ISSN 1883-1818
No.81
December 2024

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.81

目　次

C O N T E N T S

巻頭言

要件定義

安達 久博
株式会社サン・フレア

今から１６年前にあるプロジェクトを担当していました。いわゆる電子商取引（EC）の海外展開に機械翻訳を利活用して行きましょうという内容でした。当時の機械翻訳方式は統計的機械翻訳が主流になりつつありました。

海外との商取引を考える場合、大きく３つの流れを考える必要があります。１つは物の流れ（物流）、２つ目はお金の流れ(金流)、そして３つ目は情報の流れ(情報流)。この中で商品の配送（物流）と決済（金流）については既に国際間の取り決めが整備されています。

一方、情報流については商品の説明文を日本語から英語に翻訳する必要があります。事業者の多くは無料の翻訳システムを利用していたのですが翻訳精度に問題がありました。そのため、大量の商品説明文を収集し、対訳コーパスを作成し、その対訳コーパスを機械学習し、専用の機械翻訳システムで翻訳精度を改善してゆくプロジェクトです。設定した目標精度を達成し、本プロジェクトは成功しました。

翻訳精度をさらに高めるためには、さらに大量の対訳コーパスを構築することを考えましたが、事業者との話の中で予算が難しいのでポストエディット（PE）などの案がでました。今から思うと、統計的翻訳方式の限界もあったのかなと思います。

そんな時に、要件定義をお客様と交渉し、納期も決めて始めたところ、納品前に対象の商品が在庫切れになるほど売れてしまって、その商品説明文は必要なくなってしまうという想定外の出来事がありました。翻訳ビジネスの厳しさを痛感させられました。賞味期限のある生ものなのですね。

このように、お客様の担当者と仕上がりのイメージに齟齬がないように要件定義を詰めておくことはとても重要になります。特に、翻訳に関する国際標準である ISO 規格の取り扱いが課題になるのではないかと考えています。ISO 規格には、品質の高い翻訳を提供するための ISO17100、PE の品質を保証するための作業概要、プロセスなどを規定した ISO18587 があり、両規格を改定する議論が始まっています。同時に両規格を一緒にした方が良いのではないかという議論も出ています。

ニューラル機械翻訳が主流となってきている現在、機械翻訳システムの翻訳精度が飛躍的に向上しているとはいえ、翻訳出力の品質保証は人間にゆだねられています。すなわち、最終的に専門家、翻訳者のレビューによる最終確認、品質保証がとても重要になってきています。

今後、機械翻訳と生成 AI（人工知能）を組み合わせた MTPE（機械翻訳＋生成 AI）による要件定義がどのようになるのかは皆さんと今後、議論を深めてゆく必要があると思います。求められる要件定義のポイントは「品質、スピード、コスト」のバランスを発注者、受注者双方が同意できるものであることであり、文化の異なる他国の人々に単なる「ことば」や「こと」の翻訳から、伝えるべき「こころ」を正しく自然に翻訳するように努めることが肝要となります。

最期に、この原稿を執筆している今日は「翻訳の日」です。ユネスコが定めた「世界翻訳の日」であり、日本翻訳連盟（JTF）が日本記念日協会に申請し、認定されています。また、国連総会にて「国どうしを結び、平和と理解、発展を育むうえで専門の翻訳者が果たす役割」を重視する決議が採択され、9 月 30 日が国際翻訳デー（International Translation Day）として、公式に認定されています。

Requirements definition
Hisahiro Adachi
SunFlare Co., Ltd.

第 19 回 AAMT 長尾賞受賞記念

LLM-jp

黒橋 禎夫
国立情報学研究所

大規模言語モデル（Large Language Model, LLM）をベースとした ChatGPT、特に GPT-4 は驚くほど賢い。LLM とは、あるテキスト（文脈）が与えられた時に、それに続く単語をよりよく予測するように、大規模なコーパスを用いてパラメータの調整がなされた超大規模なニューラルネットワークである。このように次の単語を予測するという極めて単純な学習を行うだけであるにも関わらず、そのモデルの振舞いは従来の自然言語処理のタスク、たとえば構文の解析、意味の解析、文脈の解析、さらには機械翻訳、そしてもちろん人との自然な対話においても、これまでの長年の目標としていたレベルに到達したといってよいだろう。

実際、入力の深い理解や、言葉の使い方やニュアンスの説明まで行える翻訳能力には脱帽する。たとえば、日本語では「部長はうなぎだ」のように、「○は△だ」の構文で、あるトピック（○）について何か（△）を言う表現が広く受け入れられる。しかし、その英語への直訳"The department head is an eel"は英語の世界では不自然な表現である。高精度なニューラル翻訳システムでもこのような不自然な直訳を示すにとどまるが、GPT-4 にその翻訳を頼むと、直訳を示した上で「ただし、このフレーズは日常的な会話で直接使用されるものではなく、それが使われる文脈を考慮した訳が必要となることも考えられます」というような応答を行う。こちらが「お店で注文している場合は？」というと「そのような場面では"The manager would like eel"もしくは"The boss orders eel"となるでしょう」と教えてくれる。

このような LLM の進化によって、自然言語処理が取り組むべき課題は、その賢さを解明すること、その安全性などを検討すること、そしてそれが社会にどのように受容され人間と共存していくかをデザインしていくことなどに一気に様変わりした。しかし、ここで大きな問題がある。LLM の研究開発には膨大な計算資源・資金が必要となり、一部の外国組織の寡占状態である。そして、強い、大きなモデルの中身（アーキテクチャ、事前学習コーパス、学習ノウハウ、チューニングデータなど）はもはや公開されていない。一方で、ハルシネーションや安全性の課題など、LLM が今後社会に本格的に受け入れられていくためにはまだまだ課題がある。日本としての懸念もある。GPT-4 のような英語コーパス中心のモデルが世界標準となれば、日本の活動や文化が埋没していく懸念もある。また、外国モデルに完全に依存して日本の知的資産がただ流れていくことは経済安全保障上も看過できない。

このような問題意識から、日本でも LLM を作ろう、そのための勉強会をはじめようということで、2023 年 5 月に第一回の LLM 勉強会を開催した。LLM の研究開発は計算資源的にも人的資源的にももはやビッグサイエンスである。できるだけ多くの人が協力して取り組む必要があるので、その活動から生まれるモデル、コーパス、チューニングデータなどはもちろん、議論の過程や失敗を含めてすべてオープンにすること（商用利用も可）を方針とした。第一回の LLM 勉強会は 30 名ほどの自然言語処理研究者による、まさに勉強会としてスタートしたが、この問題意識と考え方に賛同する参加者が日に日に増加していった。

当初、計算資源の目処がなかったので、この活動の日本語名称は LLM 勉強会、英語名称は LLM-jp としてスタートした。6 月になって、2023 年 4 月から有料サービスを開始したデータ活用社会創成プラットフォーム mdx の計算資源に少し余裕があることがわかり、NII、理研 AIP、そして JHPCN（学際大規模情報基盤共同利用・共同研究拠点）で資金を持ち寄り、

LLM-jp
Sadao Kurohashi
National Institute of Informatics

40G メモリーの A100 GPU を 8 基搭載したノードを 16 台専有する計算環境を手に入れた。7 月以降はこの環境を用いて実際にモデルを作る活動をはじめたので、会の名称は日本語でも LLM-jp を使うこととした。

LLM 構築のために、コーパス構築 WG、モデル構築 WG、チューニング・評価 WG を作り、議論はそれぞれ週一回程度のオンラインミーティングと Slack で行った。また、活動の進捗に伴い、学術ドメイン WG、安全性 WG、なども立ち上げた。ハイブリッド（リアル＋オンライン）の LLM 勉強会は 2023 年 5 月以降、月 1 回のペースで続けており、LLM に関する最新のトピックの紹介、日本で構築されるさまざまなモデルの紹介、WG の活動紹介などを行っている。

LLM-jp の最初のモデルとして、2023 年 10 月 20 日に 130 億パラメータの LLM-jp-13B を公開した。モデルそのものはもちろん、コーパスやチューニングデータも公開している。さらに、モデルの入出力に類似するコーパス中のテキストを検索する機能もそなえている。つまり、何を根拠にそのような生成がなされているかを観察できる環境を整えている点に特徴がある。

さらに、2023 年 11 月には ABCI の第 2 回 LLM 構築支援プログラムに採択され、少量のコーパスではあるが 1750 億パラメータのモデル学習の実験を行った。2024 年 1 月には GENIAC 第 1 期に採択され、2024 年 4 月から 1720 億パラメータのモデルの事前学習を 2.1 兆単語のコーパスで行うこととした。

このような成果をベースに、文部科学省において「生成 AI モデルの透明性・信頼性の確保に向けた研究開発拠点形成」（事業期間：2023 年度～2028 年度）が始まり、2024 年 4 月に国立情報学研究所に大規模言語モデル研究開発センター（以下 LLM センターと略称）を設置することとなった。研究者、開発者、RA（学生 Research Assistant）、データ作成者、総勢約 70 名の規模である。LLM-jp は LLM センターの活動の一貫と位置づけ、LLM-jp のオープンな場で議論しながら、LLM の透明性・信頼性・安全性の確保、原理解明、高度化等の研究開発に取り組んでいる。

2024 年 4 月から始めた GENIAC での 1720 億パラメータ・モデルの学習では、当初、想定していたモデル性能を得ることができなかったが、徹底的な ablation study を行い、問題（パラメータ設定）を究明し、学習を最初からやり直した。このため、GENIAC の計算資源利用期限の 8 月中旬までには約 0.4 兆単語での学習しかできなかったが、幸い 8 月から利用可能となった LLM センターの計算資源を用いて学習を継続し、2024 年 9 月 17 日に、学習データを含めてすべてオープンにしたモデルとしては世界最大の、1720 億パラメータ・モデルを公開した（学習コーパス量は 0.7 兆単語）。今後も学習を継続し、12 月中には 2.1 兆単語で学習したモデルを公開する予定である。

このように、まずは LLM を作ってみないと始まらない、日本において LLM を研究開発する「場」が必要である、という問題意識で始めた LLM-jp の活動は、多くの方に賛同頂き、これまで順調に発展してきた。2024 年 10 月現在、参加者は 1800 名を超えており、2024 年度にはマルチモーダル WG、実環境インタラクション WG なども立ち上げ、活動の幅を広げている。LLM の研究開発競争は世界的にますます激化しているが、我々は今後も、多くの人が協力するオープンな活動として、そして包括的に、この問題に取り組んでいきたいと考えている。

LLM-jp の活動に対して、2024 年 6 月に第 19 回長尾賞を授賞頂いた。審査委員の皆様、AAMT の関係各位に心より御礼申し上げる。最後に、我々の活動とその精神をよく表している諺で本稿を締めくくりたい（アフリカの諺とも言われるが、起源は定かではないようである）。

If you want to go fast, go alone.

If you want to go far, go together.

第 11 回 AAMT 長尾賞学生奨励賞受賞記念

Breaking Language Barriers: Enhancing Multilingual Representation for Sentence Alignment and Translation

Zhuoyuan Mao

Sony Group Corporation

1. Introduction

In a diverse linguistic landscape where over 7,100 languages are spoken, vast swathes of digital content remain isolated within language silos, creating significant barriers to global communication. Bridging these gaps is the purview of multilingual representation learning, an emerging field within natural language processing (NLP) that seeks to develop computational models capable of understanding and translating across multiple languages. This specialized area of research aims to dismantle linguistic barriers, facilitating the flow of information and ideas in our increasingly interconnected world.

This work delves into the intricacies of multilingual representation learning, concentrating on two pivotal tasks: multilingual sentence embedding (MSE) learning and multilingual neural machine translation (NMT). These tasks are key objectives of multilingual representation learning due to their profound impact on facilitating communication across language barriers. MSE learning enables the alignment of semantically similar sentences from different languages, serving as a key enabler for applications such as cross-lingual information retrieval and parallel corpus construction. Meanwhile, multilingual NMT extends the boundaries of language translation to a multilingual context, which is crucial for real-time interpretation and content localization. Ultimately, MSE learning and multilingual NMT encapsulate the primary objectives of multilingual representation learning, underpinning innovative applications that help dissolve language barriers, thus granting more equitable access to information and fostering cross-cultural understanding in a multilingual world.

Within multilingual representation learning, specifically for applications in alignment and translation tasks, three major challenges persist: (1) high computational demands, which refers to the significant computational overhead incurred in scaling up the language coverage of a multilingual model; (2) data scarcity, the lack of sufficient and diverse language data, particularly for low-resource languages; (3) limitations in Transformer architecture, meaning the current Transformer models are not fully appropriate for the complexities of processing multiple languages. Addressing these challenges is crucial for further advancement in this field. To this end, this work seeks to provide solutions to these existing challenges while also exploring potential approaches for enhancing recent large multilingual language models (LLMs). Eventually, we expect to pave the way for more advanced and efficient multilingual representation learning, thus broadening the reach of NLP techniques to a wider audience.

2. The Challenge of High Computation Demands

To tackle the challenge of high computation demands associated with expanding the language support in training MSE models, we first introduce efficient and effective massively multilingual sentence embedding, using cross-lingual token-level reconstruction and sentence-level contrastive learning as training objectives. Compared with related studies, the proposed model can be

efficiently trained using significantly fewer parallel sentences and GPU computation resources. Secondly, we introduce a novel distilled MSE model to streamline the inference process for MSE models. Precisely, we systematically explore learning language-agnostic sentence embeddings with lightweight models. We demonstrate that a thin-deep encoder can construct robust low-dimensional sentence embeddings for 109 languages. With our proposed distillation methods, we achieve further improvements by incorporating knowledge from a teacher model.

EMS: Efficient and Effective Massively Multilingual Sentence Representation Learning [1]

Massively multilingual sentence representation (MSE) models, e.g., LASER[2], SBERT-distill[3], and LaBSE[4], help significantly improve cross-lingual downstream tasks. However, the use of a large amount of data or inefficient model architectures results in heavy computation to train a new model according to our preferred languages and domains. To resolve this issue, we introduce efficient and effective massively multilingual sentence embedding (EMS), using cross-lingual token-level reconstruction (XTR) and sentence-level contrastive learning as training objectives, as shown in Figure 1.

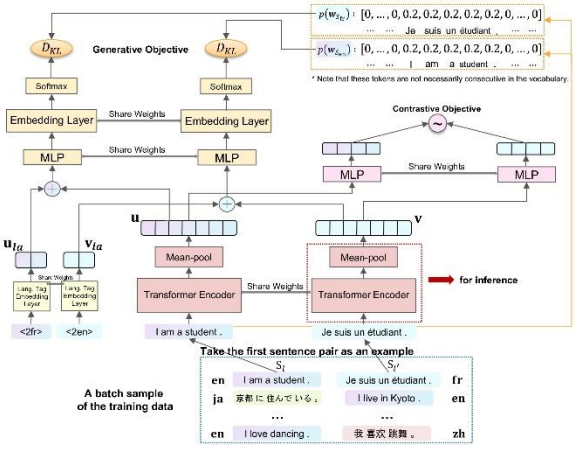


Figure 1: Training architecture of EMS. u and v are MSEs for inference, and the model components in the red dashed rectangle are used for inference. u_{la} and v_{la} are the target language embeddings. "+" denotes the hidden vector concatenation.

Compared with related studies, the proposed model can be efficiently trained using significantly fewer parallel sentences and GPU computation resources. Empirical results showed that the proposed model significantly yields better or comparable results with regard to cross-lingual sentence retrieval, zero-shot cross-lingual genre classification, and sentiment classification. Ablative analyses demonstrated the efficiency and effectiveness of each component of the proposed model. We release the codes for model training and the EMS pre-trained sentence embedding model, which supports 62 languages (https://github.com/Mao-KU/EMS).

LEALLA: Learning Lightweight Language-agnostic Sentence Embeddings with Knowledge Distillation [5]

Large-scale language-agnostic sentence embedding models such as LaBSE [4] obtain state-of-the-art performance for parallel sentence alignment. However, these large-scale models can suffer from inference speed and computation overhead.

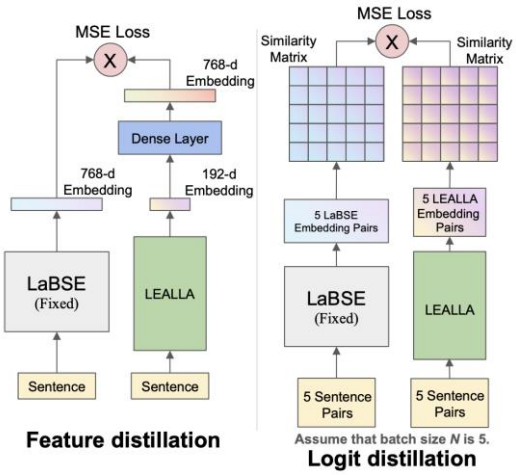


Figure 2: Feature and logit distillation from LaBSE.

This study systematically explores learning language-agnostic sentence embeddings with lightweight models. We demonstrate that a thin-deep encoder can construct robust low-dimensional

sentence embeddings for 109 languages. With our proposed distillation methods (Figure 2), we achieve further improvements by incorporating knowledge from a teacher model. Empirical results on Tatoeba, United Nations, and BUCC show the effectiveness of our lightweight models. We release our lightweight language-agnostic sentence embedding models LEALLA on TensorFlow Hub.

3. The Challenge of Data Scarcity in Low-resource Languages

To tackle the challenge of data scarcity in low-resource languages, we first introduce innovative sequence-to-sequence pre-training objectives for low-resource NMT to leverage the linguistic knowledge to compensate for the lack of training data. The proposed methods employ phrase structure masking and reordering tasks. Secondly, we propose word-level contrastive learning to leverage statistical word alignments for low-resource multilingual NMT, without the requirement to use high-quality bilingual dictionaries. Additionally, we introduce contrastive alignment instructions to address the challenge of the lack of data in low-resource languages. AlignInstruct emphasizes cross-lingual supervision via a cross-lingual discriminator built using statistical word alignments, which is empirically demonstrated superior to NMT instruction tuning baseline methods.

Linguistically-driven Multi-task Pre-training for Low-resource Neural Machine Translation [6]

In this study, we propose novel sequence-to-sequence pre-training objectives for low-resource machine translation (NMT): Japanese-specific sequence to sequence (JASS) for language pairs involving Japanese as the source or target language, and English-specific sequence to sequence (ENSS) for language pairs involving English.

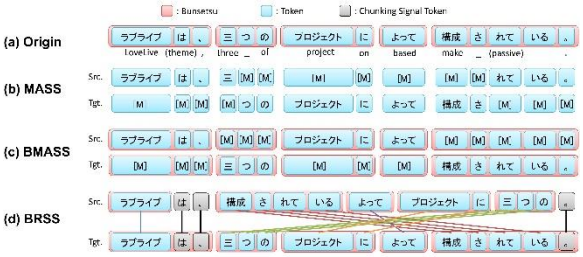


Figure 3: Example of source and target for MASS, BMASS, and BRSS with the meaning "LoveLive is made of three projects." JASS is a joint training of BMASS and BRSS.

JASS (Figure 3) focuses on masking and reordering Japanese linguistic units known as bunsetsu, whereas ENSS is proposed based on phrase structure masking and reordering tasks. Experiments on ASPEC Japanese--English & Japanese--Chinese, Wikipedia Japanese--Chinese, News English--Korean corpora demonstrate that JASS and ENSS outperform MASS and other existing language-agnostic pre-training methods by up to +2.9 BLEU points for the Japanese--English tasks, up to +7.0 BLEU points for the Japanese--Chinese tasks and up to +1.3 BLEU points for English--Korean tasks. Empirical analysis, which focuses on the relationship between individual parts in JASS and ENSS, reveals the complementary nature of the subtasks of JASS and ENSS. Adequacy evaluation using LASER [2], human evaluation, and case studies reveals that our proposed methods significantly outperform pre-training methods without injected linguistic knowledge and they have a larger positive impact on the adequacy as compared to the fluency. We release codes here: https://github.com/Mao-KU/JASS/tree/master/linguistically-driven-pretraining.

When do Contrastive Word Alignments Improve Many-to-many Neural Machine Translation? [7]

Word alignment has proven to benefit many-to-many neural machine translation (NMT). However, high-quality ground-truth bilingual dictionaries were used for pre-editing in previous methods, which are unavailable for most language pairs. Meanwhile, the

contrastive objective can implicitly utilize automatically learned word alignment, which has not been explored in many-to-many NMT. This work proposes a word-level contrastive objective to leverage word alignments for many-to-many NMT, as shown in Figure 4.

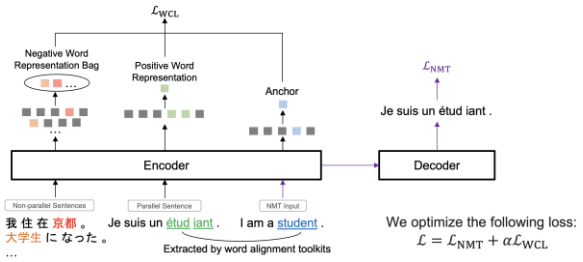


Figure 4: Word-level contrastive learning for leveraging word alignments to improve multilingual NMT.

Empirical results show that this leads to 0.8 BLEU gains for several language pairs. Analyses reveal that in many-to-many NMT, the encoder's sentence retrieval performance highly correlates with the translation quality, which explains when the proposed method impacts translation. This motivates future exploration for many-to-many NMT to improve the encoder's sentence retrieval performance.

Tuning LLMs with Contrastive Alignment Instructions for Machine Translation in Unseen, Low-resource Languages [8]

This study introduces contrastive alignment instructions (**AlignInstruct**) to address two challenges in machine translation (MT) on large language models (LLMs). One is the expansion of supported languages to previously unseen ones. The second relates to the lack of data in low-resource languages. Model fine-tuning through MT instructions (**MTInstruct**) [9] is a straightforward approach to the first challenge. However, MTInstruct is limited by weak cross-lingual signals inherent in the second challenge. AlignInstruct emphasizes cross-lingual supervision via a cross-lingual discriminator built using statistical word alignments. Our results based on fine-tuning the BLOOMZ [10] models (1b1, 3b, and 7b1) in up to 24 unseen languages showed that: (1) LLMs can effectively translate unseen languages using MTInstruct; (2) AlignInstruct led to consistent improvements in translation quality across 48 translation directions involving English, as shown in Figure 5; (3) Discriminator-based instructions outperformed their generative counterparts as cross-lingual instructions; (4) AlignInstruct improved performance in 30 zero-shot directions.

