 / 52.32 / 78.63	58.65 / 52.69 / 79.06
Comet-Kiwi (sentence)		54.64 / 49.70 / 75.35	57.08 / 50.75 / 76.32	58.86 / 51.53 / 76.97	60.22 / 52.12 / 77.49	61.36 / 52.53 / 77.91
Comet-QE (sentence)		53.98 / 49.62 / 75.52	55.97 / 50.63 / 76.55	57.12 / 51.28 / 77.24	58.00 / 51.77 / 77.76	58.58 / 52.08 / 78.07
Doc-Comet-QE		53.70 / 49.54 / 75.40	55.25 / 50.52 / 76.27	55.99 / 51.08 / 76.79	56.26 / 51.43 / 77.07	56.13 / 51.69 / 77.22
SLIDE (w=7, s=7)		55.63 / 50.05 / 75.84	58.68 / 51.44 / 77.10	60.79 / 52.40 / 77.90	62.26 / 53.09 / 78.51	63.27 / 53.49 / 78.91
SLIDE (w=7, s=1)		55.73 / 50.08 / 75.90	58.70 / 51.37 / 77.15	60.82 / 52.36 / 78.01	62.26 / 53.00 / 78.65	63.17 / 53.35 / 79.06
GEMBA-DA		56.07 / 49.79 / 75.49	59.57 / 50.93 / 76.49	62.20 / 51.77 / 77.21	64.04 / 52.41 / 77.71	65.55 / 52.94 / 78.10
EAPrompt		53.18 / 49.03 / 74.20	54.56 / 49.62 / 74.33	55.25 / 49.98 / 74.08	55.57 / 50.18 / 73.65	55.42 / 50.17 / 72.97
EAPrompt (critical)		53.51 / 49.12 / 74.56	55.13 / 49.75 / 74.90	56.04 / 50.17 / 74.99	56.65 / 50.47 / 74.95	56.92 / 50.67 / 74.80
Yaraku Translate (GEMBA-DA / BLEURT-20 / COMET-22)						
Comet-Kiwi	70.21 / 55.37 / 78.61	76.21 / 57.76 / 80.66	79.33 / 59.15 / 81.80	80.81 / 59.91 / 82.45	81.49 / 60.32 / 82.84	81.81 / 60.61 / 83.14
Comet-QE		74.92 / 57.48 / 80.63	77.01 / 58.64 / 81.69	77.66 / 59.15 / 82.25	77.85 / 59.42 / 82.59	78.18 / 59.60 / 82.83
Comet-Kiwi (sentence)		75.20 / 57.31 / 80.26	78.05 / 58.50 / 81.23	79.56 / 59.24 / 81.81	80.36 / 59.65 / 82.20	80.81 / 59.91 / 82.49
Comet-QE (sentence)		74.69 / 57.26 / 80.31	76.97 / 58.38 / 81.28	77.89 / 58.90 / 81.78	78.23 / 59.16 / 82.10	78.42 / 59.39 / 82.36
Doc-Comet-QE		73.43 / 56.84 / 79.99	74.94 / 57.64 / 80.74	75.37 / 57.97 / 81.11	75.33 / 58.03 / 81.26	74.50 / 57.80 / 81.25
SLIDE (w=7, s=7)		76.21 / 57.74 / 80.64	79.34 / 59.10 / 81.73	80.91 / 59.90 / 82.39	81.77 / 60.40 / 82.85	82.27 / 60.76 / 83.18
SLIDE (w=7, s=1)		76.23 / 57.71 / 80.64	79.41 / 59.12 / 81.77	80.89 / 59.89 / 82.41	81.77 / 60.40 / 82.87	82.28 / 60.73 / 83.18
GEMBA-DA		76.37 / 57.46 / 80.37	79.60 / 58.67 / 81.33	81.12 / 59.34 / 81.83	82.07 / 59.76 / 82.15	82.64 / 60.06 / 82.36
EAPrompt		74.30 / 56.99 / 79.86	76.62 / 57.97 / 80.57	77.79 / 58.53 / 80.87	78.35 / 58.84 / 81.00	78.41 / 58.97 / 80.93
EAPrompt (critical)		74.66 / 57.04 / 79.98	77.07 / 57.98 / 80.68	78.24 / 58.41 / 80.99	78.64 / 58.52 / 81.08	78.68 / 58.42 / 81.00

Table 1: QE reranking performance across all pool sizes reported as GEMBA-DA / BLEURT-20 / COMET-22 scores for each QE metric and translator. Best scores for each evaluator and pool size are highlighted and denoted in bold.

sequence length constraints in COMET-22 and BLEURT, which limit their ability to fully capture the advantages of LLM-based QE metrics which can handle longer sequences. Full results are given in Table 1.

Among COMET-based metrics, Comet-Kiwi consistently outperforms COMET-QE. Notably, scoring entire documents in one pass also outperforms per-sentence averaging in all settings. The gap is smallest, however, for Yaraku Translate, which enforces sentence alignment, suggesting the benefit of aligned outputs. In contrast, Doc-COMET-QE provides no benefit and ranks among the weakest QE metrics. Both configurations of SLIDE performed very similarly to each other and to standard Comet-Kiwi, only providing marginal gains in some cases. This is expected, as our SLIDE implementation is built upon Comet-Kiwi.

The LLM-based metrics show mixed results. When evaluated with BLEURT and COMET, GEMBA-DA performs well but slightly below the best-performing COMET-based metrics. This is unsurprising due to the potential overfitting risks of the COMET-based QE metrics. However, with GPT as the evaluator, GEMBA-DA achieves the best performance. Naturally, there is also an overfitting risk in this case, but the difference in backbone LLM attempts to minimise this. Surprisingly, while adding the critical category helps, EAPrompt generally showed poor

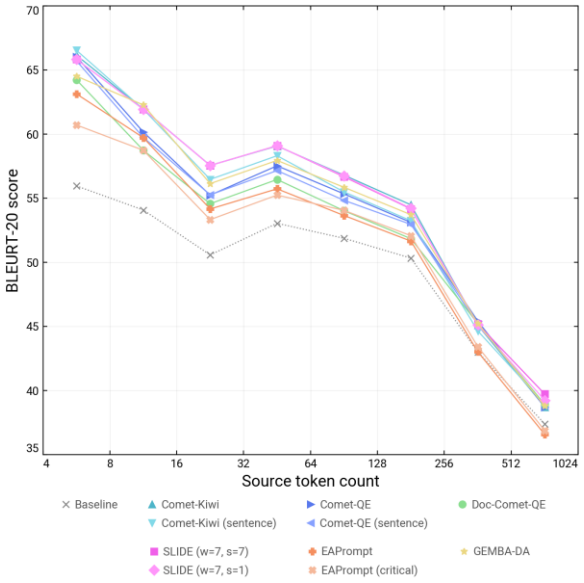


Figure 3: QE reranking performance for all QE metrics at pool size 32, averaged across all translator models. Gains diminish with longer documents but remain above the baseline (pool size 1) for most metrics.