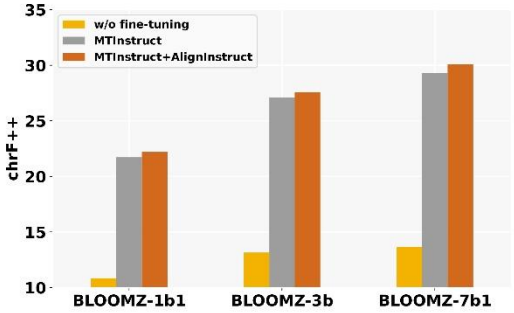


Figure 5: Average chrF++ scores of BLOOMZ models across 24 unseen languages.

4. The Challenge of Limitations in Transformer Architecture for Zero-shot NMT

To tackle the challenge of limitations in Transformer architecture for zero-shot NMT, we first unveil a novel Transformer architecture that constructs universal interlingua representations atop Transformer encoder. This development significantly enhances the performance of zero-shot NMT than standard Transformer architectures. Moreover, we comprehensively explore the effects of layer normalization on zero-shot NMT. Our results demonstrate that post-layer normalization consistently outperforms pre-layer normalization for zero-shot NMT, regardless of the language tag and residual connection settings.

Variable-length Neural Interlingua Representations for Zero-shot Neural Machine Translation [11]

The language-independency of encoded representations within multilingual neural machine translation (MNMT) models is crucial for their generalization ability on zero-shot translation. Neural interlingua

representations (Figure 6) [12] were shown as an effective method for achieving this.

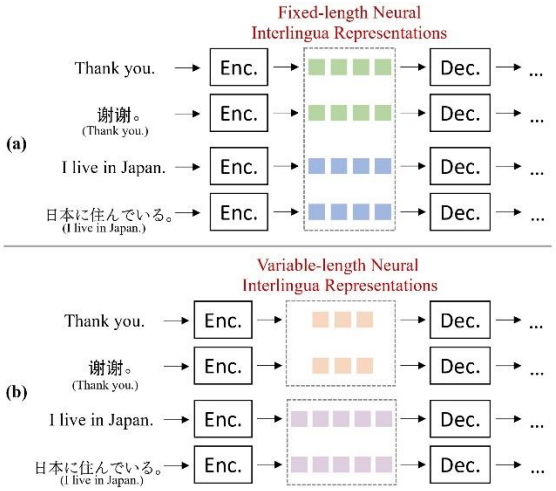


Figure 6: (a) Previous fixed-length neural interlingua representations; (b) Our proposed variable-length neural interlingua representations. Each colored box denotes the representation on the corresponding position.

However, fixed-length neural interlingua representations introduced in previous work can limit its flexibility and representation ability. In this study, we introduce a novel method to enhance neural interlingua representations by making their length variable, thereby overcoming the constraint of fixed-length neural interlingua representations. Our empirical results on zero-shot translation on OPUS, IWSLT, and Europarl datasets demonstrate stable model convergence and superior zero-shot translation results compared to fixed-length neural interlingua representations. However, our analysis reveals the suboptimal efficacy of our approach in translating from certain source languages, wherein we pinpoint the defective model component in our proposed method.

Exploring the Impact of Layer Normalization for Zero-shot Neural Machine Translation [13]

This paper studies the impact of layer normalization (LayerNorm) on zero-shot translation (ZST). Recent efforts for ZST often utilize the Transformer architecture as the backbone, with LayerNorm at the input of layers (PreNorm) [14] set as the default, as shown in Figure 7.

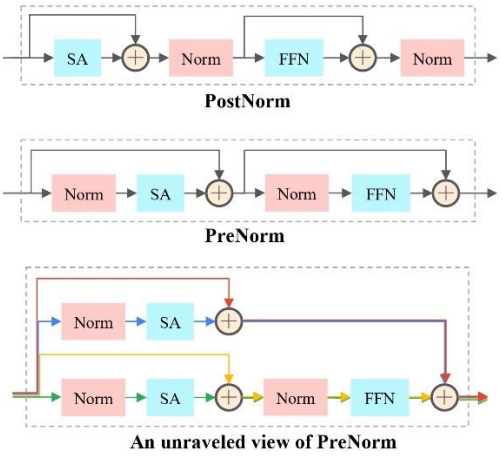


Figure 7: PostNorm, PreNorm, and an unraveled view of PreNorm in a Transformer encoder layer. "Norm," "SA," and "FFN" denote LayerNorm, self-attention, and feed-forward network. "+" is residual connection. Paths with different colors in the unraveled view of PreNorm indicate respective sub-networks.

However, Xu et al. [15] has revealed that PreNorm carries the risk of overfitting the training data. Based on this, we hypothesize that PreNorm may overfit supervised directions and thus have low generalizability for ZST. Through experiments on OPUS, IWSLT, and Europarl datasets for 54 ZST directions, we demonstrate that the original Transformer setting of LayerNorm after residual connections (PostNorm) consistently outperforms PreNorm by up to 12.3 BLEU points. We then study the performance disparities by analyzing the differences in off-target rates and structural variations between PreNorm and PostNorm. This study highlights the need for careful consideration of the LayerNorm setting for ZST.

5. Conclusion and Future Prospects

In conclusion, the research presented in this work marks a significant stride in the field of multilingual NLP. It has not only provided a deeper understanding of the challenges inherent in

multilingual representation learning but also offered innovative and practical solutions to overcome these obstacles. As the world becomes increasingly interconnected, the importance of effective multilingual communication grows. The contributions of this thesis thus hold considerable promise for future applications in global communication, information access, and beyond, fostering a world where language barriers continue to diminish.

As future prospects, firstly, the techniques presented in this work hold the potential for integration into a singular, comprehensive multilingual model, an endeavor we aim to pursue in future research. Initially, this integration would involve combining the MSE and multilingual NMT models within a unified Transformer encoder-decoder framework. Here, sentence embeddings would be generated by the encoder, while the decoder would produce translations. Following this, the proposed methods for enhancing multilingual representation, particularly those aimed at increasing efficiency and boosting performance in low-resource languages, could be combined into a single, cohesive training phase.

Secondly, the insights obtained from this work could significantly contribute to the development of robust multilingual LLMs. Our proposed training objectives, which focus on word alignment and linguistic features, could effectively facilitate better language alignment in multilingual LLMs. Moreover, the efficient MSE models we introduced could enhance the retrieval-based applications of LLMs, such as retrieval-based few-shot in-context learning. Furthermore, our findings regarding the application of Transformer architectures in multilingual contexts offer valuable guidance for future research into the Transformer architectures of multilingual LLMs.

Last but not least, looking beyond the conclusion of this work, several promising avenues for future research in multilingual NLP emerge. These prospects not only aim to broaden the scope of current methodologies but also seek to deepen the understanding and application of multilingual representation learning.

References

[1] Z. Mao, C. Chu, and S. Kurohashi. EMS: efficient and effective massively multilingual sentence representation learning. IEEE ACM Trans. Audio Speech Lang. Process. 32: 2841-2856, 2024.

[2] M. Artetxe and H. Schwenk. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. Transactions of the Association for Computational Linguistics, 7:597–610, Mar. 2019.

[3] N. Reimers and I. Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 4512–4525, Online, Nov. 2020. Association for Computational Linguistics.

[4] 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), pages 878–891, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[5] Z. Mao and T. Nakagawa. LEALLA: learning lightweight language-agnostic sentence embeddings with knowledge distillation. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 1878–1886, 2023.

[6] Z. Mao, C. Chu, and S. Kurohashi. Linguistically driven multi-task pre-training for low-resource neural machine translation.

ACM Trans. Asian Low Resour. Lang. Inf. Process., 21(4):68:1–68:29, 2022.

[7] Z. Mao, C. Chu, R. Dabre, H. Song, Z. Wan, and S. Kurohashi. When do contrastive word alignments improve many-to-many neural machine translation? In Findings of the Association for Computational Linguistics: NAACL 2022, pages 1766–1775, Seattle, United States, July 2022. Association for Computational Linguistics.

[8] Z. Mao and Y. Yu. Tuning LLMs with Contrastive Alignment Instructions for Machine Translation in Unseen, Low-resource Languages. In Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024), pages 1–25, Bangkok, Thailand. Association for Computational Linguistics.

[9] J. Li, H. Zhou, S. Huang, S. Chen, and J. Chen. Eliciting the translation ability of large language models via multilingual finetuning with translation instructions. CoRR, abs/2305.15083, 2023.

[10] N. Muennighoff, T. Wang, L. Sutawika, A. Roberts, S. Biderman, T. Le Scao, M. S. Bari, S. Shen, Z. X. Yong, H. Schoelkopf, X. Tang, D. Radev, A. F. Aji, K. Almubarak, S. Albanie, Z. Alyafeai, A. Webson, E. Raff, and C. Raffel. Crosslingual generalization through multitask fine-tuning. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15991–16111, Toronto, Canada, July 2023. Association for Computational Linguistics.

[11] Z. Mao, H. Song, R. Dabre, C. Chu, and S. Kurohashi. 2023. Variable-length Neural Interlingua Representations for Zero-shot Neural Machine Translation. In Proceedings of the 1st International Workshop on Multilingual, Multimodal and Multitask Language Generation, pages 16–25, Tampere, Finland. European Association for Machine Translation.

[12] Y. Lu, P. Keung, F. Ladhak, V. Bhardwaj, S. Zhang, and J. Sun. A neural interlingua for multilingual machine translation. In Proceedings of the Third Conference on Machine Translation: Research Papers, pages 84–92, Brussels, Belgium, Oct. 2018. Association for Computational Linguistics.

[13] Z. Mao, R. Dabre, Q. Liu, H. Song, C. Chu, and S. Kurohashi. Exploring the impact of layer normalization for zero-shot neural machine translation. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 1300–1316, Toronto, Canada, July 2023. Association for Computational Linguistics.

[14] A. Baevski and M. Auli. Adaptive input representations for neural language modeling. In 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net, 2019.

[15] J. Xu, X. Sun, Z. Zhang, G. Zhao, and J. Lin. Understanding and improving layer normalization. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alch´e-Buc, E. B. Fox, and R. Garnett, editors, Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, pages 4383–4393, 2019.

第 11 回 AAMT 長尾賞学生奨励賞受賞記念

実応用に向けたストリーミング音声機械翻訳

福田りょう
日本電信電話株式会社

1. はじめに

本稿では，第 11 回 AAMT 長尾賞学生奨励賞を受賞した博士論文 “Towards Streaming Speech Translation for Real-world Scenarios” [Fukuda, 2024] について解説する．本稿に含まれる図及び実験結果は，すべて当該の博士論文および博士論文公聴会の発表資料より引用し和訳したものである．

音声機械翻訳（Speech Translation; ST）は，異なる言語を話す人同士のコミュニケーション支援，動画の翻訳字幕作成, 会議の議事録作成など, 多言語コミュニケーションにおいて重要な役割を果たしている．近年の ST に関する研究では，発話単位で区切られた 1〜10 秒程度の短い音声を正しく翻訳することに焦点が当てられている．しかし，実用の場面では事前分割されていない音声を処理する必要があり，その長さは何十秒，何百秒と長く続くことも想定される．音声を長さに限りがないストリームデータとして捉え，適切に分割しながらリアルタイムに翻訳するシステムをストリーミング ST と呼ぶ．博士論文では，ストリーミング ST を実用化するために対処すべき 2 つの課題に取り組んだ．

- 課題 1. 音声認識誤りの伝播：ASR モデルと MT モデルを組み合わせた Cascade ST では，ASR モデルで生じた誤りが後続の MT モデルに悪影響を及ぼす．この問題は Cascade ST システムにおける音声認識誤りの伝播（ASR Error propagation）として知られる [Zhang et al., 2004]．高速な処理が求められるストリーミング ST においてこの問題は特に深刻である．本研究では，音声認識誤りを含まない人手による書き起こしと知識蒸留 [Hinton et al., 2015] を用いて音声認識誤りに頑健な MT モデルを構築する手法を提案し，英伊・西英翻訳において提案手法の有効性を示した [Fukuda et al., 2021].
- 課題 2. 音声分割手法の精度不足：ストリーミング ST では，長い音声を翻訳するために，事前に音声を翻訳単位（セグメント）に分割する必要がある．これまで，音声が存在する区間を切り出す音声区間検出（Voice Activity Detection; VAD）が広く用いられてきた [Hughes and Mierle, 2013]．しかし，音声区間は必ずしも意味的な発話のまとまりと一致しないため，文を途中で分割するような過剰分割がしばしば起こり，翻訳精度の低下に繋がる [Wan et al., 2021]．そのため本研究では，ST で高い翻訳精度を達成する音声分割モデルを提案した．その後，音声分割モデルと同時 ST モデルと組み合わせたストリーミング ST システムを構築した（図 1）．

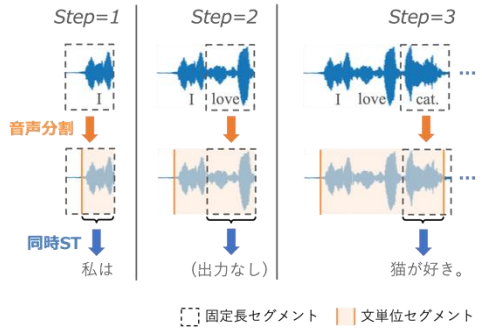


図 1. ストリーミング ST システム

本稿では, 課題 2 への取り組みとして, 2. Transformer に基づく音声分割モデル，3. Transformer に基づく同時 ST モデル, 4. ストリーミング ST システムを紹介する．

Towards Streaming Speech Translation for Real-world Scenarios
Ryo Fukuda
NTT Corporation

2. Transformer に基づく音声分割モデル [Fukuda et al., 2022]

前述した通り，VAD を用いた音声分割では過剰分割による翻訳精度の悪化が指摘されている．過剰な分割を緩和するために，一定の長さまで VAD によるセグメントを連結する方法も提案されている [Gaido et al., 2021]．しかし，音声を翻訳に適した「発話の意味的なまとまり」に分割する試みは行われていない．ST コーパスには，人手を用いて分割された文単位の音声セグメントが含まれており，ST モデルは通常これを用いて文単位の翻訳を学習する．文は意味的なまとまりであり，また ST 学習時の単位でもあるため，翻訳に適した処理単位とみなすことができる．そこで本研究では，ST コーパスの分割位置を学習し，音声から文境界を直接予測する音声分割モデルを提案する．

音声分割モデル

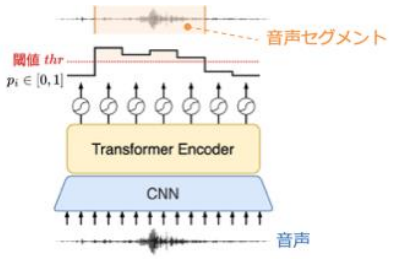


図 2. Transformer に基づく音声分割モデル

音声分割モデルの構成を図 2 に示す. モデルは CNN 層と Transformer Encoder [Vaswani et al., 2019] で構成され，音声入力に対するフレーム単位の系列ラベリング問題としてセグメント境界を予測する．固定長 T の音声を受け取り音声フレーム毎に確率pを出力し，事前に設定した閾値thrを上回るフレームの連続を音声セグメントとして扱う．

学習データ

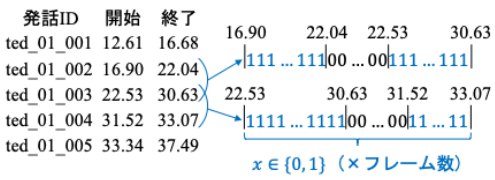


図 3. ST コーパスを用いたデータ作成

本研究では TED talks の ST コーパス MuST-C [Di Gangi et al., 2019] を使用する．学習データ作成方法を図 3 に示す．本手法では音声入力に対する系列ラベリング問題として文境界の予測を学習するため，連続する 2 つのセグメントを連結して音響特徴量の各フレームに対応するラベル$x \in \{0,1\}$を付与した．ラベル 0 と 1 はそれぞれ発話外・内に対応しており，開始・終了の時刻情報から作成する．

実験：長時間音声の英独・英日翻訳

MuST-C を用いて英独，英日翻訳の実験を行った．評価の際，TED talks の音声を音声分割手法によって分割し，分割したセグメントをそれぞれ ST で翻訳し，BLEU を測定した．ST には ASR と MT を組み合わせた Cascade ST と，直接音声をテキストに翻訳する End-to-end ST の 2 種類を用いた．これらのモデルはいずれも Transformer に基づく．本稿では End-to-end ST の結果を報告する．音声分割手法として以下の 5 つを比較した．

① 人手による文単位の分割（トップライン）
② VAD（ベースライン）
③ 固定長（ベースライン）：事前に設定した固定の長さで分割する単純なベースライン．セグメント長が一定に保たれる利点がある．長さは 4~40 秒の間で探索を行い 20 秒に設定した．
④ 音声分割モデル (提案手法) : T=20 秒に設定した．
⑤ 音声分割モデル ＋VAD（提案手法）：過剰分割を軽減するために④と②の合意で分割する手法．

	英独	英日
① 人手分割	22.50	10.60
② VAD	16.40	8.14
③ 固定長	17.96	8.52
④ 音声分割モデル	19.10	8.77
⑤ 音声分割モデル+VAD	**19.87**	**9.24**

表 1. BLEU による各音声分割手法の比較

表 1 に，各音声分割手法を用いた ST システムの翻訳精度を示す．提案手法（④）は，英独・英日翻訳においてベースライン（②③）を上回った．更に VAD と音声分割モデルを組み合わせた提案手法（⑤）は④

を上回った．一方で人手分割のトップライン（①）と比べると改善の余地があることも確認された．

3. Transformer に基づく同時 ST モデル [Fukuda et al., 2023]

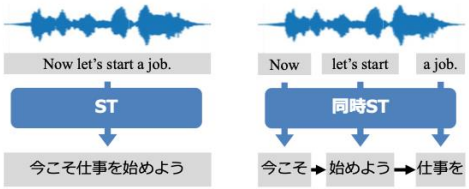


図 4. 同時 ST の概要

同時 ST は，リアルタイムに音声を翻訳することを目的とした MT 技術である [Dalvi et al., 2018]. 通常の ST の多くが文単位で翻訳を行うのに対し（図 4 左），同時 ST は文の終了を待たずに翻訳を開始する（図 4 右）．同時 ST において，音声を受け取り訳出するまでの遅延と翻訳精度にはトレードオフの関係がある．このトレードオフを改善し，低遅延で高い翻訳精度を達成しようとする取り組みが近年盛んに行われている [Ma et al., 2019]. 博士研究では, 長い音声を翻訳可能なストリーミング ST システムの構築に焦点を当てており，そうした同時 ST のトレードオフの改善には着目していない．しかし，ストリーミング ST システムの構成要素として，既存技術を組み合わせた高精度な同時 ST モデルを作成したので本章ではその内容を解説する．

同時 ST モデル

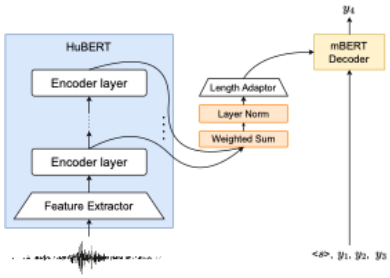


図 5. HuBERT と mBERT に基づく ST モデル

近年の ST は Transformer に基づくモデルが主流である. Transformer は自己注意機構を活用することで，長距離の依存関係を捉えることが可能であり，音声信号のような長い時系列データの特徴をモデリングすることに適している．また最近では，大量のデータで事前学習された大規模な音声エンコーダと言語モデルを組み合わせた ST モデルの構築が注目されている．代表的な音声エンコーダとして，wav2vec 2.0 [Baevski et al., 2020] や HuBERT [Hsu et al., 2021] が, 言語モデルとしては mBART [Devlin et al., 2019]が挙げられる．これらは Transformer に基づくモデルであり，事前学習によって音声や言語に関する知識を獲得している．これらのモデルを使用することで，従来の事前学習なしの Transformer に基づく ST モデルの精度を大きく上回ることが知られている．本研究では，HuBERT と mBERT を組み合わせた ST モデルを作成した（図 5）．また，この ST モデルを同時 ST に適用するために Prefix alignment [Kano et al., 2022] と Local Agreement [Liu et al., 2024] という 2 つの技術を採用した．Prefix alignment は，対訳文ペアから部分的な対訳ペア（Prefix pair）を抽出して ST モデルを学習させる方法である．例えば英独翻訳において「I have a pen. / Ich habe einen Stift.」という対訳からは,「I / Ich」「I have / Ich habe」のような Prefix pair を作成することができる．この手法は特に語順の近い言語対で有効である．一方，英日のような語順が異なる言語の場合,「I have / 私は」といった不均衡な Prefix pair がしばしば抽出される．実験では，こうした不均衡なペアが学習に及ぼす影響についても調査した．Local Agreement は，訳出を決定するためのアルゴリズムである．推論時，モデルは短い音声セグメントを受け取って翻訳文を生成するというステップを繰り返す．Local Agreement は連続するステップの翻訳文を比較し，先頭から一致する最長の部分文を訳出として確定していくことで，翻訳結果の信頼性を高める方法である．これらの手法の詳細については原著論文を参照していただきたい．

実験：英独・英日・英中の同時翻訳

MuST-C を用いて英独, 英日, 英中翻訳の実験を行った．翻訳精度を BLEU，訳出遅延を Average Lagging (AL) [Ma et al., 2019] で評価した．ここでは英独，英

日実験の結果を抜粋して説明する．異なるデータで学習された 3 つの ST モデルを比較した．

① 対訳文ペアのみで学習されたオフライン ST モデル（Offline）

② ①のモデルを同時 ST に対応させるため，Prefix Pair で追加学習したモデル（Offline+PA）

③ Prefix pair から目的言語テキストが原言語音声に比べて極端に長い不均衡なものを除外して②と同様に学習したモデル（Offline+PA+Filtering）