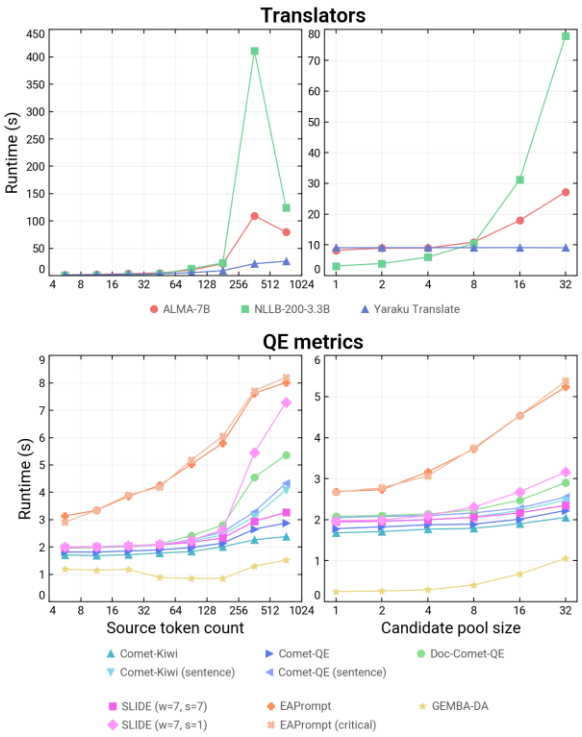


Figure 4: Runtime by source length and pool size for all QE metrics and translators. Translation runtime rises steeply for models not trained at the document level, while QE runtime remains a small fraction of the overall runtime.

performance.

4.2. Length

We expected QE performance to degrade with longer inputs: learned metrics were not trained for long sequences, and LLM attention tends to diverge over extended contexts [34]. The results shown in Figure 3 confirm this hypothesis. QE metrics perform best on short inputs, but performance remains stable up to ~256 source tokens, indicating reasonable capability in multi-sentence contexts. Beyond this point, performance rapidly declines, reflecting the 512 token limit for combined source and target texts in most QE metrics and evaluators. Nonetheless, most QE metrics continue to provide a performance gain over the baseline even at the longest tested sequence lengths.

The extent of degradation varies across metrics. SLIDE exhibits similar performance to Comet-Kiwi for short sequences but retains slightly higher performance for long sequences, thanks to the sliding window approach avoiding the token limit. Additionally, when GPT is used as an evaluator, the lead of GEMBA-DA as a QE metric becomes the biggest for long sequences, highlighting the long-context capabilities of LLMs.

4.3. Runtime

Runtime is difficult to assess in a hardware-agnostic manner, as it depends heavily on GPU memory, implementation details, batch size, and document length. In our experiments, we used a cluster of 4 NVIDIA A6000 GPUs, sharding models so that the maximum pool size could be processed in a single batch. To ensure fairness, we fixed the batch size for both translation and QE models at 32 (the largest pool size). Although many QE models could support larger batch sizes, this cap highlights the potential slowdown

of metrics requiring multiple evaluations per candidate. We show the translator and QE metric runtime in Figure 4.

Among translators, Yaraku Translate shows little sensitivity to pool size, likely reflecting the efficiency of its dynamic beam search decoding. Counterintuitively, the smaller NLLB model is slower than the larger ALMA, exploding exponentially with both pool size and input length. This is primarily due to difficulty generating stop tokens; hallucinated outputs often reach the token limit, delaying the entire batch. This highlights the value of document-level translation models and effective stopping strategies. To partially mitigate this issue, we apply an adaptive maximum token limit defined as:

$$\min\left(N_{\text{ceil}}, \left\lfloor L_{\text{in}} \cdot \alpha_{\text{m}} \cdot \frac{\mu_{\text{tgt}}}{\mu_{\text{src}}} + \alpha_{\text{a}} \right\rfloor\right)$$

where L_{in} is the input token length, $\alpha_{\text{a}} = 10$, $\alpha_{\text{m}} = 2$ are additive and multiplicative margin factors, $N_{\text{ceil}} = 2048$ is a hard ceiling, and μ_{tgt}, μ_{src} denote the average dataset token lengths for the current target and source languages, respectively. This caps hallucinated translations while retaining sufficient headroom for legitimate document-length variability.

For learned QE metrics, runtime grows modestly with pool size and sequence length. Methods requiring multiple evaluations per candidate, namely SLIDE, Doc-COMET-QE, and the sentence-based COMET variants, exhibit steeper growth on longer documents. Between SLIDE configurations, while $s = 1$ and $s = 7$ exhibit similar performance, $s = 1$ incurs significantly higher runtime, making $s = 7$ the more practical choice. Comparison between learned QE metrics and LLM-based methods is complicated by experiment setup limitations (Gemma was hosted on a GH200 NVIDIA GPU), yet GEMBA-DA runs substantially faster than both EAPrompt variants, as it requires minimal token generation. GEMBA-DA, however, exhibits higher runtime growth with larger pool sizes, likely reflecting the higher probability of triggering its failure-prevention strategy.

5. Conclusion

We investigated the applicability of QE reranking to the document translation domain and found consistent translation quality gains over standard decoding across various QE metrics and translation models. Gains increased with translation candidate pool size and were not saturated at 32, indicating further potential improvement. Methods that score full documents in one pass consistently outperform sentence-level averaging, even with QE metrics designed for sentence-level scoring. SLIDE was found to be the best performing QE metric, matching Comet-Kiwi on short inputs while being more performant on long documents. Among LLM-based methods, GEMBA-DA was found to be competitive when evaluated with COMET-22 and BLEURT, and leads under GPT evaluation. Although performance gains decline for long documents beyond 256 source tokens, QE reranking improves translation quality even for the longest documents in our dataset. Runtime analysis showed that all QE metrics represent a fraction of total translation runtime cost, allowing for near cost-free performance gains under certain conditions.