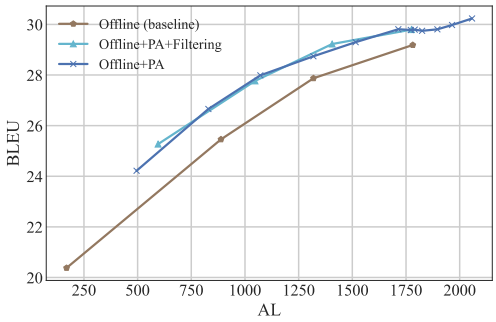


図 6. 英独の同時 ST

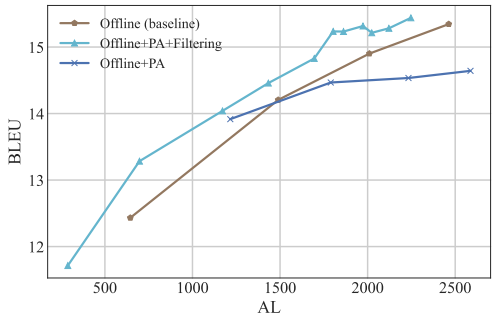


図 7. 英日の同時 ST

図 6, 7 はそれぞれ英独, 英日翻訳における遅延（AL）と精度（BLEU）のトレードオフを示す．英独において, オフライン ST（①）を Prefix Pair で追加学習（②）することで低遅延（AL<1000）から高遅延（AL>1000）まで一貫した精度の向上が得られた．語順が比較的近いペアであるため, Prefix pair のフィルタリング（③）では効果が得られなかった．一方，英日では②が①を下回り，③は①を上回った．このことから，英日のような語順の遠い言語対の ST に Prefix Alignment を適用する場合, 不均衡な Prefix pair を除外する必要があると考えられる．

4. ストリーミング ST システム

最後に音声分割モデルと同時 ST モデルを組み合わせて，長時間音声をリアルタイムに翻訳するストリーミング ST システムを構築した．音声分割モデルは，2 章で紹介した Transformer に基づくモデルを改良した，wav2vec 2.0 に基づくモデル［Fukuda et al., 2024］を用いた．同時 ST モデルは 3 章で紹介したモデルを用いた．

システムの動作を図 1 に示す．初めに音声分割モデルは固定長（e.g. 400 ms）の音声セグメントを受け取り，文の開始・終了位置を検出する．続いて，ある文単位セグメントの開始位置と終了位置に挟まれた部分のみが同時 ST モデルに入力され, 同時 ST モデルは翻訳結果を出力する．このように，音声分割と同時 ST が交互に動作する仕組みである．一つの文単位セグメントの処理が終了すると，同時 ST が保持する過去の文脈情報がリセットされる．これにより，システムはメモリエラー等を回避して無限長の音声を翻訳することができる．

実験：長時間音声の英独同時翻訳

MuST-C を用いて英独翻訳の実験を行った．固定長の音声セグメントの長さを 400~1200 ms の間で変化させ，遅延を調節した．音声分割手法として以下の 5 つを比較した．③〜⑤が提案手法の音声分割モデルであり，①と②はトップラインとベースラインである．各手法の詳細は原著論文を参照していただきたい．

① 人手による文単位の分割（Topline）

② 固定長（Fixed-length）：事前に設定した固定の長さで分割する単純なベースライン．非同時 ST において VAD より高い翻訳精度を示した（2 章）．長さは 10~20 秒の間で探索を行い 15 秒に設定した．

③ 音声分割モデル（Unmasked）：長い音声セグメント（20 秒）を入力として学習した非同時 ST 用の音声分割モデル

④ 音声分割モデル（Monotonic）：③は学習時に最大 20 秒先の音声情報まで参照できるが，短い音声セグメントを入力とする同時 ST の推論とギャップがある．このギャップを低減するため，過去の音声情

報のみ参照できる制約下で学習したモデル

⑤ 音声分割モデル（Chunk-wise）：⑥の制約を緩和し，短い音声セグメントに含まれる少し先の音声情報（最大 1200ms まで）を参照できる条件で学習したモデル

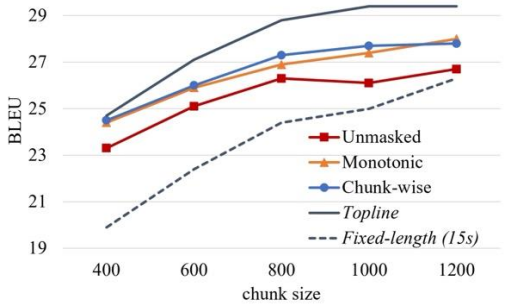


図 8. 英独のストリーミング ST

図 8 に，各音声分割手法を用いたストリーミング ST システムの遅延と精度のトレードオフを示す．ベースラインは，トップラインの 81-90%の翻訳精度を維持した．3 つの提案モデルはベースラインを上回っており，特に Chunk-wise はトップラインの 94-99%の翻訳精度を維持した．このことから，音声分割モデルは同時 ST においても高い翻訳精度を達成できたといえる．また，非同時 ST 用のモデル（Unmasked）は，Monotonic と Chunk-wise を下回った．Monotonic と Chunk-wise の間に有意な差は認められなかった．

5. まとめと今後の課題

本研究では，長い音声ストリームをリアルタイムで翻訳するための実用上の課題に取り組み，提案技術を組み込んだストリーミング音声翻訳システムを作成した．第一に，Cascade ST システムにおける音声認識誤りの伝播の緩和に取り組み，頑健な MT モデルを構築する手法を提案した．第二に，音声翻訳のための音声分割の精度改善に取り組んだ．音声を文単位セグメントに直接分割するモデルを提案し，長時間音声の翻訳精度を向上させた．第三に，音声分割モデルと同時 ST モデルを組み合わせることでストリーミング ST システムを構築し，精度を検証した．実験では，提案した音声分割モデルを使用してトップラインの翻訳精度を 94%以上維持できることを示した．これらの取り組みを通じて ST システムの発展に貢献できたと考えている．

一方で課題も多く残されている．ストリーミング ST の評価では実処理時間やモデルサイズを考慮しなかった．実用性を向上させるために，モデルの軽量化や高速化への取り組みが必要である．また，今回実験データとして比較的処理しやすい音声を使用した．背景雑音や残響を含む様々な環境下における処理や，言い淀みやフィラー独話などを多く含む自然発話の処理も重要な今後の課題である．

参考文献

[Fukuda, 2024] Ryo Fukuda, “Towards Streaming Speech Translation for Real-world Scenarios”, Thesis, Doctor (Engineering), Nara Institute of Science and Technology, 2024.

[Zhang et al., 2004] Ruiqiang Zhang, Genichiro Kikui, Hirofumi Yamamoto, Frank K Soong, Taro Watanabe, and Wai-Kit Lo, “A unified approach in speech-to-speech translation: integrating features of speech recognition and machine translation”, In Proc. COLING 2004, 2004.

[Hinton et al., 2015] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, “Distilling the knowledge in a neural network”, arXiv preprint arXiv:1503.02531, 2015.

[Fukuda et al., 2021] Ryo Fukuda, Katsuhito Sudoh and Satoshi Nakamura, “On Knowledge Distillation for Translating Erroneous Speech Transcriptions”, In Proc. IWSLT 2021, 2021.

[Hughes and Mierle, 2013] Thad Hughes and Keir Mierle, “Recurrent neural networks for voice activity detection”, In 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013.

[Wan et al., 2021] David Wan, Chris Kedzie, Faisal Ladhak, Elsbeth Turcan, Petra Galuščáková, Elena Zotkina, Zheng Ping Jiang, Peter Bell, and Kathleen

McKeown, “Seg- menting subtitles for correcting asr segmentation errors”, In Proc. EACL 2021, 2021.

[Fukuda et al., 2022] Ryo Fukuda, Katsuhito Sudoh and Satoshi Nakamura, “Speech Segmentation Optimization using Segmented Bilingual Speech Corpus for End-to-end Speech Translation”, In Proc. Interspeech 2022, 2022.

[Gaido et al., 2021] Marco Gaido, Matteo Negri, Mauro Cettolo, and Marco Turchi, “Beyond voice activity detection: Hybrid audio segmentation for direct speech translation”, CoRR, abs/2104.11710, 2021.

[Vaswani et al., 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, “Attention is all you need”, In Proc. NeurIPS 2017, 2017.

[Di Gangi et al., 2019] Mattia A.Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi, “MuST-C: a Multilingual Speech Translation Corpus. In Proc. NAACL-HLT 2019, 2019.

[Dalvi et al., 2018] Fahim Dalvi, Nadir Durrani, Hassan Sajjad, and Stephan Vogel, “Incremental decoding and training methods for simultaneous translation in neural machine translation”, In Proc. NAACL-HLT 2018, 2018.

[Ma et al., 2019] Mingbo Ma, Liang Huang, Hao Xiong, Renjie Zheng, Kaibo Liu, Baigong Zheng, Chuanqiang Zhang, Zhongjun He, Hairong Liu, Xing Li, Hua Wu, and Haifeng Wang, “STACL: Simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework”, In Proc. ACL 2019, 2019.

[Fukuda et al., 2023] Ryo Fukuda, Yuta Nishikawa, Yasumasa Kano, Yuka Ko, Tomoya Yanagita, Kosuke Doi, Mana Makinae, Sakriani Sakti, Katsuhito Sudoh and Satoshi Nakamura, “NAIST Simultaneous Speech Translation System for IWSLT 2023”, In Proc. IWSLT 2023, 2023.

[Baevski et al., 2020] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations”, In Proc. NeurIPS 2020, 2020.

[Hsu et al., 2021] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed, “Hubert: Self-supervised speech representation learning by masked prediction of hidden units”, In IEEE TASLP, 2021.

[Devlin et al., 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding”, In Proc. NAACL-HLT 2019, 2019.

[Kano et al., 2022] Yasumasa Kano, Katsuhito Sudoh, and Satoshi Nakamura, “Simultaneous neural machine translation with prefix alignment”, In Proc. IWSLT 2022, 2022.

[Liu et al., 2024] Danni Liu, Gerasimos Spanakis, and Jan Niehues, “Low-Latency Sequence-to- Sequence Speech Recognition and Translation by Partial Hypothesis Selection”, In Proc. Interspeech 2020, 2020.

[Fukuda et al., 2024] Ryo Fukuda, Katsuhito Sudoh and Satoshi Nakamura, “Improving Speech Translation Accuracy and Time Efficiency with Fine-tuned wav2vec 2.0-based Speech Segmentation”, In IEEE TASLP, 2024.

第 8 回 AAMT セミナー

第 1 回 AAMT 若手翻訳研究会 開催報告

中澤 敏明
東京大学

1. はじめに

AAMT では、機械翻訳の発展・啓蒙の一環として、MT に関するトピックの情報配信を目的とし定期的にセミナーを開催している。これまでは毎回、講師の先生をお招きし、機械翻訳の技術やその応用についての発表や研究者による最新情報、提供者側の課題など MT に関する多彩なトピックスをお届けしてきた。2024 年 3 月 22 日開催の第 8 回の AAMT セミナーでは新たな試みとして、若手の翻訳・通訳・機械翻訳研究者による最新の研究成果の発表の場を設け、研究者及び利用者間のコミュニケーションの促進を図るために第 1 回 AAMT 若手翻訳研究会を開催した。

研究会では全部で 13 件の発表があった。機械翻訳の技術的な内容に関する研究発表だけでなく、通訳や翻訳そのものに関する研究発表や、ポストエディットの実践に関する発表などもあり、発表内容は多岐に渡った。聴講者も 200 名弱に達し、翻訳に携わる様々な人の交流の場としてとても良いものになったと思う。

2. 表彰

研究会では、特に優秀であると認められた発表に関しては表彰し、副賞を進呈した。当初は 10 件程度の発表申し込みに対して、最優秀賞 1 件、優秀賞 3 件を選定する予定であったが、発表申し込み件数が 13 件と想定よりも多くなったため、優秀賞を 1 件増やして 4 件とした。選定方法としては聴講者による投票結果をもとに、AAMT セミナー委員での協議により決定した。選定結果は以下の通りである。

・最優秀賞

サブセット探索を用いた高速な kNN ニューラル機械翻訳

出口祥之（NAIST/NICT)、渡辺太郎（NAIST)、松井勇佑（東京大学)、内山将夫（NICT)、田中英輝(NICT)、隅田英一郎（NICT)

・優秀賞

日英間の機械翻訳による受容化と異質化について
木内晶基（東京工業大学)

人手翻訳から MTPE へ: 一翻訳者の所感
海老原仁美（レッドハット株式会社)

キャラクターの性格と人間関係情報を付加した映像翻訳データセットの構築
大嶽匡俊（東京大学)、加藤大地（東京大学)、野崎優斗(東京大学)、廣岡聖司（東京大学)、宮尾祐介（東京大学)、金崎朝子（東京工業大学)

大規模言語モデルに対する対訳データを用いた継続事前訓練による翻訳精度評価
近藤海夏斗（筑波大学)、宇津呂武仁（筑波大学)、森下睦（NTT)、永田昌明（NTT)

3. 終わりに

本稿では第 1 回 AAMT 若手翻訳研究会の開催報告を行った。各発表の発表概要や、一部発表の発表資料および講演の録画は AAMT セミナーのウェブサイト

Report of 1st AAMT Young Translation Research Workshop
Toshiaki Nakazawa
University of Tokyo

にて確認できる[1]。

今後、AAMT 若手翻訳研究会は年に 1 回定期的に開催することとし、第 2 回の研究会は 2025 年 3 月に開催する予定としている。第 2 回研究会では例えばスポンサーを募り、スポンサー賞を設置するなど新たな試みも実施したいと考えている。

[1] https://aamt.info/event/seminar/20240322

第 1 回 AAMT 若手翻訳研究会最優秀賞受賞記念

サブセット探索を用いた効率的なk近傍機械翻訳

出口 祥之
奈良先端科学技術大学院大学 / 情報通信研究機構

1. はじめに

ニューラル機械翻訳（Neural Machine Translation; NMT）は，訓練コーパスに十分な量のデータが含まれていないドメイン（遠ドメイン）の翻訳精度が低い．特に，医療分野のような専門文書の翻訳を担う産業翻訳では，ドメイン特有の用語やスタイル等を正確に翻訳する必要があり，既存の汎用翻訳エンジンを活用しながら，対象ドメインに効率的に適応する手法が求められる．

近年，NMT を再訓練するコストを抑えつつ遠ドメインの翻訳精度を改善する手法として，用例ベース手法を組み込んだ NMT（用例ベース NMT）が提案されている．用例ベース NMT は，Nagao によって提案された類推に基づく機械翻訳 [1]を NMT に拡張したモデルとみなせる．中でも，k近傍機械翻訳（k Nearest Neighbor Machine Translation; kNN-MT） [2] は，対象ドメインの翻訳用例を参照することにより，遠ドメインの翻訳精度を改善した．kNN-MT は既存の翻訳モデルを追加訓練することなくそのまま利用でき，さらにドメイン適応翻訳タスクにおいて最高性能を達成したことから，注目を浴びている．しかし，一単語出力するたびに用例を検索するため，一文あたりの翻訳速度が通常の NMT と比較して 100～1,000 倍程度低下するという問題がある．

本研究では，kNN-MT の翻訳速度の高速化を狙い，入力文の情報を利用して関連する事例を粗く絞り込む，サブセット探索を提案する．また，サブセット探索に適した，ルックアップテーブルを用いたベクトル間距離計算法を採用し，各翻訳事例との距離を効率的に計算することで，さらなる高速化を目指す．WMT'19 独英翻訳と複数のドメイン適応翻訳実験を行った結果，従来法と比較して，提案法は，翻訳精度が最大で 1.6BLEU%，翻訳速度が最大で 132 倍改善することを確認した．

2. k近傍機械翻訳

ニューラル機械翻訳　一般的な NMT は，原言語文 $\boldsymbol{x} = (x_1, \ldots, x_{|\boldsymbol{x}|}) \in \mathcal{V}_X^*$ が与えられたとき，目的言語文 $\boldsymbol{y} = (y_1, \ldots, y_{|\boldsymbol{y}|}) \in \mathcal{V}_Y^*$ を先頭から順に生成するように訓練される．ただし，$\mathcal{V}_X^*, \mathcal{V}_Y^*$ は，それぞれ原言語と目的言語の語彙のクリーネ閉包を表す．時刻tに出力されるトークン y_t は，原言語文 $\boldsymbol{x}$ と時刻tまでに生成した目的言語トークン系列 $\boldsymbol{y}_{<t}$ から計算される確率 $p(y_t|\boldsymbol{x}, \boldsymbol{y}_{<t})$ に基づいて生成される．

データストア構築　kNN-MT は，予め，翻訳用例を検索可能なデータ構造（データストア）に格納する．データストアはD次元ベクトルと目的言語トークンの対からなるキー・値ストア $\mathcal{M} \subseteq \mathbb{R}^D \times \mathcal{V}_Y$ で表現される．キーベクトルは訓練済み NMT モデルに教師強制で対訳文対を入力したときのデコーダ最終層の中間表現，値はキーベクトルから出力されるべき正解トークンであり，データストアは次式により定式化される．

$$\mathcal{M} = \bigcup_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{D}} \{(f(\boldsymbol{x}, \boldsymbol{y}_{<t}), y_t) | 1 \leq t \leq |\boldsymbol{y}|\}, \qquad (1)$$

なお，$\mathcal{D} \subseteq \mathcal{V}_X^* \times \mathcal{V}_Y^*$ は対訳コーパスを表し，$f: \mathcal{V}_X^* \times \mathcal{V}_Y^* \to \mathbb{R}^D$ は中間表現ベクトルを計算する NMT モデルを表す．本研究では Khandelwal ら [2]に従い，Transformer デコーダ最終層の順伝播層入力をキーベクトルに用いた．

翻訳　kNN-MT は，出力トークンの近傍事例を検索し，NMT の出力確率を補正する．具体的には，翻訳

Efficient Nearest Neighbor Machine Translation using Subset Retrieval
Hiroyuki Deguchi
Nara Institute of Science and Technology / National Institute of Information and Communications Technology

中の各時刻で計算されるキーベクトルと同じ中間表現ベクトルを検索クエリとし，データストアからk近傍トークンを探索する．探索したk近傍事例を用いて次のようなk近傍確率$p_{k\mathrm{NN}}$を計算する．

$$p_{kNN}(y_t|\boldsymbol{x},\boldsymbol{y}_{<t}) \propto \sum_{(\boldsymbol{k},v)\in\mathcal{N}_{\mathcal{M}}^{k}(f(\boldsymbol{x},\boldsymbol{y}_{<t}))} \mathbb{1}_{y_t=v}\exp\frac{-\|f(\boldsymbol{x},\boldsymbol{y}_{<t})-\boldsymbol{k}\|_2^2}{\tau}, \quad (2)$$

ただし，$\mathcal{N}_{\mathcal{M}}^{k}\colon \mathbb{R}^D \to (\mathbb{R}^D\times\mathcal{V}_Y)^k$は，与えられたクエリに対してキーベクトルとのユークリッド距離が最も近いk個の事例集合をデータストア$\mathcal{M}$から探索する関数を表し，$\mathbb{1}$は指示関数を表す．τはk近傍確率の分布の滑らかさを制御する温度パラメータである．求めたk近傍確率$p_{k\mathrm{NN}}$と NMT の予測確率p_{MT}を線形補間することでkNN-MT の出力確率を求める．

$$P(y_t|\boldsymbol{x},\boldsymbol{y}_{<t}) = \lambda p_{k\mathrm{NN}}(y_t|\boldsymbol{x},\boldsymbol{y}_{<t}) + (1-\lambda)p_{\mathrm{MT}}(y_t|\boldsymbol{x},\boldsymbol{y}_{<t}). \quad (3)$$

$\lambda\in[0,1]$はk近傍確率の重みを制御するハイパーパラメータである．

kNN-MT は，通常の NMT モデルと比較して，翻訳速度が非常に遅い [2]．これは，対訳コーパスの全目的言語トークンに対する大規模な近傍探索を，各トークンを出力するたびに計算するためである（目的言語の語彙が対象でない点に注意されたい）．すなわち，一回の近傍探索の時間計算量を$\mathcal{O}(|\mathcal{M}|)$とすると，文$\boldsymbol{y}$を生成するための時間計算量は$\mathcal{O}(|\mathcal{M}||\boldsymbol{y}|D)$となる．なお，$|\mathcal{M}|$はしばしば 10 億規模の大きさになる．

直積量子化 データストアの大きさ$|\mathcal{M}|$は対訳コーパスの全目的言語トークン数に等しいため，しばしば 10 億規模の大きさとなりうる．また，NMT モデルに Transformer big モデル [3]を用いると，$D=1024$となる．このとき，実数ベクトルを 32 bit 浮動小数点配列で表現すると，データストアの大きさは 3.7TiB となる．そのため，主記憶装置に読み込めるように，直積量子化（Product Quantization; PQ） [4]を用いてキーベクトルを圧縮する．PQ はD次元ベクトルを$\frac{D}{M}$次元ずつM個のサブベクトルに分割し，それぞれの部分空間において量子化する．量子化のためのコードブックは$\frac{D}{M}$次元部分空間ごとに学習される．m番目の部分空間のコードブック$\mathcal{C}^m=\{\boldsymbol{c}_1^m,\dots,\boldsymbol{c}_L^m\}\subset\mathbb{R}^{\frac{D}{M}}$は，データストア内のキーベクトルに対して$k$-means クラスタリングを実行して得られた$L$個のクラスタ重心集合である．なお，本研究では$L=256$とする．PQ により，ベクトル$\boldsymbol{q}\in\mathbb{R}^D$は次のようなコードベクトル$\bar{\boldsymbol{q}}\in\{1,\dots,L\}^M$に量子化される．

$$\bar{\boldsymbol{q}} = [\bar{q}^1,\dots,\bar{q}^M]^\top, \quad (4)$$

$$\bar{q}^m = \mathrm{argmin}_l\|\boldsymbol{q}^m-\boldsymbol{c}_l^m\|_2^2, \quad (5)$$

ただし，$\boldsymbol{q}=\left[\boldsymbol{q}^{1\top},\dots,\boldsymbol{q}^{M\top}\right]^\top,\boldsymbol{q}^m\in\mathbb{R}^{\frac{D}{M}}$である．先ほどのデータストアの例において，$M=64$とすると，60GiB 程度まで圧縮され，主記憶装置に読み込めるようになる．

3. 提案法：サブセットk近傍機械翻訳

本研究では，kNN-MT の翻訳速度を改善するため，サブセット探索を用いたk近傍機械翻訳（サブセットkNN-MT）を提案する．提案法は，入力文の情報を用い，翻訳開始時に近傍探索の探索空間を大幅に削減する．さらに，絞り込んだ事例からトークン単位のk近傍事例を探索する際，より効率的なアルゴリズムを用いてクエリとキーの間の距離計算を高速化する．

3.1 サブセット探索

文データストア構築 データストアを拡張した文データストア$\mathcal{S}\subseteq\mathbb{R}^{D'}\times 2^{\mathcal{M}}$を次式に従い構築する．

$$\mathcal{S}=\left\{\left(s(\boldsymbol{x}^{(i)}),\mathcal{M}^{(i)}\right)\right\}_{i=1}^{|\mathcal{D}|}, \quad (6)$$

$$\mathcal{M}^{(i)}=\left\{\left(f\left(\boldsymbol{x}^{(i)},\boldsymbol{y}_{<t}^{(i)}\right),y_t^{(i)}\right)\middle|1\le t\le|\boldsymbol{y}^{(i)}|\right\}, \quad (7)$$

ただし，$s\colon\mathcal{V}_X^*\to\mathbb{R}^{D'}$は原言語文の$D'$次元文ベクトル表現を計算する文符号化器を表し，$\left(\boldsymbol{x}^{(i)},\boldsymbol{y}^{(i)}\right)\in\mathcal{D}$は対訳コーパスの$i$番目の対訳文対を表す．すなわち，$\mathcal{M}^{(i)}\subset\mathcal{M}$は$i$番目の対訳文対のみから構築された$k$NN-MT データストアである．

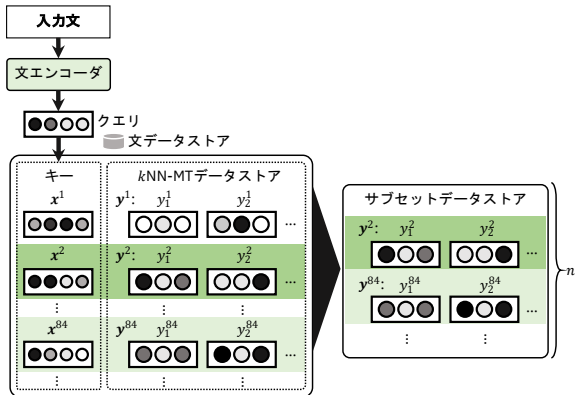


図 1: サブセット探索.