Limitations

This study has several limitations. First, although we tested multiple pool sizes, performance did not reach a clear saturation point (i.e., a peak followed by stagnation), which would have provided stronger evidence of the limits of QE reranking. Second, resource constraints prevented us from exploring more diverse LLM prompting methods for evaluation, limiting our ability to fully exploit the potential of LLMs. Third, most QE models remain constrained by a 512-token limit, restricting their applicability for longer documents. Lastly, this study lacks human evaluation, which reduces the reliability of our tested evaluation metrics, and would have allowed us to better explore the extent of overfitting.

References

[1] Q. Ma, J. Wei, O. Bojar, and Y. Graham, ‘Results of the WMT19 Metrics Shared Task: Segment-Level and Strong MT Systems Pose Big Challenges’, in Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1), 2019, pp. 62–90.

[2] B. Eikema and W. Aziz, ‘Is MAP Decoding All You Need? The Inadequacy of the Mode in Neural Machine Translation’, in Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 4506–4520.

[3] M. Freitag, D. Grangier, Q. Tan, and B. Liang, ‘High Quality Rather than High Model Probability: Minimum Bayes Risk Decoding with Neural Metrics’, Transactions of the Association for Computational Linguistics, vol. 10, pp. 811–825, 2022.

[4] P. Fernandes et al., ‘Quality-Aware Decoding for Neural Machine Translation’, in Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2022, pp. 1396–1412.

[5] S. Kumar and W. Byrne, ‘Minimum Bayes-Risk Decoding for Statistical Machine Translation’, in Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004, pp. 169−176.

[6] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, ‘Bleu: a Method for Automatic Evaluation of Machine Translation’, in Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002, pp. 311–318.

[7] R. Rei, C. Stewart, A. C. Farinha, and A. Lavie, ‘COMET: A Neural Framework for MT Evaluation’, in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 2685–2702.

[8] R. Rei et al., ‘CometKiwi: IST-Unbabel 2022 Submission for the Quality Estimation Shared Task’, in Proceedings of the Seventh Conference on Machine Translation (WMT), 2022, pp. 634–645.

[9] M. Müller and R. Sennrich, ‘Understanding the Properties of Minimum Bayes Risk Decoding in Neural Machine 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), 2021, pp. 259–272.

[10] G. Vernikos and A. Popescu-Belis, ‘Don’t Rank, Combine! Combining Machine Translation Hypotheses Using Quality Estimation’, in Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024, pp. 12087–12105.

[11] L. Wang et al., ‘Benchmarking and Improving Long-Text Translation with Large Language Models’, in Findings of the Association for Computational Linguistics: ACL 2024, 2024, pp. 7175–7187.

[12] K. Kudo et al., ‘Document-level Translation with LLM Reranking: Team-J at WMT 2024 General Translation Task’, in Proceedings of the Ninth Conference on Machine Translation, 2024, pp. 210–226.

[13] H. Xu, Y. J. Kim, A. Sharaf, and H. H. Awadalla, ‘A Paradigm Shift in Machine Translation: Boosting Translation Performance of Large Language Models’, in The Twelfth International Conference on Learning Representations, 2024.

[14] N. Team et al., ‘No Language Left Behind: Scaling Human-Centered Machine Translation’, arXiv [cs.CL]. 2022.

[15] H. Touvron et al., 'Llama 2: Open Foundation and Fine-Tuned Chat Models', arXiv [cs.CL]. 2023.

[16] M. Freitag, B. Ghorbani, and P. Fernandes, 'Epsilon Sampling Rocks: Investigating Sampling Strategies for Minimum Bayes Risk Decoding for Machine Translation', in Findings of the Association for Computational Linguistics: EMNLP 2023, 2023, pp. 9198–9209.

[17] A. K. Vijayakumar et al., 'Diverse Beam Search: Decoding Diverse Solutions from Neural Sequence Models'. 2017.

[18] R. Rei, C. Stewart, A. C. Farinha, and A. Lavie, 'Unbabel's Participation in the WMT20 Metrics Shared Task', in Proceedings of the Fifth Conference on Machine Translation, 2020, pp. 911–920.

[19] T. Kiss and J. Strunk, 'Unsupervised Multilingual Sentence Boundary Detection', Computational Linguistics, vol. 32, no. 4, pp. 485–525, 2006.

[20] D. Deutsch, J. Juraska, M. Finkelstein, and M. Freitag, 'Training and Meta-Evaluating Machine Translation Evaluation Metrics at the Paragraph Level', in Proceedings of the Eighth Conference on Machine Translation, 2023, pp. 996–1013.

[21] Z. Chi et al., 'InfoXLM: An Information-Theoretic Framework for Cross-Lingual Language Model Pre-Training', in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, pp. 3576–3588.

[22] G. Vernikos, B. Thompson, P. Mathur, and M. Federico, 'Embarrassingly Easy Document-Level MT Metrics: How to Convert Any Pretrained Metric into a Document-Level Metric', in Proceedings of the Seventh Conference on Machine Translation (WMT), 2022, pp. 118–128.

[23] V. Raunak, T. Kocmi, and M. Post, 'SLIDE: Reference-free Evaluation for Machine Translation using a Sliding Document Window', in Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), 2024, pp. 205–211.

[24] G. Team et al., 'Gemma 3 Technical Report', arXiv [cs.CL]. 2025.

[25] M. Karpinska and M. Iyyer, 'Large Language Models Effectively Leverage Document-level Context for Literary Translation, but Critical Errors Persist', in Proceedings of the Eighth Conference on Machine Translation, 2023, pp. 419–451.

[26] T. Kocmi and C. Federmann, 'Large Language Models Are State-of-the-Art Evaluators of Translation Quality', in Proceedings of the 24th Annual Conference of the European Association for Machine Translation, 2023, pp. 193–203.

[27] Q. Lu, B. Qiu, L. Ding, K. Zhang, T. Kocmi, and D. Tao, 'Error Analysis Prompting Enables Human-Like Translation Evaluation in Large Language Models', in Findings of the Association for Computational Linguistics: ACL 2024, 2024, pp. 8801–8816.

[28] M. Freitag, G. Foster, D. Grangier, V. Ratnakar, Q. Tan, and W. 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.