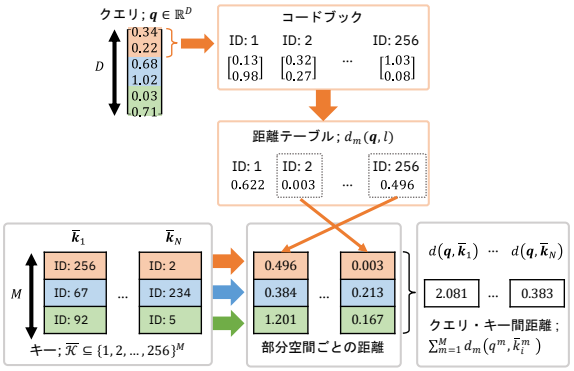


図 2: ADC によるクエリ・キーベクトル間の距離計算.

翻訳 翻訳開始時，入力文の類似文上位n ($\ll |\mathcal{D}|$)件を文データストアから検索し，k近傍探索対象の事例をn文に絞り込む．

$$\widehat{\mathcal{M}}_{\boldsymbol{x}}^{n} = \bigcup_{(s,\mathcal{M}^{(i)}) \in \mathcal{N}_{\mathcal{S}}^{n}(s(\boldsymbol{x}))} \mathcal{M}^{(i)}, \tag{8}$$

ただし，$\widehat{\mathcal{M}}_{\boldsymbol{x}}^{n} \subset \mathcal{M}$は入力文$\boldsymbol{x}$の類似文上位$n$件のみからなるデータストア$\mathcal{M}$のサブセット（サブセットデータストア）であり，$\mathcal{N}_{\mathcal{S}}^{n}: \mathbb{R}^{D'} \to \left(\mathbb{R}^{D'} \times 2^{\mathcal{M}}\right)^{n}$は，文データストア$\mathcal{S}$から入力文の類似文上位$n$件を検索する探索関数を表す．翻訳中，近傍探索に用いるデータストアは$\mathcal{M}$の代わりに$\widehat{\mathcal{M}}_{\boldsymbol{x}}^{n}$を用いる．これにより，探索対象の文の数は$|\mathcal{D}|$から$n \ll |\mathcal{D}|$へと大幅に削減される．なお，提案法は類似文検索の計算が追加されるが，対訳コーパスの文数はトークン数よりも小さく，かつ，文検索は翻訳開始時に一度だけ実行すればよいため，速度が大幅に低下することはない．

入力文$\boldsymbol{x}$の類似文の平均文長を$\overline{|\boldsymbol{x}|}$とすると，サブセット探索を用いることで，時間計算量は$\mathcal{O}(|\mathcal{M}||\boldsymbol{y}|D)$から$\mathcal{O}\left(|\mathcal{D}|D' + n\overline{|\boldsymbol{x}|}|\boldsymbol{y}|D\right)$へと削減される．

3.2 ルックアップテーブルを用いた距離計算

提案法では，クエリとサブセット内の各キーとの間の距離を効率的に計算するため，図に示すルックアップテーブル（Look-Up Table; LUT）を用いた距離計算法 asymmetric distance computation (ADC) [4]を採用する．

翻訳中の各時刻において，k近傍事例を探索するため，クエリベクトル$\boldsymbol{q} \in \mathbb{R}^D$とサブセットデータストア内の量子化キーベクトル$\{\bar{\boldsymbol{k}}_i\}_{i=1}^{n|\overline{\boldsymbol{x}}|}$との間の二乗ユークリッド距離を計算する必要がある．このとき，通常の距離計算法，すなわち，$\|\boldsymbol{q} - \boldsymbol{k}_i\|_2^2$に従って距離を計算すると，一トークン出力するために必要な時間計算量は$\mathcal{O}\left(n|\overline{\boldsymbol{x}}|D\right)$となる．ADC は，サブセット内のすべてのキーベクトルに対して計算する代わりに，予めコードブック内の代表ベクトルと距離計算しておくことで，トークン出力の時間計算量を$\mathcal{O}(LD)$まで削減する．

ADC は，はじめに，クエリ$\boldsymbol{q} \in \mathbb{R}^D$の各$\frac{D}{M}$次元サブベクトル$\boldsymbol{q}^m \in \mathbb{R}^{\frac{D}{M}}$とコードブック$\mathcal{C}^m \subset \mathbb{R}^{\frac{D}{M}}$内の代表ベクトルとの二乗ユークリッド距離を計算し，距離 LUT $\boldsymbol{A}^m \in \mathbb{R}^L$に計算結果を保存する．

$$A_l^m = \|\boldsymbol{q}^m - \boldsymbol{c}_l^m\|_2^2. \tag{9}$$

続いて，サブセット内の各キーベクトルの量子化コード自身を参照キーとし，LUT から計算済みの距離を参照する．クエリと量子化されたキー$\bar{\boldsymbol{k}} \in \{1, \dots, L\}^M$との間の二乗距離を$d^2\left(\boldsymbol{q}, \bar{\boldsymbol{k}}\right)$とすると，次式のようになる．

$$d^2\left(\boldsymbol{q}, \bar{\boldsymbol{k}}\right) = \sum_{m=1}^{M} d_m^2\left(\boldsymbol{q}^m, \bar{k}^m\right) = \sum_{m=1}^{M} A_{\bar{k}^m}^m. \tag{10}$$

ADC により，各キーベクトルを復号することなく，LUT 構築に要する時間$\mathcal{O}(LD)$と距離参照に要する定数時間$\mathcal{O}(1)$でクエリ・キー間距離が得られる．また，ADC は，事前に索引を構築するような他の探索アルゴリズムと異なり，探索対象が入力に応じて動的に変化するサブセットkNN-MT に対して適した手法となっている．

サブセット探索と ADC を併用することにより，時間計算量は$\mathcal{O}(|\mathcal{D}|D' + L|\boldsymbol{y}|D)$まで削減される．

4. 実験

本節では，従来法のkNN-MT と提案法のサブセットkNN-MT の翻訳品質と翻訳速度を比較するため，翻訳モデルの訓練コーパスと近いドメイン（近ドメイン）および訓練コーパスと遠いドメイン（遠ドメイン）において翻訳実験を行った．

4.1 実験設定

翻訳 翻訳品質は sacreBLEU(%) [5]と COMET[1](%) [6]によって，翻訳速度は一秒間に生成するトークン数（トークン毎秒; tok/s）によって評価した．すべての実験で NVIDIA V100 GPU を 1 基使用した．速度評価において，バッチサイズは，オフライン翻訳を想定した 12,000 トークン（B_∞），および，オンライン翻訳を想定した 1 文（B_1）の 2 つの設定を比較した．翻訳文はビーム探索により生成し，ビーム幅は 5，文長正規化パラメータは 1.0 とした．

***k*近傍探索** 探索する近傍事例数は$k = 16$，$p_{k\mathrm{NN}}$の温度パラメータは$\tau = 100$に設定した．$p_{k\mathrm{NN}}$の計算におけるクエリ・キー間距離は量子化ベクトルから算出される近似距離を用いた．kNN-MT のトークン探索とサブセットkNN-MT の類似文探索には Faiss [7]を用い，optimized PQ [8]と IVFPQ [4]によってインデックスを構築した．サブセットkNN-MT のデータストアは主成分分析によって 1,024 次元から 256 次元に次元削減した後に PQ によって量子化した．すべての手法において PQ のサブベクトル数は$M = 64$とした．

文符号化器 文符号化器は大きく分けてニューラルモデルと非ニューラルモデルを採用し，それぞれ性能を比較した．ニューラルモデルには訓練済み多言語文符号化器 LaBSE と，NMT モデル自身のエンコーダ中間表現の平均ベクトル AvgEnc を採用した．非ニューラルモデルには，TF-IDF と BM25 を採用し，各トークンの重みを算出した後，L2 正規化と特異値分解を適用し， 256 次元ベクトル表現を獲得した．

表 1: 近ドメイン（WMT'19 独英）翻訳の実験結果．

			↑tok/s	
モデル	BLEU	COMET	B_∞	B_1
Base MT	39.2	<u>84.6</u>	**6375.2**	**129.1**
***k*NN-MT**	<u>40.1</u>	**84.7**	19.6	2.5
C***k*NN-MT**	39.5	84.3	74.6	22.3
F***k*NN-MT**	**40.3**	**84.7**	286.9	27.1
サブセット***k*NN-MT (ours)**				
s: LaBSE	<u>40.1</u>	**84.7**	2191.4	<u>118.4</u>
s: AvgEnc	39.9	**84.7**	1816.8	97.3
s: TF-IDF	40.0	<u>84.6</u>	<u>2199.1</u>	113.0
s: BM25	40.0	<u>84.6</u>	1903.9	108.4

4.2 近ドメイン翻訳

WMT'19 独英翻訳タスクを用い，翻訳品質と翻訳速度を評価した．提案法の有効性を検証するため，通常の Transformer MT [3] (Base MT)，kNN-MT [2]，Chunk-based k NN-MT (C k NN-MT) [9]，Fast kNN-MT [10] (FkNN-MT)と比較した．

データストアの構築には WMT'19 独英翻訳の対訳コーパスを用い，サブワード化後の文長が 250 以下かつ対訳文の文長比が 1.5 以内の 29M 文から得られた 862M トークンから構築した．kNN 確率の重みは$\lambda = 0.3$とした．FkNN-MT の原言語側探索は近傍 512 トークンを探索し，CkNN-MT のチャンク数は 16 に設定した．提案法で探索する近傍文数は$n = 512$文とした．

実験結果を表 1 に示す．なお，評価指標ごとの最高スコアを太字で，2 番目に高いスコアを下線で記した．表より，kNN-MT は追加学習なしで Base MT より翻訳精度を 0.9 BLEU%改善しているが，翻訳速度は B_∞ において 325 倍低下している．一方で，提案法のサブセットkNN-MT は，近傍文探索に LaBSE を用いた際，kNN-MT の翻訳精度を維持しつつ，kNN-MT の翻訳

[1] Unbabel/wmt22-comet-da

表 2: 遠ドメイン翻訳の翻訳品質と翻訳速度.

	IT		コーラン		法		医療		字幕	
モデル	BLEU	tok/s	BLEU	tok/s	BLEU	tok/s	BLEU	tok/s	BLEU	tok/s
Base MT	38.7	**4433.2**	17.1	**5295.0**	46.1	**4294.0**	42.1	**4392.1**	29.4	**6310.5**
kNN-MT	41.0	22.3	19.5	19.3	52.6	18.6	48.2	19.8	29.6	30.3
サブセット***k***NN-MT										
s: LaBSE	**41.9**	2362.2	**20.1**	2551.3	**53.6**	2258.0	**49.8**	2328.3	29.9	3058.4
s: AvgEnc	**41.9**	2197.8	19.9	2318.4	53.2	1878.8	49.2	2059.9	**30.0**	3113.0
s: TF-IDF	40.0	2289.0	19.3	2489.5	51.4	2264.3	47.5	2326.6	29.3	2574.4
s: BM25	40.0	1582.4	19.1	2089.5	50.8	1946.3	47.4	1835.6	29.4	1567.7

速度を 112 倍（B_∞）/ 47 倍（B_1）改善することを確認した．また，他の先行研究と比較しても，提案法は 8 倍から 29 倍程度（B_∞）高速に翻訳できることを確認した．

4.3 遠ドメイン翻訳

遠ドメインにおける翻訳性能を評価するため，ドメイン適応独英翻訳タスク [11]を用いて翻訳品質と翻訳速度を評価した．具体的には，汎用な対訳コーパスで訓練された翻訳モデルを用い，IT，コーラン，法，医療，字幕の 5 つのドメインの翻訳実験を実施した．データストアは，汎用な対訳コーパスと 5 つの対象ドメインの対訳コーパスをすべて混ぜ，896M トークンからなる対訳コーパスから構築した．確率分布の補間パラメータは$\lambda = 0.5$に設定した．提案法で探索する近傍文数は$n = 256$文に設定した．翻訳品質は BLEU，翻訳速度は B_∞で評価した．

実験結果を表 2 に示す．表より，*k*NN-MT の翻訳品質の改善幅は，近ドメイン翻訳よりも遠ドメイン翻訳のほうが大きいことがわかる．しかし，近ドメイン翻訳と同様に，速度が 2 桁低下している．一方で，サブセット*k*NN-MT は, 遠ドメイン翻訳においては従来の*k*NN-MT よりもさらに翻訳品質を改善し，かつ，Base MT の 40～50%程度の速度，最大で*k*NN-MT の 132 倍の速度で生成することを確認した．

5. 考察

4.3 節において，提案法は従来法よりも探索対象が削減されているにもかかわらず, 翻訳品質が改善した．本節では，実際の翻訳例を用いて，提案法により翻訳品質が改善した理由を考察する．

*k*NN-MT とサブセット*k*NN-MT の医療ドメインの翻訳例を表 4 に示す．参照訳より，このドメインでは「Co-administration」を訳出すべきであるが，Base MT および*k*NN-MT は「A joint use」と翻訳している．一方で, サブセット*k*NN-MT は,「Co-administration」を正しく訳出している．この事例のサブセット探索の結果を表 3 に示す．表中の「S-1/S-2/S-3」は，それぞれ，入力文の近傍文の検索結果上位 1 位/2 位/3 位を，「T-1/T-2/T-3」は，それぞれ，S-1/S-2/S-3 の対訳文，すなわち, サブセット*k*NN-MT の探索対象となる目的言語文を示す．表 3 より，上位 3 件中 T-1 と T-3 には訳出されたい「Co-administration」が含まれている．また，実際の翻訳時は，近傍上位 256 文を探索対象としたが，「Co-administration」はサブセット中に 30 件含まれていた．さらに，「A joint use」の「joint」はサブセット中に 1 件も含まれていなかった．このことから，サブセット*k*NN-MT は，探索対象を近傍文のみに絞ることにより，ノイズとなりうる事例を探索対象から除き，ドメイン特有の用語やスタイルを訳出しやすくなったことにより，遠ドメイン翻訳における翻訳品質が改善したと考えられる．

表 4: 医療ドメインにおける***k***NN-MT とサブセット***k***NN-MT の翻訳例の比較.

入力文	Eine gemeinsame Anwendung von Nifedipin und Rifampicin ist daher kontraindiziert.
参照訳	Co-administration of nifedipine with rifampicin is therefore contra-indicated.
Base MT	A joint use of nifedipine and rifampicin is therefore contraindicated.
kNN-MT	A joint use of nifedipine and rifampicin is therefore contraindicated.
サブセット***k***NN-MT	Co-administration of nifedipine and rifampicin is therefore contraindicated.

表 3: 表 4 の事例におけるサブセット探索の検索結果上位 3 件.

S-1	Die gemeinsame Anwendung von Ciprofloxacin und Tizanidin ist kontraindiziert.
S-2	Rifampicin und Nilotinib sollten nicht gleichzeitig angewendet werden.
S-3	Die gleichzeitige Anwendung von Ribavirin und Didanosin wird nicht empfohlen.
T-1	Co-administration of ciprofloxacin and tizanidine is contra-indicated.
T-2	Rifampicin and nilotinib should not be used concomitantly.
T-3	Co-administration of ribavirin and didanosine is not recommended.

6. おわりに

本研究では，用例検索を組み込んだニューラル機械翻訳*k*NN-MT の翻訳速度を 100 倍以上改善するサブセット*k*NN-MT を提案した．提案法は，検索対象を入力文の近傍に絞り込み，さらにルックアップテーブルを用いた効率的な距離計算法を採用することで，翻訳速度を改善した．また，遠ドメイン翻訳実験においては，翻訳速度を改善するだけでなく，翻訳品質まで改善することを確認した．

今後は，提案法を拡張し，音声翻訳モデルなどに向けたマルチモーダル化や，大規模言語モデルを用いた翻訳を含むテキスト生成タスクへの応用を検討したい．

謝辞

本稿は，第 1 回 AAMT 若手翻訳研究会最優秀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した．選考に関わってくださった皆様に感謝いたします．また，本研究の一部は JSPS 科研費 JP22J11279 と JP22KJ2286 の助成を受けたものです．ここに謝意を表します．

参考文献

[1] 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.

[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] 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.

[4] 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.

[5] 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.

[6] 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.

[7] M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. Mazaré, M. Lomeli, L. Hosseini and H. Jégou, "The Faiss library," arxiv, 2024.

[8] T. Ge, K. He, Q. Ke and J. Sun, "Optimized Product Quantization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 4, pp. 744--755, 2014.

[9] P. H. Martins, Z. Marinho and A. F. T. Martins, "Chunk-based Nearest Neighbor Machine Translation," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, United Arab Emirates, Association for Computer Linguistics, 2022, pp. 4228--4245.

[10] Y. Meng, X. Li, X. Zheng, F. Wu, X. Sun, T. Zhang and J. Li, "Fast Nearest Neighbor Machine Translation," in *Findings of the Association for Computational Linguistics: ACL 2022*, Dublin, Ireland, Association for Computer Linguistics, 2022, pp. 555--565.

[11] R. Aharoni , Y. Goldberg, “Unsupervised Domain Clusters in Pretrained Language Models,” 著: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, 2020, pp. 7747--7763.