[29] M. Freitag et al., 'Results of WMT23 Metrics Shared Task: Metrics Might Be Guilty but References Are Not Innocent', in Proceedings of the Eighth Conference on Machine Translation, 2023, pp. 578–628.

[30] T. Sellam, D. Das, and A. Parikh, 'BLEURT: Learning Robust Metrics for Text Generation', in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 7881–7892.

[31] R. Rei et al., 'COMET-22: Unbabel-IST 2022 Submission for the Metrics Shared Task', in Proceedings of

the Seventh Conference on Machine Translation (WMT), 2022, pp. 578–585.

[32] M. Freitag et al., ‘Results of WMT22 Metrics Shared Task: Stop Using BLEU -- Neural Metrics Are Better and More Robust’, in Proceedings of the Seventh Conference on Machine Translation (WMT), 2022, pp. 46–68.

[33] OpenAI et al., ‘GPT-4 Technical Report’, arXiv [cs.CL]. 2024.

[34] F. Barbero et al., ‘Why do LLMs attend to the first token?’, in Second Conference on Language Modeling, 2025.

イベント報告

MTSummit2025 参加報告

田中英輝
情報通信研究機構

1. はじめに

2025 年 6 月 23 日から 27 日までスイス、ジュネーブ大学の Uni Mail キャンパスで MTSummit2025 が開催されました。MTSummit はアジア太平洋機械翻訳協会、ヨーロッパ機械翻訳協会、アメリカ機械翻訳協会が持ち回りで 2 年に 1 度開催しており、今回はヨーロッパ機械翻訳協会が主催しました。

会議の概要は以下の通りで、機械翻訳に関する幅広い内容が報告されました。また、久しぶりの対面のみの開催で、会場は多数の参加者で賑わいました。

23 日、24 日　ワークショップとチュートリアル
25 日－27 日　本会議
参加者数　320 名
発表件数（採択率）

Research	
- Technical	23（51%）
- Translators and Users	13（81%）
Sponsored Talk	2
Implementations and Case Studies	8
Products and Projects	20

2. 会議の動向

以下、会議で気づいた動向を報告します。

手話翻訳

手話翻訳に関する研究発表は一件だけでしたが、キーノートスピーチ、ワークショップで扱われたため印象に残りました。マルチモーダル翻訳への技術的興味が高まる中、手話翻訳は実用的な意味が大きく、今後ますます注目されることが予想されます。ただし、手話利用者は少なくデータ収集は容易ではありません。技術開発に加えてデータの開発、共有が大きな課題になります。

文学翻訳

文学の翻訳には創造性が必須なため機械翻訳が手を出せない領域だと考えられてきましたが、今回、これに取り組む研究がいくつか報告されました。それぞれ扱う言語、領域が違うため一般的な結論とは言えませんが、興味深い第一歩だと思います。以下、3 編の研究を紹介します。

Optimising ChatGPT for creativity in literary translation: A case study from English into Dutch, Chinese, Catalan and Spanish, Du et al.

ChatGPT を用いた文学翻訳において創造性を最大化する方法を探るため、英語の SF 小説のオランダ語・中国語・カタルーニャ語・スペイン語への翻訳を対象に、プロンプトの工夫や温度設定など 6 つの条件を比較しています。創造性は語彙・構文・意味の忠実性の 3 観点から評価され、最も創造的な翻訳は「創造的に翻訳して」と指示した単純なプロンプトと温度 1.0 の設定で得られました。DeepL よりも創造性に優れる場合もありましたが、人間翻訳には一貫して及ばないという結果を得ています。

Extending CREAMT: Leveraging Large Language Models for Literary Translation Post-Editing, Castaldo et al.

文学翻訳を対象に大規模言語モデル（LLM）を活用したポストエディット（PE）手法の有効性を検証しています。研究では、英語の小説のイタリア語翻訳を対象に、プロの文学翻訳者と協力し、編集時間・品質・創造性の観点から LLM 翻訳の PE を評価しました。この結果、人手翻訳と同等の創造性を保ちつつ、編集時

MTSummit2025 Attendance Report
Hideki Tanaka
National Institute of Information and Communications Technology

間を大幅に短縮できることが明らかになり、LLM は文学翻訳者の生産性向上に貢献する可能性が示唆されました。

*The Challenge of Translating Culture-Specific Items: Evaluating MT and LLMs Compared to Human Translator*s, Budimir

文化特有の表現（Culture-Specific Items, CSIs）の翻訳における課題を、ChatGPT-4o、Gemini 1.5 Flash、Google Translate と人間翻訳を比較することで明らかにしています。対象言語は、フラマン語（オランダ語の方言）とセルビア語という低資源言語ペアで、3 つの小説から抽出した CSIs を「物質文化（食物名など）」「固有名詞」「社会文化（スポーツ名など）」の 3 領域に分類しています。これらを対象にシステムが 8 種類の翻訳戦略（音訳、説明、目的言語での類似語による翻訳など）のどれを適用したかを分析しています。

この結果、Gemini が人間の翻訳戦略に最も近く、Google Translate は言語モデルに比べて課題を抱えていることがわかりました。また特に固有名詞の翻訳は、どのシステムにとっても困難であることが明らかになりました。CSIs の翻訳は依然として機械翻訳にとって大きな課題であると結論しています。

WIPO における機械翻訳の実用状況

今回は大規模言語モデルの翻訳応用に注目が集まっていましたが、ニューラル機械翻訳（NMT：Neural Machine Translation）の利用が確実に進んでいることを実感できる報告がありました。

WIPO（World Intellectual Property Organization：世界知的所有権機関）はジュネーブに本部を置き、特許などの知的財産に関わる大量の文書を日々翻訳しており、今回、スポンサートークの中で機械翻訳システムの利用状況を報告しました。

WIPO では統計翻訳の時代から積極的にシステムの開発と導入を進めており、現在は Marian ベースの NMT を 17 言語で開発済みで、日々の翻訳、後編集に利用しています。機械翻訳の導入教育、データ収集、システムの学習、フィードバックなど日々実施しており、膨大な実用のノウハウを有しています。

特に興味深かったのが、日々の後編集作業をモニターするダッシュボードです。翻訳結果と後編集結果の編集距離を常時モニターすることで、システムの不具合や後編集作業の問題を有効に追跡可能だと報告していました。簡単な仕組みですが有用なノウハウだと思います。

WIPO では、LLM の導入には至っていませんが、現在、導入の検討を進めているとのことでした。

3. おわりに

会期中に次の各賞が授与されました。

- **IAMT Award of Honor**
  Prof. Mikel L.Forcada
- **Best Thesis Award（最優秀博士論文賞）**
  Ricard Costa Dias Rei
  Robust, Interpretable and Efficient MT Evaluation with Fine-tuned Metrics
- **Best Paper Award（最優秀論文賞）**
  Zhi Qu, Chenchen Ding, and Taro Watanabe
  Languages Transferred Within the Encoder: On Representation Transfer in Zero-Shot Multilingual Translation

Best Thesis Award の対象となったのは近年注目されている翻訳評価指標、COMET に関する研究でした。意味を捉えることができる評価指標ということで、最近よく使われています。受賞者は挨拶の冒頭で「COMET はすでに古くなっており、新たな手法が必要だ」と言って会場を笑わせていました。機械翻訳研究の進歩の速さを感じさせる一言でした。

次回の MTSummit2027 は、アメリカ機械翻訳協会が主催し、候補地としてカナダを検討しています。大規模言語モデルによって機械翻訳の世界がどのように進化していくのか、MTSummit2027 が楽しみです。

イベント報告

第 11 回特許・技術文書翻訳ワークショップ（PSLT2025）開催報告

後藤 功雄[1]　須藤 克仁[2]　綱川 隆司[3]
[1]愛媛大学　[2]奈良女子大学　[3]静岡大学

1. 開催概要

スイス・ジュネーブにて開催された機械翻訳サミット（Machine Translation Summit）2025 の併催ワークショップの一つとして、特許・技術文書翻訳ワークショップ（The 11th Workshop on Patent and Scientific Literature Translation; PSLT 2025）が 2025 年 6 月 24 日に開催された。本ワークショップは、アジア太平洋機械翻訳協会（AAMT）、および一般財団法人日本特許情報機構（Japio）による AAMT/Japio 特許翻訳研究会が中心となり 2005 年から隔年で開催しており今回で 11 回目を数える。ワークショップは現地開催で、現地に来られない発表者に対してはオンライン発表にも対応した形での開催となった。

本ワークショップでは日本国特許庁と WIPO からそれぞれ招待講演をいただき、2 件の一般講演があった。現地にて 20 名程度の参加があり、質疑も活発であり盛況であったと言える。

2. 招待講演

招待講演 1 件目は日本特許庁の村上遼太氏から、特許庁の特許情報を用いた情報サービスに関するご講演をいただいた。特許庁では特許をデータベースに蓄積して、特許情報プラットフォーム（J-PlatPat）を通してユーザに特許情報を提供しており、J-PlatPat では機械翻訳を利用して日本語の特許を英語で提供したり、外国語の特許を日本語で提供していることが紹介された。さらにこのサービスの最近の追加機能や、日本語―インドネシア語の 500 万文対の対訳コーパスを機械翻訳システムで追加学習する効果などについて説明があった。

招待講演 2 件目はジュネーブに本部がある世界知的所有権機関（WIPO）の Bruno Pouilquen 氏から WIPO の機械翻訳に関してご講演をいただいた。WIPO の機械翻訳の歴史、WIPO の機械翻訳をドイツ、韓国、ユーラシア特許機構などの各国の特許庁が利用していることや、機械翻訳が機械学習により構築されていること、機械翻訳が IPC ドメインの情報を利用していること、訓練データの言語対毎のデータ量、他の翻訳システムとの BLEU スコア比較評価などについて解説いただいた。また IPC の自動分類や画像の自動分類についても紹介があった。

3. 一般講演

一般講演 1 件目は Longhui Zou らによる英語から中国語への学術論文の翻訳での ChatGPT-4o と DeepSeek-V3 との比較評価についてであった。評価尺度として、参照訳を使わない自動評価（COMET-KIWI）、語彙多様性、構文の複雑さを用いた。結果は COMET-KIWI の平均スコアは DeepSeek-V3 の方が高く、語彙多様性は GPT-4o の方が高く、構文の複雑さも GPT-4o の方が高かった。LLM を学術翻訳に利用する実務者にとって、本研究の成果は、テキストの特性に基づくモデルの選定に必要であることを示唆しているとのことでした。

一般講演 2 件目は Thomas Moerman らの発表で、彼らの手法は既存の 2 つの手法の組み合わせで、組み合わせた手法は、多分野にわたるデータから関連するデータを抽出するトピックフィルタリングと、利用可能なデータをより効率的に活用するためのファジーマッチ（FM）拡張である。3 つの科学分野における英語からフランス語への翻訳実験から、トピックフィル

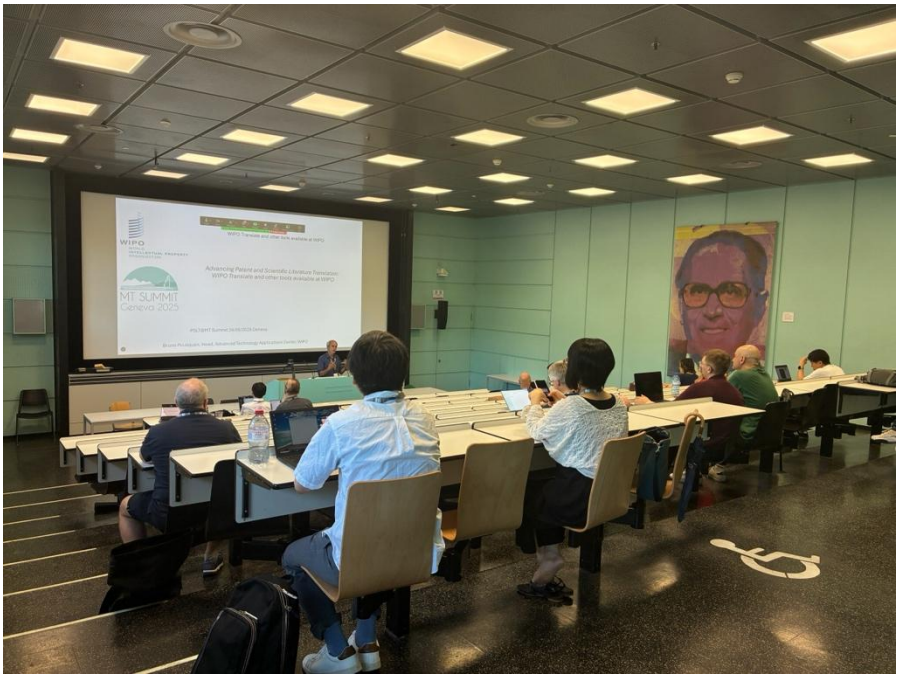
Report of the 11th Workshop on Patent and Scientific Literature Translation (PSLT 2025)
Isao Goto, Katsuhito Sudoh, Takashi Tsunakawa
Ehime University, Nara Woman's University, Shizuoka University