第 1 回 AAMT 若手翻訳研究会優秀賞受賞記念

キャラクターの性格と人間関係情報を付加した映像翻訳データセットの構築

大嶽匡俊†、加藤大地†、野﨑 雄斗†、廣岡 聖司†、宮尾祐介†、金崎朝子‡
東京大学(†)、東京工業大学(‡)

1. 概要

本研究では、映像翻訳におけるキャラクターの「性格」や「関係性」といったメタ情報が、どれほど翻訳結果に寄与するのかを探求するため、発話者のメタ情報を集めたデータセットと、その発話者の名前を付与した日英対訳データセットを構築した。

2. はじめに

機械翻訳は深層学習の進展により急速に進歩しており、特に Transformer ベースのモデルは高性能で広く応用されている。しかし、これは主に直訳に関する成果であり、追加の解釈が必要な翻訳は依然として難しい。ドラマや映画の翻訳を扱う映像翻訳も、直訳では対応することが難しく、工夫が必要である[i]。一つの工夫として、付加的な情報を入力に追加して精度を向上させようという目論見が存在する。例えば映像情報を追加した翻訳を行う研究は存在するが、大きく精度を改善することはできていない[ii]。

我々の事前の翻訳者へのインタビューによると、実際の翻訳の現場においては、登場人物の「性格」と、登場人物同士の「関係性」が重要な役割を果たしていることがわかった。「性格」に関連する研究として、役割語やキャラクター言語という考え方が存在する[iii]。役割語は社会的・文化的なステレオタイプに基づいて使われる話し方であり、キャラクター言語はさらにそれを拡張した、各個人の独特の話し言葉のスタイルを指す。ステレオタイプを強化しないよう慎重な取り扱いが求められるものの、セリフのスタイルにこの考え方を取り入れることで、登場人物の個性が際立つ。これにより物語が分かりやすくなることが期待されるため、翻訳の際に用いられることがある。

本研究では、映像翻訳におけるキャラクターの「性格」や「関係性」などのメタ情報がどれほど翻訳結果に寄与するのかを探求するため、発話者のメタ情報を集めたデータセットと、その発話者の名前を付与した日英対訳データセットを構築する。

3. 提案手法

本研究は、図 1 のように、「発話者のメタデータ」「発話者名付き対訳データ」という 2 種類の大規模なデータセットを構築する手法を提案する。これらは、「脚本データ」「字幕データ」から、「発話者情報」「セリフ情報」「整列情報」という中間情報を経由して構築される。

(ア) クリーニングとアライメント

脚本データとして、 Forever Dreamings[iv]という書き起こし（脚本）データを用いる。まず、Forever Dreaming から、英語の脚本データをダウンロードする。Forever Dreaming は TV シリーズのファンフォーラムとして設立された書き起こしウェブサイトで研究や教育のために利用されている。

その後、Chen ら iv に倣い、発話者情報が含まれていると判定された書き起こしデータを抽出する。発話者情報とは発話者の名前を指す。「作品内の各行を:又は] で分割し、分割数が 2 となったもの」を「発話」と判断し、前後をそれぞれ発話者とセリフ情報とする。発話と判断された行が 100 以上となった書き起こしデータのみを用いた。

日本語の字幕データは Opensubtitles[v]が提供する API から収集した。API で検索された日本語の字幕データと、前節で作成したセリフ情報との間でアライメントを取り、対応する文対を示す整列情報を作成する。アライメント手法は文埋め込みの類似度を用いる vecalign[vi]を用い、文埋め込みには LASER[vii]を用いた。

Construction of a Video Translation Dataset with Added Character Personality and Interpersonal Relationship Information
Masatoshi Otake†, Daichi Kato†, Yuto Nozaki†, Satoshi Hirooka†, Yusuke Miyao†, Asako Kanezaki‡
The University of Tokyo(†), Tokyo Institute of Technology(‡)

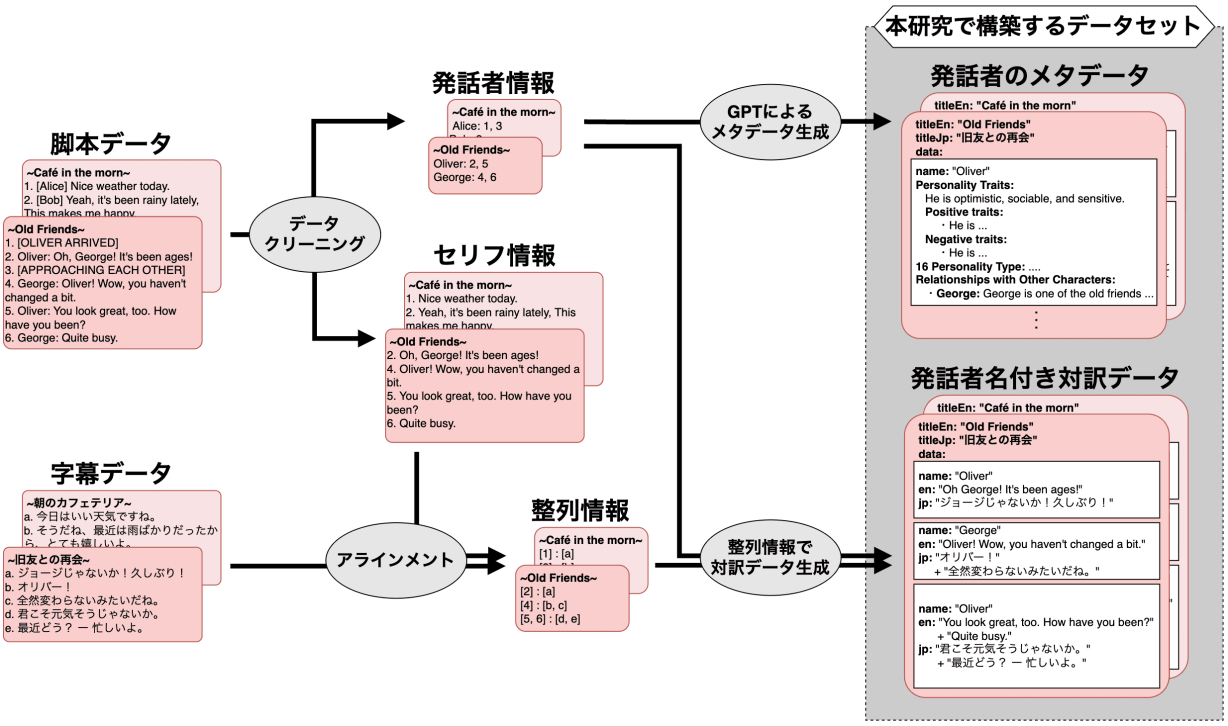


図 1 提案手法の概念図

(イ) 発話者のメタデータ・発話者名付き対訳データ生成

「発話者のメタデータ」は、作品内の発話者の「性格」と「人間関係」を含む。これらのデータは、発話者情報から得られる発話者名と作品名を使って作成されるプロンプトを用いて、gpt-3.5-turbo により生成する。性格情報は一般的な情報と、美点(Positive Traits)、欠点(Negative Traits)、性格タイプ(Personality Type)に分けて作成される。人間関係情報(Relationship)は、対象の人物と関係する人物について 10 人分生成する。

4. 今後の展望

今後の展望として、会話シーンの映像情報や、これまでの会話の履歴を加えるなどして、メタデータの種類を拡張していくことが挙げられる。また、どの種類のメタデータが、どの程度翻訳の質に影響を与えるのかについて、網羅的な評価を行うことも、非常に重要な研究となりうる。

5. 謝辞

本研究は、国立研究開発法人産業技術総合研究所事業の令和 5 年度覚醒プロジェクトの助成を受けたものです。

[i] Longyue Wang, Zhaopeng Tu, Yan Gu, Siyou Liu, Dian Yu, Qingsong Ma, Chenyang Lyu, Liting Zhou, Chao-Hong Liu, Yufeng Ma, Weiyu Chen, Yvette Graham, Bonnie Webber, Philipp Koehn, Andy Way, Yulin Yuan, and Shuming Shi. 2023b. Findings of the WMT 2023 shared task on discourse-level literary translation: A fresh orb in the cosmos of LLMs.(WMT 2023)

[ii] Zhishen Yang, Tosho Hirasawa, Mamoru Komachi, and Naoaki Okazaki. 2022. Why videos do not guide translations in video-guided machine translation? An empirical evaluation of video-guided machine translation dataset. Journal of Information Processing, 30:388–396.

[iii] Satoshi Kinsui and Hiroko Yamakido. 2015. Role language and character language. Acta Linguistica Asiatica, 5(2):29–42.

[iv] Chen, Zewei Chu, Sam Wiseman, and Kevin Gimpel. 2022. SummScreen: A dataset for abstractive screenplay summarization. (ACL 2022)

[v] Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. 2018. OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora.(LREC 2018)

[vi] Brian Thompson and Philipp Koehn. 2019. Vecalign: Improved sentence alignment in linear time and space. (EMNLP-IJCNLP 2019)

[vii] Mikel Artetxe and Holger Schwenk. 2019. Margin based parallel corpus mining with multilingual sentence embeddings.(ACL 2019)

第 1 回 AAMT 若手翻訳研究会優秀賞受賞記念

人手翻訳から MTPE へ：一翻訳者の所感

海老原仁美
レッドハット株式会社

1. はじめに

この度、若手翻訳者研究発表会にて登壇の機会をいただき、大変光栄なことに優秀賞を賜った。発表の機会をいただいた AAMT の皆様、ご清聴いただいた皆様、発表をサポートしていただいたローカリゼーションチームのメンバーに、この場を借りて心より御礼申し上げたい。

筆者は、長年にわたり人手翻訳に従事した後、MTPE[i]の世界に足を踏み入れた。本記事では発表内容をまとめ直す形で、人手翻訳と MTPE の両方を経験した翻訳者の視点での MTPE に対する所感を述べたい。

なお、以下の内容はあくまで私見であり、分野等によっても状況は異なる可能性がある点にご留意いただきたい。あくまで一企業の一翻訳者の例としてご笑覧いただければ幸甚である。

2. 筆者のバックグラウンド

2015 年 4 月に、新卒で特許翻訳者として法律事務所に入所。主に中間処理文書や特許訴訟の書面などの翻訳に従事する。様々な事情から、翻訳支援ツール（CAT ツール）や機械翻訳（MT）は一切使用せず、完全に人手による翻訳を行っていた。

8 年半ほど勤務したのち、2023 年 9 月にレッドハット株式会社に入社。同社は、オープンソースソフトウェアの事業を展開する IT 企業である。ここではテクニカルトランスレーターとして、製品ドキュメントやナレッジベースといったコンテンツの和訳を行っている。前職と打って変わって CAT ツールや MT を駆使しており、ほぼ 100%が MTPE である。

3. MTPE に対する所感

さて、このように人手翻訳から MTPE に転向して数ヶ月が経過した。以下、現時点で筆者が MTPE に対して抱いている所感を共有する。

人手翻訳との比較

まず、人手翻訳と MTPE を比較する。なお、特許と IT という異なる分野の経験に基づくため、必ずしも厳密な比較ではない点をご了承いただきたい。

スピードについては、基本的に MTPE の圧勝であると感じている。人手翻訳時代の和訳スピードは 1 時間あたり 250-300 ワードだったが、MTPE では 1 時間あたり 500-750 ワード前後[ii]となっている。分野は異なるものの、2-3 倍のスピードである。

正確性は、ほぼ同等の印象である。MT では時折突拍子もないミスもあるが、翻訳者がレビューを行う限りそれらのミスは修正できる。結果として、人手翻訳と同様の正確性は担保できる印象である。

流暢さについては人手の方が優勢といえる。ただし、以前と比べて MT の流暢さは向上している。読みやすさが重視される分野でフル活用するのはまだ難しいかもしれないが、マニュアルなどの技術文書では MTPE により十分な自然さを実現できると感じている。

翻訳プロセスの違いについても触れておきたい。筆者の翻訳プロセスを単純に説明すると、人手翻訳の場合はまず原文を読み込み、リサーチをしつつ翻訳・推敲し、最後に見直して仕上げるという流れである。一方 MTPE では、最初から MT の出力があるため、原文と照合しつつ MT の訳文を読み、適宜リサーチをしながら修正・推敲を行っている。どちらが良いというわけではないが、訳文作成のプロセスには大きな差異がある。

From Human Translation to MTPE: A Translator's Perspective
Hitomi Ebihara
Red Hat K.K.

MTPE のメリット・デメリット

次に、MTPE のメリットとデメリットを検討する。

まずメリットとして挙げられるのは、やはり効率化である。分野にもよるが、意訳が必要であったり、原文が難解であったりしない限り、かなり効率は上がると考えられる。効率化の結果、翻訳を諦めていた文書にも手が回るようになるという副次的効果も期待できる。

また、MT の出力が訳語のヒントになるというのもメリットである。当然ながら訳語の裏取りは必要であるが、特に専門知識がない分野においては、リサーチの手がかりがあるというのは大きな利点となる。

さらに、用語統一の負担が軽減される点も挙げられる。指定した用語を使用するように MT を設定することで、用語統一の手間を省くことができる。特に、ひとつの文書を複数人で翻訳する場合に恩恵が大きい。

デメリットとしては、まずミスを見落としやすい点が挙げられる。MTPE では、ある程度整った訳文が目の前にある状態で作業を始めることになる。そうすると、誤訳や訳抜けがあるにもかかわらず「それらしく書かれているから大丈夫だろう」と流してしまいやすい。MTPE においては、原文や訳文を疑う姿勢が人手翻訳よりも一層重要だと感じている。

どこまで手を入れるか迷うというのもデメリットとなり得る。例えば「内容は正確だが、語順を入れ替えた方が読み易いかもしれない」という時に、一手間をかけて修正すべきか悩み、心理的な負担になる可能性がある。ただし、この点については、修正の基準を明確に定めることである程度クリアできると思われる。

また、自然さを第一に追求する場合は大幅な書き直しが必要となるため、MTPE よりも人手で翻訳をした方が効率・品質の両側面で望ましい可能性がある。

翻訳力への影響

上記のデメリットに加えてよく耳にするのが、「MTPE に従事することで翻訳力が落ちる」という懸念である。この点に関して先行研究がないか探してみたものの、翻訳力低下を示す客観的なデータは見つけることができなかった。それでは、なぜ MTPE が翻訳力低下につながると感じるのだろうか。

まず考えられるのが、上述した訳文作成プロセスの違いに起因するものである。これは単純に、人手翻訳と MTPE ではプロセスが異なるため、MTPE のみを行っていれば人手翻訳の腕がなまる、ということである。

MT の出力への慣れも一要素として考えられる。具体的には、MT の単調な表現や、MT と CAT ツール併用時のぶつ切り訳などに対する慣れである。前者について具体例をあげると、MT では“How can I … ?”という表現が、一律「...するにはどうすればよいですか？」と訳される傾向があり、表現の幅に乏しいといえる。後者については、CAT ツールでは原文を文単位に区切るため、CAT ツールと MT を併用した場合、MT も文単位で訳文を出力する。結果として、MT の訳文は文同士のつながりが不自然になりやすい。こうした MT の出力に慣れてしまい、自分で翻訳する際にも違和感を覚えなくなってしまうと、翻訳の品質に影響する可能性は否定できない。

では、翻訳力を落とさないためにはどうすればよいのだろうか。まず、MTPE だけではなく人手翻訳も行ってスキル維持に努めることが重要である。さらに、ターゲット言語で書かれた良質な文章を読む、客観的なフィードバックを得るといったことも有益であろう。結局のところ、MTPE による翻訳力低下のリスクを下げるには、翻訳力を維持・向上する基本的な方法を実践するに尽きると考える。

MTPE への適性

次に、適性について考えてみたい。MTPE はやはり人手翻訳と大きく異なるものであり、すべての翻訳者が向いているとは言い難いかもしれない。

適性としてまず考えられるのは、レビューや校閲を楽しめることである。筆者自身は、翻訳の添削などにもやりがいを感じるため、MTPE にも楽しんで取り組

んでいる。一方で、一から自分で訳したい翻訳者は、あまり充実感を得られないかもしれない。

また、MTPE に限られないが、ニーズに合わせて柔軟に対応できるという点も重要である。高品質を追求したいと考えるのが多くの翻訳者の性であろうが、MTPE の場合、訳文を完璧に仕上げるのに十分な時間は確保できないことが多い。「品質よりもスピードを優先してほしい」と言われた場合に、依頼者のニーズを優先し、割り切って対応できる翻訳者は向いていると考えられる。

MTPE に移行して

さて、「人手翻訳から MTPE に移ってどう感じているのか、後悔はないのか」と聞かれることがある。回答としては、移行してみて正解だったと感じている。まず何より、人手翻訳に加えて MTPE のスキルがあれば、単純に仕事の幅が広がる。また、MTPE という手法は多くの依頼者のニーズと合致すると考えられ、今後は CAT ツールのスキルと同じように、MTPE のスキルも必須になっていく可能性がある。したがって、将来性という点でも、MTPE の経験を積むメリットは大きいと考えている。

4. MTPE の普及に向けて

最後に、翻訳者の間で MTPE を普及させるため必要なことを検討して終わりたい。

まず重要なのは、MTPE に対する理解の促進である。人手翻訳に従事していると、MTPE のメリットを体感する機会がないまま、何となくネガティブな印象を抱きがちである（過去の筆者も例外ではない）。このような先入観を払拭することは、MTPE 普及のために不可欠と考えられる。具体策としては、MTPE 体験会の実施などが挙げられる。興味のある人が実際に MTPE を試すことのできるイベントのようなイメージである。実際、隅田（2023）によれば、翻訳者に MTPE を体験してもらった結果、大部分の翻訳者がポジティブな感想を抱いたという。このように MTPE のメリットを体感してもらうことで、MTPE の印象改善につながることが期待される。翻訳者の中には、「過去に性能の低い MT のせいで苦労したので、もう MTPE には手を出さない」という人も存在するであろうが、ここ数年で MT の性能は著しく向上している。現在の MT の性能を知ってもらうという意味でも、体験する場を設けるメリットは大きいと考える。

依頼に際しては、依頼者と翻訳者の間でしっかりと認識を擦り合わせることが重要である。MTPE の場合、経済的合理性を重視する依頼者と、高品質を追求する翻訳者との間で認識が乖離しがちである。このようなすれ違いの結果、翻訳者は MTPE に対して「不当に安く買い叩かれる、無茶振りをされる」といったネガティブな印象を抱いてしまう。したがって、例えば「品質は下げてよいのでスピードを上げてほしい」といった要望があるとすれば、依頼者と翻訳者の間で品質や納期に関する共通認識を形成しておく、翻訳会社が間に入る場合は、翻訳会社が擦り合わせをしっかりと行うことが必要である。こうすることで、MTPE の依頼で生じがちな摩擦を防ぎ、ひいては MTPE に対する好印象につながるのではないだろうか。

5. まとめ

以上が、MTPE に対する現時点での所感である。MTPE に懐疑的な方や肯定的な方、関心のある方など様々かと思うが、筆者の経験が少しでも参考になれば幸いである。

6. 参考文献

隅田英一郎（2023）「ポストエディットの真実」*AAMT journal* 79: 37-40.

[i] 機械翻訳の出力を修正して訳文を作成する翻訳手法のこと。

[ii] レッドハット株式会社ではスピードを元に MTPE

の品質基準を定めており、1 時間あたり 500 ワードと 750 ワードの 2 つの基準が存在する。前者はフルポストエディット、後者はライトポストエディットに近いものである。各基準で修正ルールが詳細に定められている。

第 1 回 AAMT 若手翻訳研究会優秀賞受賞記念

大規模言語モデルに対する対訳データを用いた継続事前訓練による翻訳精度評価

近藤 海夏斗[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

本稿では、第 1 回 AAMT 若手翻訳研究会で発表した内容と、発表後に IWSLT2024 へ採択された論文の内容について解説する。本稿ではページ数の都合上，一部の内容のみ言及するため，詳細は IWSLT2024 の採択論文を参照していただきたい．

近年，GPT をはじめとする大規模言語モデルが，多くの自然言語処理タスクで成果を収めている．この大規模言語モデルの進展は，機械翻訳タスクにも影響を及ぼしている．例えば，昨年行われた機械翻訳コンペティション WMT23 では，GPT-4 が多くの言語対で，他の参加システムやオンライン翻訳サイトの結果を人手評価で上回ったことが明らかとなった [1]．一方で，LLaMA-2 7B，13B をはじめとするパラメータ数が 100 億前後の大規模言語モデルでは，既存の encoder-decoder モデルより翻訳精度が大きく劣ることが報告されている [2]．

そこで本研究では，パラメータ数が 100 億前後の大規模言語モデルの翻訳精度を高める手法として，2 段階の訓練を提案する．1 段階目では，原言語文と目的言語文が交互に出現するデータで継続事前訓練を行う．そして 2 段階目では，少量の高品質な対訳データで supervised fine-tuning を行う．提案法を適用した大規模言語モデルを，WMT22 General Machine Translation Task のテストデータをはじめとする 13 種のテストセットで翻訳精度を評価した．その結果，以下の 2 点が明らかとなった．

- 原言語文と目的言語文が交互に出現するデータで継続事前訓練した場合のみ，翻訳精度が向上する．そして，原言語文と目的言語文の順番によって，翻訳精度が向上する翻訳方向が変化する．
- 大規模言語モデルに基づく翻訳モデルは，従来手法である transformer に比べ，ノイズや話し言葉を含む文章の翻訳に頑健である．

2. 関連研究

大規模言語モデルを事前訓練する場合，単言語データを用いることが一般的である．しかし，対訳データを継続事前訓練データに取り入れることで，下流タスクの精度が向上するという報告がある．Briakou ら [3]は，大規模言語モデルの事前訓練に対訳データを用いることで，0-shot および 5-shot の翻訳精度が向上することを示した．

また，Xu ら [4]は LLaMA-2 のような主に英語で事前訓練されたモデルは，たとえ fine-tuning をしたとしても英語以外への翻訳精度が低いことを示した．この問題に対し，彼らは 1 段階目に単言語データで継続事前訓練し，2 段階目に少量の高品質な対訳データで supervised fine-tuning を行うという ALMA という手法を提案した．また，1 段階目の訓練において，単言語データだけでなく，対訳データも利用することで，さらに翻訳精度が向上することが報告されている [5, 6]．

これまで行われてきた大規模言語モデルに基づく翻訳モデルの研究は，WMT General Machine Translation Task および Flores-200 [7]のテストデータでのみ評価されている．したがって，従来手法である encoder-decoder モデルとの比較が十分に検証されていない．さらに，継続事前訓練のデータが翻訳精度

Evaluation of Translation Accuracy by Continual Pre-Training using Parallel Data for Large Language Models
Minato Kondo[1], Takehito Utsuro[1], Makoto Morishita[2], Masaaki Nagata[2]
[1] Deg. Prog. Sys.&Inf. Eng., Grad. Sch. Sci.&Tech., University of Tsukuba, [2] NTT Communication Science Laboratories, NTT Corporation, Japan

に与える影響についても不明である．本稿では，この 2 点について述べる．

3. 提案手法

本稿では，大規模言語モデルの翻訳精度を高める手法として，2 段階の訓練手法を提案する．1 段階目では，原言語文と目的言語文が交互に出現するデータで継続事前訓練を行う．そして 2 段階目では，少量の高品質な対訳データで supervised fine-tuning を行う．

3.1 原言語文と目的言語文が交互に出現するデータで継続事前訓練

原言語文を$\{x_1, \ldots, x_n\}$，目的言語文を$\{y_1, \ldots, y_n\}$とする．そして，$\{x_1, y_1, \ldots, x_n, y_n\}$のように，原言語文と目的言語文を交互に結合したデータを作成する．作成したデータのトークンを$\mathbf{z} = \{z_1, \ldots, z_n\}$とすると，以下の誤差関数が最小となるモデルのパラメータθを訓練する．

$$\mathcal{L}_1(\theta) = -\sum_t \log P(z_t | z_{t-c}, \ldots, z_{t-1}; \theta) \qquad (1)$$

ここで，cはモデルの最大入力トークン数である context length を表す．$\mathcal{L}_1(\theta)$は一般的な causal language modeling loss であり，直前までのトークンをもとに次のトークンを予測する．$\mathbf{z}$からc個のトークンを抽出するため，抽出したデータの先頭と末尾が，原言語文もしくは目的言語文の途中になる可能性がある．