タリングと FM 拡張を組み合わせることにより、スクラッチから訓練したニューラル機械翻訳（NMT）モデルの性能が向上した。NMT システムは計算リソースが限られた状況下で科学文献を翻訳するための有力な選択肢となり得るが、翻訳性能と大規模言語モデル（LLM）のパラメータ数の増加に正の相関があることから、ファインチューニングなしでも大規模な LLM は、このような特化型 NMT モデルよりも優れた翻訳性能を発揮する可能性があることも示唆しているとのことでした。

4. 所感

MT Summit 2025 には 5 つの併設ワークショップがあり、PSLT 2025 は同じ時間帯に他の 2 つのワークショップと同時開催になったものの比較的多くの方にご参加いただけ、また参加者は現地参加であったことから現地会場での質疑が活発であった。プログラム最後の総合討論の時間では、今年のアジア言語の翻訳ワークショップ（WAT 2025）で新たに始める特許請求項翻訳タスクについての紹介があり、参加者と情報共有できた。

今後、大規模言語モデルが特許や技術文書について翻訳に限らず作成支援や要約など様々な用途で活用されるようになると考えられ、大規模言語モデルの活用についても幅広く技術的課題の解決に向けた議論の場として本ワークショップをご活用いただくべく、募集分野やワークショップ名の再検討を行った上で引き続き実施していきたいと考えている。



法人会員 PR

ビジネス課題に特化した専用翻訳機の可能性

織田 稔之
株式会社 IP DREAM

1. はじめに

株式会社 IP DREAM[1]は多言語 AI コミュニケーションサービス「VoiceOn」を提供しています。このサービスには、ブラウザだけで利用できるクラウド型と、専用ハードウェアを用いた端末型の 2 種類があり、いずれも、法人業務において、外国人のお客様に正確な情報を伝える翻訳ツールとして活用されています。

本稿では、接客窓口専用翻訳機「VoiceOn Information Desk」、スタンドアロン翻訳機「VoiceOn Station」を取り上げ、専用機として提供する目的、ビジネスシーンにおける機械翻訳の役割、そして専用機ならではの拡張性について、活用事例を交えながら解説します。

2. 接客窓口専用翻訳機

サービス利用方法を案内する接客窓口では、ゲストにサービス内容を正確に伝え、満足度の高い利用体験を提供することが最終的な目的です。

注目した接客現場のコミュニケーション要件	
1	**すぐに接客をスタートできるか** 端末準備、起動、言語設定、…
2	**ゲストが迷わず操作できるか** 手順理解、ボタン操作、大きな画面、…
3	**テンポよく質疑応答できるか** ホスト・ゲストの切り替え、短文の対話、…
4	**サービス内容を正確に伝えられるか** 業務用語、サービスメニュー、逆翻訳、…
5	**質疑応答を記録できるか** サービス改善のヒント

そのためには、スマートフォンなどの個人端末による翻訳サービスよりも、専用の翻訳機を設置する方が効果的だと考え、「VoiceOn Information Desk」を開発・商品化しました。

この商品は、公共施設の窓口、ゴルフ場、シェアオフィス、カラオケ店など、さまざまな受付業務で活用されています。

写真：カラオケレインボー渋谷店 受付
月間 5,000 名のインバウンド利用がある人気店。

＜利用シーン：カラオケ店＞

タブレット 2 台を背中合わせに設置してコンパクトに常設。ゲストが言語メニューをタッチすると、大きな画面でチャットが始まり、操作はマイクボタンのみ。

カラオケ店では、ドリンクコースの説明が必要で、「ドリンクメニューの指差し＋翻訳機の併用」により、サービス内容を正確に伝えて、満足度の高い選択をサポートしています。よく使うメニュー説明フレーズも登録して簡単に呼び出せます。

従来使用していたパーソナル翻訳端末と比べて、大画面で即時対応が可能。ルームにタブレットを持ち込んでの接客にも活用されています。

Possibilities for exclusive translation machines that specialize in business issues.
Toshiyuki Oda
IP DREAM Inc.

3. スタンドアロン翻訳機

秘匿性が高い重要な会議や、通信環境が不安定な場所では、クラウド型よりもスタンドアロン型の翻訳サービスが適しています。

スタンドアロン型が適するビジネスシーン	
1	**秘匿性が高い会話を外部に漏らしたくない** 重要な会議、情報セキュリティ基準、…
2	**通信速度が悪い場所でも使える** 洋上、地下、海外、…
3	**被災で通信が完全遮断されても使える** 防災拠点、避難所、…

こうした多様なビジネスシーンをさらに分類・整理し、スタンドアロン翻訳機「VoiceOn Station」の 3 つのモデルを開発・商品化しました。いずれのモデルも、クラウド型サービスの機能を端末に搭載しており、それぞれの端末にお客様専用の独自辞書を適用できます。

VoiceOn Stationの3モデル	
1	**対面交渉用** ハンズフリーで双方同時に発話・翻訳
2	**プレゼンテーション用** 商品説明・提案が主で、質疑応答が従
3	**避難所用** 大人数の参加、一斉通知と個別相談

写真：VoiceOn Station 対面交渉用モデル

翻訳機本体に、マイクとスピーカーをそれぞれ 2 セット、さらにミニキーボードを接続します。電源を入れると、自動的に双方向の同時通訳が始まり、会話中に操作する必要は一切ありません。

対面交渉用モデルは、マイクとスピーカーをセットで提供することで、利用者がそれぞれ母国語で同時に話しながら、翻訳された音声をリアルタイムで聞くことができます。

音声入力は、音声認識に直結するため、会話中に飛び交う音声を専用機器で的確に拾うことが重要です。

4. 専用翻訳機の可能性

例に示したビジネスシーンのいずれの場面でも共通して求められるのは、「いざという時に確実に使えること」と、「重要な情報を的確に伝えられること」です。

さらに、業務の効率化、サービス体験の向上を図るためには、次のような機能拡張が効果的です。

(1) システム連携による対応力の強化

受付業務に関連する各種業務システムや、自動応答型の AI ガイド[2]と連携することで、より多くのお客様にスムーズに対応できるようになります。

(2) ハンズフリーデバイスによるサービス体験の向上

たとえば観光ガイドツアーでは、スマートフォンを操作せずに使えるスマートグラスの活用が有効です。こうしたデバイスをセットで提供することで、より確実で快適なサービス体験を実現できます。