3.2 少量の高品質な対訳データで Supervised Fine-Tuning

継続事前訓練を行ったあと，少量の高品質な対訳データで supervised fine-tuning を行う．原言語文を$\mathbf{x}$，$\mathbf{x}$の目的言語文を$\mathbf{y}$，そしてプロンプトを$I(\mathbf{x})$とする．Supervised fine-tuning では，以下の誤差関数が最小となるモデルのパラメータθを訓練する．

$$\mathcal{L}_2(\theta) = -\sum_{t=1}^{T} \log P(y_t | \mathbf{y}_{<t}, I(\mathbf{x}); \theta) \qquad (2)$$

ここで，Tは目的言語文のトークン数，$\mathbf{y}_t$は目的言語文のt番目のトークンを表す．$\mathcal{L}_2(\theta)$も一般的な causal language modeling loss であるが，目的言語文の出力のみ誤差を計算する．例えば，「Translate “Good morning into Japanese: おはよう”」とモデルに入力するとする．すると，モデルはプロンプトを含む全ての入力トークンの次トークンを予測する．しかし，プロンプトの出力は推論時に使用しないため，誤差関数から除外する．

4. 提案手法

4.1 概要

今回の評価実験では，単言語データで事前訓練された大規模言語モデルとして rinna/bilingual-gpt-neox-4b[1] (以下 rinna-4b という)を使用した．rinna-4b は，日本語が 1,730 億トークン，英語が 2,930 億トークンで事前訓練された 38 億パラメータの大規模言語モデルである．LLaMA-2 のように，主に英語で事前訓練された大規模言語モデルで翻訳モデルを構築する際には，単言語データと対訳データの両方で継続事前訓練する必要があることが報告されている [5, 6]．しかし，前述のとおり rinna-4b は，日本語と英語の単言語データで事前訓練が十分に行われているため，単言語データでの継続事前訓練は必要ないと考えられる．

4.2 データセット

4.2.1 継続事前訓練

継続事前訓練のデータとして，JParaCrawl v3.0 [8]，開発データとして，WMT20 の開発，テストデータ，および WMT21 のテストデータを使用した．なお，継続事前訓練で使用する JParaCrawl v3.0 は，LEALLA-large[2][9]で取得した文埋め込みベクトルのコサイン類似度をもとに，2,080 万文対をサンプリングした．これにより，rinna-4b のトークナイザーで 18 億トークンの対訳データとなった．

4.2.2 Supervised Fine-Tuning

Supervised fine-tuning の訓練データは，WMT20 および Flores-200 の開発，テストデータから作成した．なお，KFTT [10]の訓練データは，全データから 10,000

[1] https://huggingface.co/rinna/bilingual-gpt-neox-4b
[2] https://huggingface.co/setu4993/LEALLA-large

件をランダムサンプリングした．作成した訓練データの数は，英日と日英それぞれ約 15,000 件ずつとなった．また，supervised fine-tuning の開発データは WMT21 のテストデータを使用した．これらのデータに対し，以下のように，原言語で書かれたプロンプトを適用した．

英日翻訳のプロンプト

Translate this from English to Japanese:
English: {原言語文}
Japanese: {目的言語文}

日英翻訳のプロンプト

これを日本語から英語に翻訳してください:
日本語: {原言語文}
英語: {目的言語文}

4.2.3 テストセット

Supervised fine-tuning を行ったモデルの翻訳性能を評価するため，JParaCrawl v3.0 の評価で用いられたテストセットを使用した．なお，WMT20 および WMT21 のテストデータは，継続事前訓練および supervised fine-tuning の訓練・開発データに含まれるため除外し，WMT22 のテストデータを追加した．これにより，英日・日英のテストセットが 5 種 (ASPEC，JESC，KFTT，TED (tst2015)，Business Scene Dialogue Corpus (BSD))，英日翻訳のみのテストセットが 5 種 (WMT19，20 Robustness Task，IWSLT21 Simultaneous Translation Dev，WMT22 General Machine Translation Task)，そして日英のみのテストセットが 3 種 (WMT19，20 Robustness Task，WMT22 General Machine Translation Task)の全 13 種となった．

4.3 比較モデル

4.3.1 ベースライン

本稿ではベースラインモデルとして 2 つのモデルを作成した．これらのモデルの訓練データは 4.1.1 節および 4.1.2 節の訓練データを，開発データは WMT21 のテストデータを使用した．なお，訓練データに使用した JParaCrawl v3.0 は，英日と日英のデータとして，全データの半分にあたる 1,040 万文対ずつを重複がないようランダムサンプリングしたものを使用した．

Transformer このモデルはスクラッチから訓練した 10 億パラメータの transformer モデルである．モデルの構造は mT5-large[3] [11]のうち，vocab_size を 250,112 から rinna-4b と同じ値である 65,536 へ，feed-forward network の次元数を 2,816 から 4,096 へ変更した．これにより，モデルは encoder と decoder をそれぞれ 24 層もち，モデルの次元数は 1,024，attention の head 数が 16，そして dropout が 0.1 となった．トークナイザーは sentencepiece[4] [13]ライブラリを用いて作成し，サブワード分割手法を unigram，character coverage を 0.995，そして byte-fallback を有効にした．訓練はバッチサイズを 4,096 として 15 エポック (38,160 ステップ)行い，1,000 ステップごとに開発データの誤差を計測した．そして，開発データの誤差の最小値が 3 回更新しなければ学習を終了した．optimizer として AdamW ($\beta_1 = 0.9, \beta_2 = 0.98, \epsilon = 1.0 \times 10^{-8}$) [14]を使用し，weight decay と label smoothing を 0.1，gradient clipping を 1.0 とした．また，最大の学習率を1.0×10^{-3}，warmup ratio を 0.1 とし，inverse square root scheduler を適用した．さら訓練中は，bfloat16，gradient checkpointing，そして deepspeed [15] ZeRO stage 2 を適用した．訓練は 2 台の NVIDIA A6000 を用いて 17 日かかった．

Direct-SFT このモデルは，rinna-4b に LoRA [16]を適用してそのまま supervised fine-tuning したモデルである．このモデルの supervised fine-tuning は，4.2.2 節で述べたプロンプトを適用して行った．さらに，full fine-tuning と条件を近づけるため，LoRA を self-attention の query, key, valud, そして output の線形層，ならびに feed-formard network の 2 つの線形層に適用した．これにより，学習可能パラメータは 2,590 万となった．

[3] https://huggingface.co/google/mt5-large
[4] https://github.com/google/sentencepiece

表 1. BLEU および COMET の平均スコアと，Transformer と有意差を示したテストセットの個数．各行で最高のスコアを**太字**，ベースラインを上回るスコアを緑色で強調している．

(e) 英日翻訳

		ベースライン		継続事前訓練+Supervised fine-tuning							
				Mono		En-Ja		Ja-En		Mix	
評価指標		Transformer	Direct-SFT	full	LoRA	full	LoRA	full	LoRA	full	LoRA
BLEU	平均スコア	13.9	12.2	6.3	5.9	15.4	**15.5**	7.3	7.2	14.7	14.9
	有意差の数	-	1	0	0	8	**9**	0	0	7	8
COMET	平均スコア	79.0	79.6	75.6	74.8	**83.5**	83.3	76.9	76.8	82.9	82.9
	有意差の数	-	7	0	0	**8**	**8**	0	0	**8**	**8**

(b) 日英翻訳

		ベースライン		継続事前訓練+Supervised fine-tuning							
				Mono		En-Ja		Ja-En		Mix	
評価指標		Transformer	Direct-SFT	full	LoRA	full	LoRA	full	LoRA	full	LoRA
BLEU	平均スコア	16.8	12.5	7.9	7.1	7.8	7.6	**17.0**	**17.0**	15.9	15.8
	有意差の数	-	1	0	0	0	0	**2**	**2**	**2**	1
COMET	平均スコア	76.0	75.0	70.4	69.7	70.3	70.0	**77.8**	77.7	77.1	76.9
	有意差の数	-	7	0	0	0	0	**7**	**7**	6	6

4.3.2 継続事前訓練データの原言語文と目的言語文の順番

本稿では，以下のように原言語文と目的言語文の順番を変えた 4 パターンのデータを作成し，継続事前訓練を行った．

Mono 対訳データを日本語と英語の単言語データとみなす．すなわち，モデルへ入力される固定長のトークンは，日本語もしくは英語のみで構成される．

En-Ja 英日方向のみ対訳となる．すなわち，英文 1 サンプルの直後に，英文に対応する和訳文を結合する．

Ja-En 日英方向のみ対訳となる．すなわち，和文 1 サンプルの直後に，和文に対応する英訳文を結合する．

Mix En-Ja および Ja-En で作成したデータから，重複がないよう 50%ずつランダムサンプリングする．

これら 4 パターンで継続事前訓練を行ったあと，4.2.2 節で述べたデータとプロンプトで supervised fine-tuning を行った．

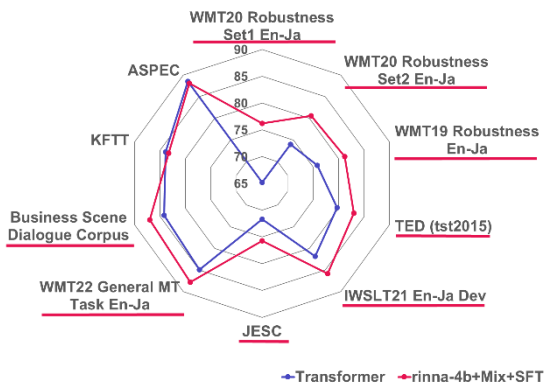
4.4 ハイパーパラメータ

4.4.1 継続事前訓練

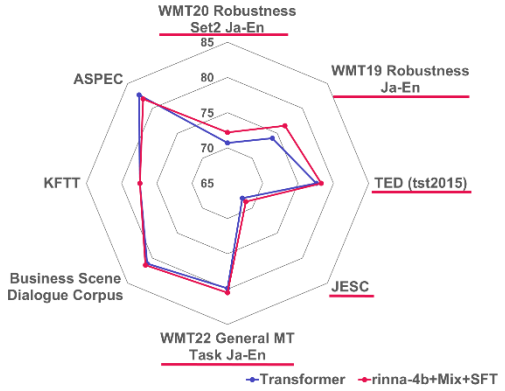
継続事前訓練では，optimizer として AdamW ($\beta_1 = 0.9, \beta_2 = 0.95, \epsilon = 1.0 \times 10^{-8}$)を使用し，weight decay を 0.1，gradient clipping を 1.0 とした．そして，rinna-4b のと同様に，context length を 2,048 とし，1 エポック訓練しながら，100 ステップごとに開発データの誤差を計算した．また，最大の学習率を1.5×10^{-4}とし，warmup ratio を 1%の cosine scheduler とした．訓練は NVIDIA RTX A6000 を 2 台用いて行い，それぞれの GPU のミニバッチサイズを 1，gradient accumulation step を 128 とすることで，トータルバッチサイズが 256 となる．訓練中には，bfloat16 および deepspeed ZeRO stage 2 を適用した．これらのパラメータにより，継続事前訓練は 10 日かかった．

4.4.2 Supervised Fine-Tuning

Supervised fine-tuning は，継続事前訓練にて開発データの誤差が最小となるモデルに対して行った．Supervised fine-tuning でも，optimizer として AdamW を使用したが，継続事前訓練で使用したパラメータのうち，β_2を 0.95 から 0.999 に変更した．Weight decay と gradient clipping については継続事前訓練と同じ値とした．最大の学習率を full fine-tuning では3.0×10^{-5}, LoRA では2.0×10^{-4}とし，warmup ratio を 1%の inverse square root scheduler とした．また，Direct-SFT ではエポック数を 1，バッチサイズを 256 とし，それ以外ではエポック数を 5，



(f) 英日翻訳



(b) 日英翻訳

図 1. COMET スコアのレーダーチャート．青線が Transformer，赤線が Mix 形式のデータで継続事前訓練したあと supervised fine-tuning を行ったモデルを表している．下線は，Transformer と有意差を示したテストセットを表している．

バッチサイズを 256 とした．LoRA のパラメータは，$r = 16, \alpha = 32, \text{dropout} = 0.05$とし，query, key, value の線形層に適用した．これにより，LoRA 適用時の学習可能パラメータは 640 万となった．

4.4.3 推論

全てのモデルは，開発データの誤差が最小となるモデルを用いて，bfloat16 を適用して推論した．Transformer は beam search で推論し，beam size を 4 とした．他のモデルについては，4.2.2 節で述べたプロンプトを用いて，greedy decoding を適用して推論した．

4.5 評価指標

評価指標として，BLEU [17]および COMET[5] [18]を使用した．BLEU は sacreBLEU[6] [19]を用いて計測した．COMET のモデルは wmt22-comet-da を使用した．なお，有意差の判定は，有意水準 5% ($p < 0.05$)で行った．

5. 結果

5.1 原言語文と目的言語文の順番による影響

表 1 は，BLEU と COMET スコアの平均と，Transformer との有意差を示したテストセットの個数を表している．この結果から，そのまま supervised fine-tuning を行った Direct-SFT より，原言語文と目的言語文が交互に出現するデータで継続事前訓練を行ったのち，supervised fine-tuning を行った En-Ja, Ja-En, Mix の方が高い精度であることが明らかとなった．また，今回の実験では，言語方向を一切明示していないにもかかわらず，Mix は英日・日英の両方向とも高い翻訳精度となった．この結果は，大規模言語モデルが訓練データに混在する対訳データから，原言語文と目的言語文の順番と一致する翻訳方向の知識として活用されることを示唆している．

5.2 テストセットごとの精度比較

図 1 は，Transformer と Mix 形式のデータで継続事前訓練をしたあと，supervised fine-tuning を行ったモデルの COMET スコアをレーダーチャートにした図である．この結果から，大規模言語モデルに基づく翻訳モデルは，Reddit をドメインとする WMT19, 20 Robustness Task および，TED Talk もしくは映画字幕をドメインとする TED (tst2015)，IWSLT21 En-Ja, JESC において Transformer を大きく上回っている．この結果は，大規模言語モデルに基づく翻訳モデルが，ノイズや話し言葉を多く含むデータに対して，従来の encoder-decoder モデルよりも頑健であることを示唆している．

[5] https://github.com/Unbabel/COMET
[6] https://github.com/mjpost/sacrebleu

6. おわりに

本稿では，原言語文と目的言語文が交互に出現するデータで継続事前訓練を行ったあと，少量の高品質な対訳データで supervised fine-tuning を行うという 2 段階の訓練を提案した．提案法を適用した大規模言語モデルを, 13 種のテストセットで評価した. その結果, 継続事前訓練データの原言語文と目的言語文の順番と同じ翻訳方向のみ翻訳精度が向上することが明らかとなった．さらに，大規模言語モデルに基づく翻訳モデルは，従来の encoder-decoder モデルに比べ，話し言葉を多く含む文章の翻訳に対して頑健であることも示した．本稿では，大規模言語モデルとして rinna-4b を用い, 英日・日英翻訳のみ評価を 1 文単位で行った．したがって，他の大規模言語モデルや言語対に対しても同様の結果となるか，そして長文翻訳における提案法の有効性を調査することが今後の課題である．

参考文献

[1] T. Kocmi, et al. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proc. of the 8th WMT*, pp. 1–42, 2023.

[2] W. Zhu, et al. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of NAACL2024*, 2024.

[3] E. Briakou, C. Cherry, and G. Foster. Searching for needles in a haystack: On the role of incidental bilingualism in PaLM's translation capability. In *Proc. of the 61st ACL*, pp. 9432–9452, 2023.

[4] H. Xu, Y. Kim, A. Sharaf, and H. Awadalla. A paradigm shift in machine translation: Boosting translation performance of large language models. In *Proc. of the 12th ICLR*, 2023.

[5] D. Alves, et al. Tower: An open multilingual large language model for translation-related tasks. *arXiv:2402.17733*, 2024.

[6] J. Guo, et al. A novel paradigm boosting translation capabilities of large language models. In *Findings of th NAACL2024*, pp. 639–640, 2024.

[7] NLLB Team et al. No language left behind: Scaling human-centered machine translation. *arXiv:2207.04672*, 2022.

[8] 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.

[9] Z. Mao and T. Nakagawa. LEALLA: Learning lightweight language agnostic sentence embeddings with knowledge distillation. *In Proc. 17th EACL*, pp. 1886–1894, 2023.

[10] G. Neubig. The Kyoto free translation task. http://www.phontron.com/ kftt, 2011.

[11] L. Xue, et al. mT5: A massively multilingual pre-trained text-to-text transformer. *In Proc. of the NAACL2021*, pp. 483–498, 2021.

[12] T. Kudo and J. Richardson. SentencePiece: a simple and language independent subword tokenizer and detokenizer for neural text processing. *In Proc. EMNLP2018*, pp. 66-71, 2018.

[13] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *In Proc. 7th ICLR*, 2019.

[14] J. Rasley, S. Rajbhandari, O. Ruwase, and Y. He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. *In Proc. 26th ACM SIGKDD*, pp. 3505–3506, 2020.

[15] E. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-rank adaptation of large language models. *In Proc. 10th ICLR*, 2022.

[16] 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.

[17] 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.

M. Post. A call for clarity in reporting BLEU scores. *In Proc. 3th WMT*, pp. 186–191, 2018.

第 1 回 AAMT 若手翻訳研究会優秀賞受賞記念

日英間の機械翻訳における受容化と異質化について

木内晶基
東京科学大学

1. はじめに

従来，機械翻訳は主に産業翻訳を中心に議論されてきた[1]が，最近では動画サイトの字幕翻訳や SNS の翻訳機能など，より口語的な表現を含んだ，目的の異なる翻訳にも用途は拡大傾向にある．それらの場面では PE（ポストエディット）なしで機械翻訳による出力が人々の目に触れられる機会も多い．近年では，出力文に見られるバイアスを明らかにする分析研究も活発化しており[2,3]，これらは機械翻訳がユーザーや社会に与えうる影響について注意を促し，今後の開発指針を示す上で重要な役割を持つ．

一方で，それらのバイアス研究で着目されるトピックは，出力文の語彙の多様性やジェンダー表現など限定的である．特に，翻訳が本質的にもつ文化媒体としての役割に着目し，機械翻訳が社会や文化に与える，または受ける影響について議論をするケースは少ない．そこで，本稿では文化情報の翻訳に着目し，翻訳学の概念を援用しながら機械翻訳が語彙レベルでどのような翻訳をしているのかを観察し，その特徴について考察を試みる．本稿は第 1 回 AAMT 若手翻訳研究会にて，著者の修士論文[4]の内容を元にして発表をした「日英間の機械翻訳における受容化と異質化について」の解説記事に当たる．発表内では取り扱えなかった説明を加え，今後の更なる議論の契機となることを期待する．

2. 受容化と異質化

初めに，機械翻訳の出力分析に移る前に，関連する理論について導入を行う．翻訳理論家のヴェヌティは「受容化（domestication）」と「異質化（foreignization）」という 2 つの翻訳方略について論じている[5]．受容化とは，元からその言語で書かれたかのような，流暢な目標文化志向の翻訳で，異質化は，起点テクストの外的な帰属を感じさせる，流暢ではない異質な起点文化志向の翻訳である．これはシュライアーマハーの同化と異化の翻訳論[6]からその議論を発展させたものである．ここでヴェヌティは翻訳者の不可視性（invisibility）を取り上げ，翻訳者の存在が見えないような流暢な翻訳ではなく，それが見えるような異質な翻訳をとるべきだと主張した．

ヴェヌティは，アングロ=アメリカ圏の文芸翻訳に見られる受容的な翻訳の風習を自民族中心的だと批判し，異国のアイデンティティをなかったかのようにする受容化の暴力性に抵抗して，異質化によって言語的な差異や文化的な差異を示す必要があると訴える．また，受容化・異質化を権力やイデオロギーに関わる問題として上で，そのような目標言語の読者に読みやすい本を作ろうとする翻訳業界や，その元で低賃金で働く翻訳者の生存競争にも触れ，受容化の傾向が個人だけでなく社会構造の中から生み出されると主張した．

これに関して，ヴェヌティの理論を日本語へ適用する際に，受容化が日本語圏では自民族中心主義と結びつかないという議論がある[7, 8]．ヴェヌティ自身は受容化や異質化の意味は時や場所が変われば変化しうるとしている．つまり，アングロ=アメリカ圏で受容化が自民族中心主義と結びつくように，日本語圏においても受容化や異質化が持つ特有の意味が存在するといえる．「受容化」と「異質化」の持つ意味合いに関する定義上の問題が指摘されながら，いずれにしても変わらないのは，他文化を読みやすいように置き換える翻訳と他文化の異質さをそのまま残す翻訳が起きて

Domestication and Foreignization in Japanese-English Machine Translation
Masaki Kinouchi
Institute of Science Tokyo

いて，それらがその先の社会や文化と相互に影響を与え合っているという点である．

3. 機械翻訳への応用

本研究ではこの受容化と異質化を援用して分析を行う．機械翻訳が翻訳の方略や意図に当たるものを持たないとしても，原文と訳文を比較して翻訳の手法を分類し，その傾向を調べることで，そこに見られるバイアスから擬似的に翻訳者としての特徴を知ることは可能である．既に述べたように日本語を目標言語とした翻訳で，受容化が自民族中心主義と結びつかないことを除いても，受容化や異質化が背後の権力やイデオロギーと結びつけられて考察されることは重要である．それは特に，データの量や技術力・人材など様々な外周的な要素に，権力に関する議論が発展する可能性がある機械翻訳において，より一層重要だと考える．そのため，本研究では，依然としてヴェヌティの受容化・異質化を中心概念として据え，分析を試みる．

また，ヴェヌティの指摘する受容化の傾向は文芸翻訳において見られ，産業翻訳については情報を正確に伝えることが求められるため状況は異なるとされる．その点では，本研究で扱う機械翻訳は，決して文芸翻訳をすることが目的とされていない[1]ため，ヴェヌティ自身の言及する翻訳の対象とはならないかもしれない．しかし，現状として映像字幕やSNSなどに自動翻訳が使われ始めている以上，その目的は徐々に拡大され，機械翻訳は少しずつ文芸にも寄った翻訳を担うようになっている．この点で，機械翻訳についても受容化・異質化の視点からその傾向を明らかにし，社会や文化にどのような影響を与え得るのか議論をしていく正当性があると考えられる．

4. 翻訳タスク

ここで，実際に語彙レベルでどのような受容化や異質化が起きるかを確かめる．映像字幕に対して機械翻訳タスクを行い，出力に対してテクスト分析を行った．

4.1. 訳文の生成

訳文の生成には基礎的な構造をとるモデルとしてTransformer[9]を，流通している商業的なモデルとしてGoogle翻訳とDeepLを用意した．訳文に見られる特徴が根幹の Transformer からそれをベースとした強力なモデルにまで一貫して見られる特徴であるか，また，エンドユーザーが触れる商業的なサービスの現状がどのようであるかを確かめる．Transformer の実装には Web 上の大規模なテキストデータから構築された JparaCrawl[10]の学習済みパラメータを用いた．また，テストデータにはTEDとTEDxJapanの字幕翻訳を用意した．英日方向については既存の TED コーパス （IWSLT2017[11]，1452 対）を用いたが，日英方向については日本版 TED から翻訳された字幕用テストセットが存在しなかったため，本研究の中で新しく TEDx Japan から日英の字幕用テストセットを用意した（1472 対）．これらを用いて文単位で翻訳タスクを実行した．

4.2. 異文化要素の抽出

受容化と異質化の傾向を調べる方法の一つとして，異文化要素の訳出方法を分類する研究手法がある．ある文化に固有な表現について，翻訳研究の中で様々な定義がされてきた[12,13,14]．本研究ではアイセラによる定義を用いて，異文化要素は「目標テクストの読者の文化体系に参照される項目が存在しない，もしくは文章内の位置づけが異なること」によって翻訳時に問題を引き起こす表現[13]，として定義する．同様な概念について様々な呼称があるがここでは「異文化要素」に統一する．異文化要素の抽出には先行研究[14, 15,16]を参照しながら精査し，人名と地名は訳出が固定化されるため対象から除いた．

表 1. アイセラの翻訳手法の分類

異質化	複写	目標テクストにそのまま複製する“Harlem Grown”→“Harlem Grown”
	表記の適応化	目標言語のアルファベット表記への変換“food stamp” → 「フードスタンプ」
	言語内翻訳	言語的な透過による翻訳，起点言語の面影がわかる，起点文化の要素を残した翻訳 “brothers and sis- ters” →「兄弟姉妹」
	テクスト外注釈	起点テクストの要素を残しつつ目標テクストの外に注釈を足す，フットノートや訳者あとがきでの追記
	テクスト内注釈	起点テクストの要素を残しつつ目標テクスト内に注釈を加える，より詳細な情報を付与する 「東京」→ “the Japanese capital, Tokyo”
受容化	類義語	目標テクストの中で繰り返しを避けるために類義語や他の表現で置き換える “Harlem Grown” → 「そのプログラム」
	限定一般化	起点文化内の範囲で，目標文化にとって馴染みのある他の異文化要素に置き換える “grand” →「1000 ドル」
	絶対一般化	文化要素が除かれたより一般的な言葉や上位語への置き換え “r-word” →「差別用語」
	帰化	目標文化の中におけるおおよそ同等な異文化要素への置き換え “dollar” →「円」
	削除	異文化要素を省略，削除する “the pews”→ (省略)
	創造	目標テクスト内で新しく異文化要素をつくる “Fantastic Four” → 「宇宙忍者ゴームズ」

4.3. 翻訳手法の分類

異文化要素がどのように翻訳されたか，アイセラの分類[13]を用いて翻訳手法を同定したのちに，異質化（複写，表記の適応化，言語的翻訳，テクスト外注釈，テクスト内注釈）と受容化（類義，限定一般化，絶対一般化，帰化，削除，創造）の翻訳方略・翻訳手法に分類してその出現頻度をみる．各翻訳手法については表 1 を参照されたい．その後，人手の翻訳と機械による翻訳を質的にテクスト分析して比べながら，英日と日英方向にて見られた特徴を整理する．

また，人手の翻訳と機械による翻訳の決定的な違いとして「エラー」が挙げられる．機械翻訳に関しては誤訳や訳抜けが想定されるため，本研究では分析中に遭遇した機械翻訳におけるエラーのパターンとして誤訳，未訳，訳抜けをアイセラの分類に統合して分析を行う．

5. 異文化要素の訳出結果・考察

実際に異文化要素の抽出を行い，それらが訳出された箇所について分析を行った．異文化要素は英日方向で 194 個，日英方向で 95 個確認された．また，表 2 にそれぞれの翻訳手法の出現回数を示す．

本稿ではすべての翻訳方法について例文を示しながら詳説することを避け，特に考察を要する特徴のみを取り上げて解説を行う．機械翻訳による異文化要素の訳出について，以下のような特徴が挙げられた．

(1) 言葉の言い換えをしない

一般に機械翻訳について言われる特徴ではあるが，見られた特徴の一つとしてここに並べて取り上げたい．人手の翻訳と機械翻訳を比較したときに，機械翻訳文では字句通りの訳や逐語訳に当たる「言語的翻訳」が増え，情報を付け加える「テクスト内注釈」や平易な表現に置換する「絶対的一般化」，繰り返しの表現を避ける「類義語」が減っている．逆に人手ではそれらが柔軟に行われる．下に実際の訳文の例を示す. HT のような置換とは対照に Transformer ではそのまま訳された．

[例文 1]

ST Then we dropped them off in schools, seventh through eleventh grade.

HT 次に これを中学 高校に持ち込みました

MT(T) その後、7 年生から 11 年生まで、学校に送り出しました。

(2) 英日方向での言語規範を誇張するような出力

次の特徴として，英日方向の機械翻訳では人手に比べて多くの「複写」と「表記の適応化」が見られた．転写と表記の適応化の例を下にそれぞれ示す．

[例文 2，複写]

ST It’s called the Meta 2.

HT これは「Meta 2」です

表 2. 各翻訳手法の出現回数

（英日）		HT	MT(T)	MT(G)	MT(D)
異質化	複写	10	26	16	13
	表記の適応化	29	56	45	59
	言語内翻訳	64	87	98	83
	テクスト内注釈	13	2	8	11
受容化	類義語	3	-	-	-
	限定一般化	1	-	1	1
	絶対一般化	56	11	25	19
	帰化	13	-	-	-
	削除	5	-	-	-
エラー	未訳	-	5	-	-
	訳抜け	-	3	-	6
	誤訳	-	4	1	2

（日英）		HT	MT(T)	MT(G)	MT(D)
異質化	複写	1	3	3	3
	表記の適応化	14	13	15	13
	言語内翻訳	24	56	49	48
	テクスト内注釈	5	-	-	-
受容化	類義語	11	3	-	3
	限定一般化	-	-	1	1
	絶対一般化	33	13	27	24
	帰化	3	-	-	1
	削除	4	2	-	2
エラー	未訳	-	-	-	-
	訳抜け	-	2	-	-
	誤訳	-	3	-	-

*HT: 人手の翻訳, MT(T): Transformer, MT(G): Google 翻訳, MT(D): DeepL
（1 度も確認されなかった翻訳手法については表から除いた）

[例文 3, 表記の適応化]

ST (...), and my pastor called us out of the pews and down to the altar because ...

MT(T) (...)私の牧師は(...)、私たちをピューから祭壇に呼んでくれました。

これらは，英日方向では人手よりも機械による翻訳で多く確認されたものの，逆の日英方向では頻度に大きな差が確認されなかった．機械翻訳の複写については，頭字語(ESPN や TSN など)やそのほかに組織，サービスなどの固有名詞がそのまま和訳の中に移されたものが多く現れた．我々が日本語の文中に英語表現を度々用いる一方で，英語の文中に日本語表現を見ないように，言語システムによって他のある言語を文中に受容する度合いは異なる．人手の翻訳であれば，ある程度，日本語読者に馴染みのない頭字語を正式名称に戻すなどの対応がされるが，機械翻訳では頭字語に代表されるような固有名詞がそのまま写されるという形が頻繁に見受けられた．

表記の適応化では，コラーゲンフィンガープリント(collagen fingerprinting) やアクションアイテム(action item) など，馴染みのあるカタカナを組み合わせた複合語が多く見られた．これらの複合語は機械翻訳が創造的にカタカナ化したのではなく，いずれも日本語の辞書には登録されていないが WEB 上で調べればカタカナ表記で現れる言葉たちである．つまり，今まさに日本語の中で受容されようとしている表現たちと言える．それらを人間の翻訳家は「コラーゲン鑑定」や「活動項目のリスト」など日本語に馴染ませて訳すのに対して機械翻訳はそのままカタカナで表した．このカタカナの用い方に人手と機械で大きな差が見られる．これについても，その傾向は固有名詞に対して強く現れた．逆に，日英方向では表記の適応化にあたるローマ字への変換の頻度について，人間と機械で大きな差はなかった．カタカナによって外来語を取り入れる日本語体系の規範が機械翻訳では誇張されて現れているように見て取れる．

(3) 日英方向でのみ観察される機械翻訳による削除

また，日英方向の翻訳では，人手だけでなく機械翻訳による「削除」も確認された．削除は受容化の翻訳手法の中でも，特に受容的な手法として位置づけられており[13]，どのような情報を削除するかは翻訳者の特徴が色濃く現れると言っても良いだろう．下に削除の例文を並べる．

[例文 4]

ST これは 13 校の高校 中学校 そして 60 人以上の中高生スタッフが 集まって作り上げた ひとつの学校祭型イベントで 学校の垣根を超えて 私の地元の帯広広小路商店街で行いました

MT(T) This is a school festival-type event created by 13 high schools, junior high schools, and more than 60 junior high and high school staff.

[例文 5]

ST 野菜栽培するだけではなくて バーベキュー大会で料理をしたり(…)

MT(D) We not only grew vegetables, but also cooked at barbecues, (…)

[例文 6]

ST 私は(…)お客様に「いらっしゃいませ」「ありがとうございました」といわば 看板犬ならぬ看板娘でした

HT (…) and I would greet customers saying ”welcome” and ”thank you.” I was basically the ”mascot dog” of the store, so to speak

Transformer では[例文 4]のような文章単位の削除が見られた．例文では文章後半部分「学校の垣根を超えて（...）で行いました」に該当する訳がない．しかしこのような削除は，前後の文脈を参照して意味が通っているか確認した上でようやく，それが削除か訳抜けかを判断することができる．この実験では，機械翻訳自体は前後の文を参照することができないため，それらが訳抜けではなく削除として働いたことはかなり恣意的な結果と言える．

一方で DeepL では[例文 5]のように単語やフレーズ単位で削除が見られた．これは，[例文 6]のような人手の翻訳に見られる削除と特徴が近い．(1)で述べたように機械翻訳の特徴の一つとして字句通りに訳す性質が挙げられるが，そのような性質とは対極にある削除を機械翻訳が実行していることは興味深い．どのような情報を削除して，それが何に起因するのか，人間による削除とどのように異なるのか，といったことは今回の実験では削除による翻訳のサンプルが少なかったため確かめられず，今後の課題として残される．

6. おわりに

ここまで，語彙レベルで実際にどのように受容化や異質化が起きるのかを調べるため，異文化要素の訳出に着目してその傾向を分析した．まず(1)にまとめたように，語彙レベルで機械翻訳は異質化をとる傾向がありながら，(2)や(3)でまとめたように，それが翻訳する言語ペア（本研究で言えば翻訳する方向）によって質的に大きく異なることが分かった．

ヴェヌティの議論に沿うならば，(1)の機械翻訳の全体的な異質化傾向は，アングロ＝アメリカ圏の自民族中心主義的な姿勢を緩和する方向に機械翻訳が働いている可能性があることを示す．日本語圏では 2 章で触れたように受容化と自民族中心主義は関連されないが，どちらの翻訳の方向についても，語彙レベルで異文化的情報は保持される傾向にある．これは(1)に述べたように，置き換えや情報の付与がされない技術的な制約が原因となっていると考えられる．

ただし，同じ異質化でも程度に差があり，表 2 に見られるような頻度の違いや，(2)や(3)に挙げたような質的な特徴の違いがある．これらから，英語を目標言語としたときに相対的に受容的であったと言える．ここでは学習データの性質やそれを形成する言語資源に関する社会的制約についても議論が予想されるが，本研究の中では扱わず今後の課題として考えたい．本研究では，文化的な情報を対象とした機械翻訳の翻訳手法について，そのバイアスを明らかにした．しかし，ここでは用いたデータ，モデルは限定的であったため，それらが異なる場合や，ポストエディットが介在する人的な営みとしての機械翻訳についても今後調べる必要がある．また，実際に訳文がどのような場所でどのように読者に受容され，最終的に人々の生活や思想，言語使用などにどのような影響を与えるのかを調べていくことで，「社会への影響」を評価していく必要があるだろう．

本研究で提示したのは，機械翻訳の文化媒体としての特性をテクストから実践的に分析する 1 つの手法であり，その際に翻訳学の受容化と異質化を用いた．機械による翻訳が精度を上げて一般に普及するほど，既存の評価指標のように人の翻訳と言語的に比べるのみならず，社会の一部として，翻訳の産物としてその影響を見る必要が出てくるように思われる．このような

研究により，今後益々翻訳学と機械翻訳研究の融合的な研究が促されることを期待して，解説記事の結びとする．

付録

テストデータに用いた TEDxJapan の映像資料

1. Encouragement for a Goal-Free Life, Mitsufumi Nishu, TEDxAnjo
2. Something Old Yet New, Shinichi Fukuyama, TEDxAwaji
3. わくわくの輪郭, Hitomi Matsui, TEDxDoshishaU
4. Mathematics and Us, Takehiko Nakama, TEDxDoshishaU
5. 「だからこそ」の可能性, Mina Tanaka, TEDxDoshishaU
6. 突き動かしたものは何か, daisaku harasaki, TEDxNagasakiU
7. Wishing for the happiness of others., Hiroshi Hatano, TEDxHamamatsu
8. Talent, death and love, KITANO Yuiga, TEDxKobe
9. Invention of swarming molecular robots and its infinite possibilities, Akira Kakugo,TEDxSapporo
10. 両想いの法則, Ayumi Yamamoto, TEDxSapporo
11. 宇宙のゴミを減らすために, Miki Ito, TEDxKyoto
12. スポーツの持つ力, Hiroko Morohashi, TEDxKyoto
13. 紛争地で見つめ直した看護の力, Yuko Shirakawa, TEDxKyoto
14. Not talk ABOUT refugee but talk WITH refugee, Sayaka Watanabe, TEDxHamamatsu

参考文献

[1] T. Poibeau, Machine Translation, The MIT Press, 2017.

[2] E. Vanmassenhove, D. Shterionov and M. Gwilliam, “Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation,” in *Proc. of the 16th Conf. of the European Chapter of the Association for Computational Linguistics: Main Volume (EACL)*, [Online], 2021, pp.2203–2213.

[3] G. Stanovsky, N.A. Smith, and L. Zettlemoyer, “Evaluating gender bias in machine translation,” in *Proc. of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, Italy, 2019, pp.1679–1684,

[4] 木内晶基, “機械翻訳における受容化と異質化に関する研究,” 修士論文, 東京工業大学, 2024

[5] L. Venuti, The Translator’s Invisibility: A History of Translation, 1st ed. Routledge, 1994.

[6] F. Schleiermacher, "4. From On the Different Methods of Translating," in *Theories of Translation: An Anthology of Essays from Dryden to Derrida*, University of Chicago Press, 1992, pp. 36-54.

[7] 水野的, “本アンソロジーを読むために,” 著: *日本の翻訳論 アンソロジーと解題*, 柳父章, 水野的 , 長沼美香子, 共同編集, 法政大学出版局, 2010, pp. 36-53.

[8] 明石元子, H. James, “著名翻訳家・テクスト分析・可視性概念,” *通訳翻訳研究*, 第 14, pp.183-201, 2014.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” 2017.

[10] M. Morishita, J. Suzuki, and M. Nagata, “JParaCrawl: A large scale web-based English-Japanese parallel corpus,” in *Proc. of The 12th Language Resources and Evaluation Conference*, France, 2020, pp.3603–3609.

[11] M. Cettolo, C. Girardi, and M. Federico, “Wit3: Web inventory of transcribed and translated talks,” 2012.

[12] P. Newmark, A Textbook of Translation, Prentice Hall, 1988.

[13] J.F. Aixelá, “Culture-specific items in translation,” in *Translation, Power, Subversion*, R. Alvarez and M.A. Vidalın, Eds. Multilingual Matters, 1996, pp.52–78.

[14] E.E. Davies, “A goblin or a dirty nose?,” *The Translator*, vol.9, no.1, pp.65–100, 2003.

[15] J. Marco, “The translation of food-related culture-specific items in the valencian corpus of translated literature (covalt) corpus: a study of techniques and factors,” *Perspectives*, vol.27, pp.20–41, 2019.

[16] J. Pedersen, Subtitling norms for television, John Benjamins, 2011.

温故知新

温故知新 5

後藤 功雄
AAMT

AAMT では、AAMT 創立３０周年記念事業として、過去の AAMT ジャーナルおよび JAMT ジャーナル（AAMT の前身である JAMT（日本機械翻訳協会）の会誌）を PDF 化して公開しています。

https：//www.aamt.info/act/journal/

公開されているジャーナルは次の通りです。

- JAMT ジャーナル No.01,1991 年 7 月～No.07,1992 年 8 月（今回 PDF 化）
- AAMT ジャーナル No.01,1992 年 11 月～No.70,2019 年 6 月（今回 PDF 化）
- AAMT ジャーナル「機械翻訳」No.71,2019 年 12 月～（当初より PDF 版を公開）

「温故知新」シリーズでは、過去の AAMT ジャーナルおよび JAMT ジャーナルの記事を紹介します。

過去のジャーナル記事を読むと、当時の MT の事情が分かるとともに、現在も解決していない課題が残っていることが分かるなど、これからの MT 発展に役立つものが多く見られます。

本号では、1993 年 2 月と 1993 年 5 月号の AAMT ジャーナルから 2 つの記事を転載して紹介します。これらは、PDF から OCR で読み取ったテキストを修正したものです。画像等を含むオリジナルの原稿については、上記サイトをぜひご覧ください。

【1993 年 2 月 AAMT ジャーナル No.2 より】

翻訳の現場から

機械翻訳時代の翻訳者の役割は？

元吉 宏子

原稿用紙（実はパソコン）を前に機械翻訳についてあらためて考えてみた。締切が近づくと「ああコンピュータに原稿を読み込ませるだけで翻訳が仕上がり、後はコーヒーを飲みながらのんびりチェックすれば..なんてことにならないかな」と想像することはよくあるが、具体的な知識はほとんどない。

衛星放送で「機械翻訳」という但し書きがついた字幕のニュースを見たことがあるが、とても楽しい頭の体操になった。というのは、字幕の日本語から、原文の英語の輪郭がはっきり読み取れ、自分も多分はこう訳出してから、前後を考えてもっと適切な単語に変えるとか、文章の順序を動かすかなというプロセスが次々と頭に浮かんできたからである。でも、英語を日本語におきかえるという作業に日頃なやまされている翻訳者としては、行間ににじみでている翻訳機の苦労にすくなからず親近感をおぼえ、とてもひとごとと思えない共感を感じた。

そんなことを考えながら、学生時代に読んだときは文の流れがつながらず、途中で放り出し「ジェーン・エア」の日本語訳をもう一度読んでみた。そこには、あの機械翻訳と同じようにあたかも英文を読んでいる気分になるほど「英語らしい（?）」日本語の文章があふれていた。でも翻訳者としての自分も知らず知らずに同じ間違いをおかし、お読みになる方々に迷惑をかけているんだろうなと反省しきりで、またしても途中で放り出してしまったが..