多言語 AI コミュニケーションサービス「VoiceOn」は、翻訳 AI・生成 AI・Web テクノロジーを融合したクラウド型および端末型のラインナップを展開しています。さらに、関連サービスや最新技術との連携により、さまざまなビジネスシーンに柔軟に対応できるソリューションを提供しています。

参照：

[1]株式会社 IP DREAM（https://www.ip-dream.co.jp/）

[2]地域一体で育成する「多言語対応 AI コンシェルジュ」プロジェクト（東京都発表）
https://www.digitalservice.metro.tokyo.lg.jp/business/data-utilization/case-study/project-r602

法人会員 PR

翻訳支援ツールにおける LLM の活用：ProTranslator Neo

本間 奨
日本特許翻訳株式会社 [1]

1. 翻訳支援ツールの課題

翻訳メモリ（Translation Memory; TM）

従来の翻訳支援ツールにおいて、翻訳メモリ（TM）は重要な役割を果たしているものの、ファジーマッチ TM には根本的な課題があります。マッチ率が閾値未満のセグメントのファジーマッチ TM 訳については、翻訳者による修正作業が必要となり、結果的に翻訳効率の向上に限界がありました。特に、類似表現があるにもかかわらず、わずかな差異により修正作業が必要になるケースが多数存在していました。

用語ベース（Term Base; TB）

用語ベース（TB）についても、用語の不一致が発生した場合、翻訳者が手動で探し出し、訳語を TB に合わせるという編集作業が必要でした。この作業は時間を要するだけでなく、人的ミスの原因ともなっていました。

ニューラル機械翻訳（Neural Machine Translation; NMT）

NMT は翻訳品質の向上をもたらしましたが、以下のような課題が残存しています：

1. **長文における訳抜け**：複雑な構造の長文において、重要な情報が抜け落ちる
2. **訳文の過剰生成（湧き出し）**：原文にない表現が追加されることがまれに発生する
3. **未知語処理**：医薬品名やレアな化合物名などが未知語になり易い
4. **数値・単位・記号・参照符号の不整合**：医薬や特許・財務関係の文書において致命的な誤記が発生することがある

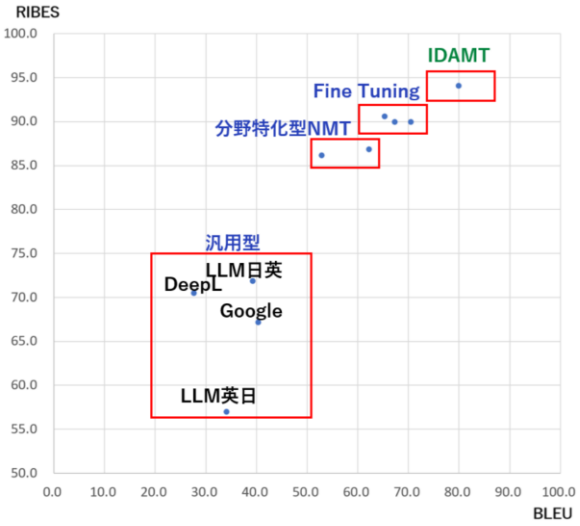
これらの課題により、NMT 使用時にはポストエディット（PE）作業が前提となっていました。

QA チェックの課題

従来の QA チェック機能は限定的で、数字チェックにおいて序数（First→第一）の対応ができない、誤訳の検知機能が不十分といった問題がありました。特に、文脈に依存する誤訳や専門用語の不適切な使用を検出することは困難でした。

2. LLM による課題解決

IDAMT（Instant Domain Adaptive Machine Translation）



当社が開発した IDAMT は、翻訳メモリの課題を根本的に解決する革新的な技術です。翻訳プロジェクトのファジーマッチを含む翻訳メモリから 100 個程度の対訳を抽出し、LLM に翻訳時にその場学習させます。**学習時間ゼロ**で、そのプロジェクトの原文特有の用語・文体に特化した翻訳が可能となり、RIBES 評価において 94.1 というスコアを達成しています。これは、Fine

Leveraging Large Language Models in Computer-Assisted Translation Systems
Susumu Honma
Nihon Patent Translation Co., Ltd.

Tuning（90.0）、分野特化型 NMT（86.9）、汎用型 LLM（71.9）を上回る精度を達成しています。(IPC G06F の特許請求項 500 文・自動評価)

NMT 参照 LLM 翻訳

ファジーマッチ翻訳メモリが利用できないセグメントは、NMT 訳が訳文の元になっていますが、NMT の課題があり、それに対しては、当社独自の「NMT 参照 LLM 翻訳」技術で対応しています。この技術は以下の 3 ステッププロセスで実行されます：

1. **ドメイン・サブドメイン・文書形式の特定**：原文全文を LLM に読み込ませ、専門分野と文書特性を識別
2. **NMT 翻訳の実行**：基本となる NMT 訳を生成
3. **LLM による高度な翻訳生成**：ドメイン情報と NMT 訳を参照し、専門性と自然さを両立した翻訳を実現

LLMTransPro [2]**:** 低めの翻訳メモリ閾値を設定し、閾値以上のセグメントについては IDAMT を、閾値未満のセグメントについては NMT 参照 LLM 翻訳を適用することで、全セグメントに対して高精度な翻訳を可能としました。またプロジェクト指定の TB を LLM で自動で調整することも可能としています。

TransCheckPro [2] **による品質評価革新**

QA チェックの課題解決として、当社が開発した TransCheckPro は、最大 7 つの評価観点からの詳細分析を実現します。複数セグメントを一括処理することで効率性を向上させながら、低品質と判定されたセグメントに対して修正訳を自動提案します。

3. 統合翻訳支援ツール ProTranslator Neo

統合プラットフォームの革新性

ProTranslator Neo は、3 つの革新的な「Pro」機能を統合した次世代翻訳プラットフォームです：

LLMTransPro：IDAMT と NMT 参照 LLM 翻訳のハイブリッドシステムにより、翻訳メモリの閾値に基づいて最適な翻訳方式を自動選択

PostEditPro™ [2]：機械翻訳の出力を自動改善し、未知語、訳文の過剰生成、訳抜け、数値・単位・記号・参照符号の不整合、用語統制、インラインタグ処理など、NMT の主要課題を自動解決

TransCheckPro：複数観点評価と用語統一機能を備えた AI 翻訳品質評価システム

セキュリティ面での優位性

ProTranslator Neo は、ISO27001/27017 認証を取得し、国内自社データセンターで運用されています。NMT 翻訳サーバー、LLM システム、memoQ オンプレミスシステムが同一データセンター内の自社サーバーで委託を伴わない自社従業員のみで運用されているため、機密性の高い特許出願書や医薬申請書類も安心してご利用いただけます。

クラウドサービスでありながら、顧客データの海外サーバーへの転送を行わない運用設計とされており、特許・医薬業界で求められる高度なセキュリティ要件を満たしています。

実証された効果

欧州特許公報（化学系明細書）の翻訳において、未知語、スペース欠損、誤訳、訳文の過剰生成、インラインタグ不正などの問題を 94%削減することを当社検証で確認しました。翻訳プロジェクト全体の所要時間を最大 85%削減し、特に人手によるポストエディットや品質チェックにかかるコストを大幅に削減します。

まとめ

ProTranslator Neo は、LLM 技術を活用して翻訳メモリおよび用語ベースの既存翻訳資産を最大限活用しながら、NMT と LLM の長所を融合させ、当社検証で確認された高品質な翻訳を実現します。特に特許・医薬・技術文書などの専門分野において、従来の翻訳ワークフローでは実現困難だった品質向上と効率化を同時に達成し、翻訳業界の未来を切り開く翻訳支援ツールです。

1. 日本特許翻訳株式会社 (https://npat.co.jp)
2. 商標：**PostEditPro™** は日本特許翻訳株式会社の商標です。**LLMTransPro／TransCheckPro** は特許・商標出願中。

編集後記

編集後記

石川 弘美
AAMT 編集委員会

つい先日、日本では新しい自民党総裁が誕生し、ほどなく新たな総理大臣が誕生することになります。この雑誌が刊行される頃には新内閣がスタートしていることでしょう。いま、社会全体が大きな変化の只中にあることを日々実感します。価値観や制度、そして人間の生き方そのものが再構築されつつある時代です。その変化の大きな要因の一つが、人工知能（AI）の急速な進歩でしょう。

機械翻訳は、AI 技術の中でも特に社会への実装が進んだ分野の一つです。近年は、観光案内や鉄道・交通機関、飲食店、行政サービスなど、あらゆる場面で機械翻訳が自然に使われています。もはや「翻訳機を使う」という意識すら希薄になり、ごく当たり前のこととなりました。翻訳という行為が、社会インフラの一部へと静かに溶け込んでいったのです。

子どものころ、アニメ『ドラえもん』に登場する「ほんやくコンニャク」に夢を抱いた方も多いでしょう。言葉の壁を超える道具は、当時は空想上の象徴でした。しかしいまや、私たちはスマートフォンを介して、リアルタイムで異なる言語の人々と意思疎通ができます。人の想像力が技術を生み出し、技術が再び人の想像力を刺激する。その循環の中に、翻訳技術の発展が位置づけられていると感じます。

今号では、AAMT 長尾賞学生奨励賞の受賞者や AAMT 若手研究会など、若い世代の方々による寄稿が多く集まりました。いずれの論考からも、既存の枠組みにとらわれない新しい視点と、実践的な問題意識が感じられます。若手研究会は今後も毎年３月に行いますので、機械翻訳に限らず翻訳に関する知見のある方の発表をお待ちしています。

機械翻訳の技術は、この十数年で飛躍的な進歩を遂げました。統計的機械翻訳からニューラル機械翻訳（NMT）へ、そして大規模言語モデルによる生成型翻訳の時代へと移り変わっています。その過程で、翻訳の品質だけでなく、「翻訳という行為の定義」そのものが変化しつつあります。AI が言語の意味構造を理解し、人間がその文脈的・文化的意義を補完する――そうした共創的翻訳のあり方が、今後ますます重要になると考えます。

そして、心で思ったことがそのまま伝達されるような未来が、もしかするとそれほど遠くないかもしれません。そのとき私たちは、何を「伝える」と感じ、何を「言葉」と呼ぶのでしょうか。AI が媒介する言語の未来は、人間の思考と感情の在り方をも問い直す可能性を秘めています。恐れを抱きつつも、理性的な好奇心をもって、次の時代を見つめていきたいと思います。

なお、今号から誌面のフォーマットを一部変更し、読みやすさと視認性の向上を図りました。余白のとり方、フォントの調整など、細部にわたって改良を重ねています。ご意見やご感想がありましたら、ぜひ編集部までお寄せください。読者の皆さまの声をもとに、より開かれた、知の共有の場としての誌面づくりを目指してまいります。

技術がいかに進歩しても、言葉を介して人と人が理解し合うという営みの本質は変わりません。機械翻訳はその過程を支える知的基盤として、今後も社会とともに進化していくでしょう。AAMT は来年で創立３５周年を迎えます。本誌が、これからもその歩みを記録し、次代の研究者・実務者にとっての羅針盤となることを願っております。

Editor’s note
Hiromi Ishikawa
AAMT Editorial Board

AAMTジャーナル「機械翻訳」No. 83

【 発 行 日 】2025年11月29日
【 発 行 】アジア太平洋機械翻訳協会（AAMT）
ホームページ：https://aamt.info/
【 住 所 】〒160-0004
東京都新宿区四谷4-7新宿ヒロセビル5F
一般社団法人アジア太平洋機械翻訳協会（AAMT）事務局
【編集委員会】内山将夫　後藤功雄　中澤敏明　新田順也　園尾聡　森口功造　隅田英一郎
石川弘美　早川威士　出内将夫
【表紙デザイン】泉谷東十郎
【 題 字 】長尾真
【 事 務 局 】奥麻里
【 印 刷 所 】株式会社 プリントパック

Asia-Pacific Association for Machine Translation (AAMT)
Shinjuku Hirose Bldg. 5F, 4-7 Yotsuya, Shinjuku-ku, Tokyo 160-0004 Japan


AAMT
Asia-Pacific Association for Machine Translation
アジア太平洋機械翻訳協会