理科系出身者は科学全般に強いという大いなる幻想? のおかげもあり、翻訳者としてラッキーなスター

Learning from the past 5
Isao Goto
AAMT

トがきれた。小説の登場人物が何を考えているのか、この文章の作者の意図は、などのいわゆる「現代国語」に高校時代あまり関心がもてなかったという単純な理由で理科系に進んだこともあるが、科学的文章の翻訳の方が論理という強い見方があるため取り組みやすい。Native speaker 以外の人が書いた英文は文法的ミスもあり、文章が完結していないものさえあるが、科学的文章の場合は、資料を調べ、文脈をたどればある程度自信をもって著者の意図を推測し、翻訳の段階で文章の完成を手伝えるという楽しみも稀にある（間違った楽しみかもしれないが)。著者の独りよがりではなく、論理をたどれば必ず理解できるという手ごたえがある。また、同じ意味を表現するのに、どの言葉、どんな文章を選ぶかにそれほどなやまなくても、事実と論理という強力な助っ人の力をかりて忠実な翻訳文が作れるという心強さもある。

もちろん、科学的文章がかならずしも無味乾燥なあじわいのないものではなく、記憶が違わなければポーリング博士（ノーベル賞受賞の化学者）が大学１年生向けに書いた教科書の英文はとても美しく、人間性がにじみでるような香り高い文章だったと思う。同じテーマでも、書き手が違うとまったく趣が違う。翻訳者も正確さにくわえて香り高い文章が書ければ一人前になれるのだろうが、むずかしい作業であることは毎日痛感させられている。

それとは矛盾するが、遠くない将来に機械が大きな力を発揮するようになると思う。特に科学的文章では。ただし、入力する文章はできるだけ短く区切る、意味に曖昧さのない文章であることが必要になろう。これは言葉としての美しさを損なうのでは、という不安もあるが、ボーダレス時代の膨大な情報を言語の壁をこえて活用するには機械翻訳は不可欠であり、そのためには言語そのものを翻訳しやすい構造に変えざるをえない状況が生まれるかもしれない。多少の（この表現には異議があるかも知れないが）犠牲を払っても、充分その意義がある時代になるのかもしれない。

あまりうれしくない予測ではあるが、翻訳者は将来、機械が理解できるような文章作りを助けるインターフェースの役割に甘んじることになるのだろうか。

著者の経歴

お茶の水女子大学理学部化学科卒業。三菱レーヨンに勤務した後、現在は、フリーランスの翻訳者として、科学、医学・薬学、経済等の翻訳を手がける。

【1993 年 5 月 AAMT ジャーナル No.3 より】

利用技術研究会ヒアリング

人間翻訳と機械翻訳

TCC 代表 安藤 進

人間翻訳と機械翻訳について、お話をしたいと思います。資料としまして、雑誌「bit」（共立出版）に寄稿した記事のコピー（1992 年 12 月号および 1993 年 4 月号）をお配りします。これは日本語と英語の翻訳で、人間翻訳と機械翻訳のレベルはどの程度違うのかというテーマでまとめたものです。

私が翻訳の世界に入りましたのは、今から約 15 年ほど前になります。最初の 5 年間は、富士通研究所で機械翻訳の研究開発をやっておりました。その後は、翻訳の発注側で日英翻訳のチェックを英語のネイティブスピーカ 10 人ほどの仲間といっしょにしておりました。それから、㈱十印翻訳部の部長として、主として英語から日本語に翻訳する仕事をしてきました。

そこで、本日は、これまでの経験を踏まえて私なりの問題を提起しまして、後はみなさんのご質問に答える形で話しを進めて行こうと考えています。

****

◆［日英翻訳について］

最初に、日本語から英語に翻訳する場合についてお話しいたします。

人間翻訳の例としては、(社）日本翻訳連盟の機関誌で、会員を対象にした翻訳講座を 3 年ばかりやってお

りまして、その講座の応募答案を人間翻訳の例として採用しました。

機械翻訳（MT）の例としては、現在商用化している MT システムによる翻訳の例を採用しました。MT システムの出力結果そのままと（一次結果）、辞書登録や原文を書き換えるなどの前編集をした後の翻訳結果（二次結果）を用意して、人間翻訳と比較してみたのです。

さて、応募答案の数は 40、答案作成者の年齢は 20 代から 50 歳代、国籍は日本がほとんどですが、米国から応募された人も数人おりました。原文はプリンタのマニュアルです。応募者の得意分野は必ずしも技術分野ではありませんし、翻訳経験もさまざまです。翻訳者を志して勉強中の人もいます。

翻訳の品質を評価する方法については、いろいろな観点があると思いますが、ここでは実務で通用する翻訳、つまり商売として通用する翻訳という観点で評価して見ました。

応募答案を、A、B、C、D の 4 つのレベルに分けて評価しました。A レベルの答案というのは、そのまま使えるというものです。残念ながら、A レベルのものはありませんでした。B レベルは、冠詞、前置詞など、英語のネイティブスピーカが原文を見ずに簡単に修正できるレベルです。B レベルの答案は、4 例、全体の 10%でした。C レベルは、原文を参照すれば比較的簡単な修正で通用するという意味です。このレベルが 8 例、全体の 20%になります。

したがって、実務的に通用するというのは全体の 30%位ということになります。

次に、機械翻訳の出力を評価しますと、一次結果は、利用価値はほとんどありませんでした。しかし二次結果はマァマァの出来でした。ちょっと主観的かもしれませんが 40～50%の出来とでも言いましょうか、その位のレベルになっていました。

人間翻訳でも、機械翻訳の二次結果のレベルに達していない答案が全体の 30%ぐらいありました。この結果からみると、人間翻訳が必ずしもよいというわけではないことがわかります。また、機械翻訳だからといって必ずしもまったくダメだというわけでもないと言えると思います。

◆［英日翻訳について］

英語から日本語に翻訳する場合について、お話しいたします。

実例としましては、私が翻訳会社におりました頃今から一年ほど前になりますが、実施したトライアルを紹介します。新聞広告を出し、翻訳者と機械翻訳の後編集者の募集をしたのです。全国から 600 人近い応募者がありました。トライアル問題の題材はコンピュータ、通信、半導体でした。

応募者に問題を送りましたところ、約半数の人がギブアップしてしまいました。つまり、問題の内容を見て自信をなくしてしまったと思われます。一般に、翻訳者になりたいと考えている人の数は多いのですが、技術的なバックグラウンドを持っている人が少ないということがわかります。それでも、300 人近い人が答案を返送してきました。その答案を A、B、C、D（不合格）という 4 つにわけて採点しました。その結果、A レベルはなし、B レベルが 5 名 C レベルは 15 名となりました。合否すれすれの中から比較的良いと思われる人も含めて、約 50 名の人と面接をしました。最終的には、約 20 名の方を採用することになりました。

◆［できる翻訳が足りない］

この数字を見ますと現在実務の世界で必要とされる翻訳力がある方は全体の 10%もいないということになります。英訳についてもほぼ同じことが言えます。これが、翻訳業界の現実だと思います。

もうひとつ例を紹介します。(社)日本翻訳連盟では、〈翻訳検定〉試験を実施していますが、この 1、2、3 級は先程の A、B、C とほぼ同じレベルです。毎回、200 人近い人が応募されますが、その中で、1 級レベルになる人はほんの数人です。2 級が 15～20%位です。3 級までで全体の 30%ぐらいという数字になります。

先ほど、私がかつて翻訳会社にいたといいましたが、当時、社内で見習い翻訳者を養成する制度がありまし

て、私がその指導を担当しておりました。当時の経験によりますと、少し頑張れば、また、適切な指導があれば伸びるということがわかりました。

残念ながら、今はその制度はなくなりました。最近は、翻訳会社でもその余裕がなくなってきているようです。当時の教え子は、今、第一線で活躍しています。

いずれにしましても、即実践で使える人を探すのはなかなか難しい。また、そのような方がいても、その数は大変少ないのです。

◆［機械翻訳の大衆化］

自然発生的に良い翻訳者が出てくるのを待ち望んでも現状を変えることは難しいのではないかと思います。

翻訳は特定のプロがやるものだという見方をする方が多いと思います。しかし、最近コンピュータ関連のマニュアルなど大量の技術文書の翻訳の仕事がありまして、従来の職人芸的な翻訳では、業界の要望に応えられないという事態になっております。

また、NIFTY-SERVE という商用ネットワークがあります。そこで、機械翻訳のサービスを提供しています。このサービスを利用すれば、簡単に機械翻訳を一般の人でも体験できるようになってきました。このような機械翻訳というツールを使って翻訳をやって行くというのがこれからの大きな流れではないかと思います。

****

（以下、出席者からの質問に対する応答）

◆機械翻訳における前編集の位置付けは？

英語から日本語へ翻訳する場合は、後編集に重点を置き、日本語から英語に翻訳する場合は、前編集をするというのが一般的だろうと思います。

いずれにしても、翻訳の品質は、前編集者または後編集者の力量に依存します。

実務的には、翻訳の全工程の中で前編集と後編集をどのように位置付けるのかが問題となります。翻訳工程を複雑にすると、それだけコストと時間がかかるからです。

前編集について見ますと、問題がたくさんあります。7～8 年前、日英翻訳の前編集者の指導をしたことがあります。その経験からいいますと、作業をする人は比較的こり性な人が多かったと思います。

例えば、かかり受けの括弧をいろいろ試しても思いどおりの訳文が得られないので、何度もやり直している。その結果、1 文に何時間もかかってしまうことがありました。

その結果、やればやるほど時間とコストがかかってしまうことになり、採算が合わなくなってしまうのです。また、機械翻訳の出力が悪いために、前編集者が無理をしてしまう。しかし、そのわりに成果が上がらないというケースも多かったと思います。

機械翻訳は高速で翻訳するというメリットがあるのですが、人間が介在するために、逆に、ブレーキになってしまう。

しかし、長い目で見れば、辞書に単語を登録したり、前編集をしながらシステムを使い込んでいくしかないと思います。

◆前処理者の要件は英語力なのか専門分野の経験なのかどちらが重要か？

結論から言えば、両方ともある程度は必要だと思います。私の経験から言うと、英語力は、例えば、英検の準 1 級ぐらいがちょうどいいと思います。専門分野の知識としては、その分野の基礎知識があれば十分だと思います。

例えば、コンピュータ関連では、CPU とか I/O とかいう用語のイメージがわかる程度でいいと思うのです。でも、コンピュータというものを見たことも聞いたこともないという人では無理だと思います。

さて、翻訳業界には神話がたくさんあります。そのひとつが「専門知識」なのです。翻訳を発注する側、また翻訳会社でもそうですが、採用試験のときに経歴書を見ます。例えば、コンピュータ関係ですと、情報処理系の学部を出た人、プログラミングの経験者だと無条件に安心してしまう人が多い。

しかし、私の経験から言うと、なにか違うという気がしています。翻訳をするには、ある程度の専門知識は必要ですが、技術者指向の人と翻訳者志向の人とではやはりタイプが違うと思います。

専門知識があるか無いかでアプリオに裁断するのは実情に合っていないと思います。

むしろ、しっかりした明快な文章が書ける力のある人の方が適していると思います。英語力も専門知識もそこそこという人の方が原文をしっかり読んで明快な訳文を作成できるのです。

英語力のすぐれている人は、つい自分の英語力を過信して辞書を引く手間を惜しむ傾向があります。また、専門知識のある方は、自分の古い知識で判断しようとする傾向があります。

いわゆる産業翻訳というのは、ほとんど現代が対象です。執筆者も同じ時代に生きている人間です。翻訳の読者も同時代の人です。例えば、マニュアルの翻訳では、一般に対象読者は一般のユーザーですから、専門技術者ではない人の方がわかりやすい訳文が作れると思うのです。

また、専門用語について云いますと、例えば、科学技術関連では、何十万語という用語があります。医学分野でも何 10 万、いや何 100 万語とかの用語があるといいます。人間はとても覚え切れない。

したがって、専門用語はコンピュータに任せて、人間は考えることに専念する。どう表現すれば、原文の意味が読者に明快に伝わるか、という方向に向かっていかなければいけないと思います。さて、MT システムは、それぞれのメーカによって多少違いますが、今のところ、人間が相当手直しをしなければ通用しません。ただ、人間の手が加わると人間のレベルを越えられないということに注意していただきたいのです。

とくに力の低い人が後編集した場合には総じて機械翻訳のレベルより悪くなる。MT システムでできないところは手が出ないからそのままになります。一方、MT の方がよく知っていてよくできている部分を手直ししてしまうので、結果的には改悪してしまうのです。

ですから、機械翻訳の前編集や後編集をする人は翻訳ができる人でなければならないのです。翻訳のできない人には無理な仕事なのです。MT システムを利用すれば、素人でもできるというのは、現代の間違った神話のひとつだと思います。

◆英検の資格保持者と翻訳品質との関連は？

これも神話のひとつです。「おまえは英検の 1 級をもっているから、これを翻訳してくれ」とか、「あの人は英検 1 級をもっているから翻訳は安心できる」とかいう人が意外に多いのですが、ほとんど偏見といっていいと思います。

むしろ、先ほどお話ししましたように、自称英語のできる人の場合、悪い面の方が目につくような気がします。確かに、英語力の一般的な力は相当なものなのですが、つい自分の力を過信してしまい、辞書を引く労力を省いてしまいがちなのです。

多分試験制度にも原因があるのかもしれません。従来の試験では、試験場という密閉された場所で、しかも辞書の持ち込みも禁止され、頭のなかにあるものが試されるのです。

ところが、例えば、（社）日本翻訳連盟が実施している〈ほんやく〉検定という試験がありますが、この試験では、パソコン通信でも受験できるように辞書、参考書、ノートなど何を見てもよいという制度になっています。

大切なのは、知識ではなく、理解力と表現力だというわけです。まだ、小規模な試験なので、一般の人にはあまり知られていませんが、今後は広まっていくと思います。

◆これからの翻訳の理想像は？

個人的には、執筆者と翻訳者とリライタが一堂に集まって話合いながら訳していくのが理想だと思います。私は、現在、フリーの翻訳者としてやっているわけですが、執筆者がわかる場合、電子メールで質問をしています。ほとんど、すぐ返事がきます。

質問の内容は、原文の解釈が複数ある場合、原文の誤りと思われる箇所、対象読者が違うので、こんな意味で訳したいがどうだろうか、...といったことです。

日英翻訳の場合、原文は日本人を対象にして書かれています。これを英語に翻訳する場合、対象読者が、例えば、米国人、英国人、オーストラリア人なのか、あるいはアジアの人なのかで表現が違ってくるからです。

翻訳者は、原文の最初の読者になります。また、リライタは2番目の読者ということになります。この人たちに理解できないようでは、一般読者はまず理解できないはずです。

英日翻訳でも同じことが言えます。英語の原文はあくまで、英語を母国語にしている人を対象にして書かれています。これを日本語にする場合、読者は日本人になります。

いわゆる直訳では、なかなか意味がわからない。逆説的な言い方をすれば、英語のわかるひとならわかるという訳文が多いのです。

さて、産業界全体から見ると、例えば、マニュアルの翻訳では､ダンボール何箱分という大量になります。通常は半年とか1年はかかっていたのですがそれでは間に合わない。

タイムリーな翻訳でないと情報としての価値がなくなります。また、翻訳の個性が強いと、全体としての統一がとれなくなります。

そこでどうしても､従来の職人芸的なやりかたでは、時代の要請に応えられなくなっています。やはり、言語処理技術を活用しなければ、問題は解決できないのです。

ただ、現在商用化されているシステムは、どうもこのような現実を踏まえていないようです｡したがって、実用化が足踏みしているといった状態だと言えます。

◆大量翻訳のあるべき姿は？

現状を踏まえると、ローカルな柔らかいネットワークでグループのコミュニケーションをとりながらやっていくという形態がいいと思います。文体、用語、などは、一括して管理しない限り、大最の翻訳を短期間でしかも一定の品質を維持するのは困難です。また、最近の英文は必ずしも明快な文章ばかりではありませんので、意味不明な場合の対処の仕方も大切になります。実務の世界では、いつも時間とコストが問題になります。原文に忠実な翻訳を原点にして、読者にわかりやすい翻訳まで、さまざまなレベルがありますが、そのレベルに応じて翻訳料金も上がるというシステムのコンセンサスが必要だと思います。

翻訳者とチェッカーの役割分担、発注者と翻訳会社の役割分担などを明確にしていく必要もあります特に、発注側に翻訳の品質を客観的に判断できる方があまりいないことも大きな問題です。

◆翻訳は技術的バックグランドがなくても可能か？

結論から言えば、可能だと思います。先ほどもいいましたが、この業界には困った神話があります。専門技術者と翻訳者とはタイプが違うと思います。

例えば自動車の翻訳をする人は、整備士の免許をもっている人でないとできないのか、あるいは、運転免許証をもっていなければできないのか。そんなことはないと思います。

同じ車といいましても、みなさんが利用しているマイカーから見る世界と、タクシーやトラックなどの職業として運転している人たちの世界はかなり違います。当然、仲間内でしゃべる言葉も違います。

ところが、翻訳という仕事は、自分たちの狭い世界から別の人々に情報を伝えることなのです。

極端な言い方をすれば、技術者は単眼、翻訳者は複眼、であると思います。

したがって、例えば、技術分野の翻訳者にとって必要な資質は、技術者と話しができることなのです。技術者がうまく表現できないこと、原文の表現のままでは読者に意味が十分伝わらないと思われる部分を質問できる力、あるいはそういう方向性をもっている人が適しているのです。

今後は、翻訳を発注する側も受注する側も、お互いに納得できる座標軸が必要になると思います。良い翻訳と悪い翻訳を区別する基準も必要です。そうしてはじめて今後の展望が開けると思うのですが、まだ、時間がかかると思います。そのためにも、機械翻訳協会とか、翻訳連盟のような公的な機関が率先して試案を世間に提示していかなければならないと、考えております。

（この原稿は、5 月 12 日利用技術研究会で講演した記録テープをもとに筆者が書き直したものです）

編集後記

編集後記

隅田 英一郎
AAMT ジャーナル編集委員会

1. LLM は泡と消える流行語じゃない

巷間を賑わす **LLM** とは、大規模なコーパスに基づいて次単語予測をするように学習されたニューラルネットワークであり、対話、機械翻訳等の応用と、構文解析、意味解析、文脈解析等の基礎との両面で（多くの人の想像を超える勢いで）性能を更新し続けています。

中国を除くアジアは、**LLM** の開発・活用において現在劣後気味ですが、インド、シンガポール、日本は国をあげて盛り返そうとしています。開発サイドの方々は伸び代が大きいと見ています。活用サイドの多くの方々は **LLM** との付き合い方に悩んでいます。そんな中、日本において **LLM** の開発・活用を加速すべく、日本政府は「AI 事業者ガイドライン（第 1.0 版）」を公表しています（←ご参考の情報です）。

AAMT は機械翻訳や **LLM** の最新情報をジャーナルやセミナーや年次大会で提供しています。**LLM** は泡と消えないことを映して本号においても **LLM** に関連する話題も多いです。

2. 本号記事の振り返り

巻頭言 AAMT 会長である安達久博様に「要件定義」という題でご寄稿をいただきました。16 年前の事例からはじめて、品質、スピード、コストに関して翻訳の関係者が共通理解を持ち、それらのバランスについて同意することの重要性を説かれています。ISO 規格にも触れつつ翻訳産業では今後は機械翻訳だけでなく **LLM** の活用も含める形になると指摘されています。

AAMT 長尾賞 AAMT では、本協会の創立者名を冠した賞を設けています。賞の趣旨はユニークで「機械翻訳システムの実用化の促進および実用化のための研究開発に貢献した個人あるいはグループを表彰します。いわゆる学会の論文賞や発表賞といった学術賞ではなく、たとえば、高性能の機械翻訳システムを商品化した、機械翻訳システムを使った新しいサービスを開始した、といった貢献を対象とします。もちろん学術的にも意味のある成果を除外するものではありません。」となっています。今年の受賞は「**LLM**-jp によるオープンかつ日本語に強い大規模言語モデルの研究開発」(https://aamt.info/news/nagao-2/)です。

LLM の研究開発には膨大な計算パワーが必要であることから外国組織の寡占状態になっています。高い性能を否定するのは馬鹿げていますが、一方で、固有文化が歪曲されたり、経済安全保障上のリスクもあり、**LLM** の副作用が懸念されています。これらの問題意識から、日本でも **LLM** を独自に作ろうと 2023 年 5 月に立ち上げられた勉強会 **LLM**-jp は 1 年半で二桁大きな規模(数十人から 2 千人弱)になりました。この大勉強会は、期待を裏切らず、オープンな形で次々に成果を出し続けています。同活動のリーダーである黒橋禎夫様に熱い思いを語っていただきました。

AAMT 長尾賞/学生奨励賞 AAMT は、機械翻訳研究に携わる優秀な研究者と判断される学生を表彰するために長尾賞の拡張として学生奨励賞を 2014 年に新設しました。近年、推薦数、受賞数ともに増えており頼もしく感じるところです。二宮崇長尾賞委員長の報告から引用します(https://aamt.info/news/nagao-student/)。

① 毛卓遠様、Breaking Language Barriers: Enhancing Multilingual Representation for Sentence Alignment and Translation、「低資源

Editor's note
Eiichiro SUMITA
AAMT Editorial Board

言語も含めた多言語機械翻訳を視野に、多言語埋め込み表現の学習、多言語機械翻訳のためのモデル構成と学習方法」

② 福田りょう様、「Towards Streaming Speech Translation for Real-world Scenarios、長い音声ストリームに対するオンライン機械翻訳のために、話し言葉機械翻訳や音声翻訳のための入力音声の自動分割」

今年の 2 件の受賞者に詳しい解説を依頼しましたのでお楽しみください。

長尾賞と長尾賞学生奨励賞は、毎年 3 月から 4 月第一週の間、賞に相応しい候補者・グループの推薦を受け付けております。上の 3 つの記事をご参考に次年度の推薦をご検討いただけましたら、コミュニティを更に盛り上げることに繋がると思いますのでよろしくお願いいたします。

AAMT 若手翻訳研究会 翻訳産業を更に大きく発展させるためには、次世代の研究者・開発者・翻訳者・利用者をコミュニティ全体で育成するのが大事です。これを怠るとやがて活力を削がれ衰退してしまう虞があります。若手育成のための AAMT 施策の一つに AAMT 若手翻訳研究会を 2024 年 3 月 22 日に AAMT セミナーとして実施しました。「若手による短時間のプレゼンとセミナー視聴者による評価」のイベントで、初開催でしたので改善点もありますが、斬新な発表をお愉しみいただけたと思います。中澤敏明理事に「第１回 AAMT 若手翻訳研究会 開催報告」としてご寄稿いただきました。実施に至る経緯や多数の参加を得て成功したことが分かります。また、下記の 5 件の受賞者一人一人に記事を書いていただきました。

●最優秀賞

① サブセット探索を用いた高速な kNN ニューラル機械翻訳、出口祥之様（NAIST）

●優秀賞

② 日英間の機械翻訳による受容化と異質化について、木内晶基様(東京工業大学)

③ 人手翻訳から MTPE へ: 一翻訳者の所感、海老原仁美様（レッドハット株式会社)

④ キャラクターの性格と人間関係情報を付加した映像翻訳データセットの構築、大嶽匡俊様（東京大学)

⑤ 大規模言語モデルに対する対訳データを用いた継続事前訓練による翻訳精度評価 近藤海夏斗(筑波大学)、

温故知新 紙で発行されてきた会誌を PDF 化し公開したことが内山将夫編集委員長が発案したシリーズ記事「温故知新」につながりました。シリーズ名通り、過去から学べることも多く面白いです。今回は後藤功雄編集委員が過去記事から二つ選んで転載しています。

① 1993 年 2 月の「翻訳の現場から、機械翻訳時代の翻訳者の役割は？（元吉宏子著）」の超要約は「「コンピュータに原稿を読み込ませるだけで翻訳が仕上がれば」と想像することがあるが＜中略＞翻訳者は将来、機械が理解できる文章作りを助ける役割に甘んじることになるのだろうか。」です。翻訳アルアル満載ですのでお楽しみください。

② 1993 年 5 月の「利用技術研究会ヒアリング『人間翻訳と機械翻訳』(安藤進著)」の内容は、同氏のご講演と、それに対する問い、例えば「前処理者の要件は英語力なのか専門分野の経験なのかどちらが重要か？」、とその回答をまとめたもので、読み応えがあります。

3. 編集後記の振り返り

画竜点睛したのは人間ですが、本稿は「**LLM** が相当量書きました」。**LLM** に巧みに書かせるのに紆余曲折したとはいえ **LLM** 抜きでは本稿の完成はなかったことから、編集委員も **LLM** の言語能力に感心したところです。

LLM がアジアを含む世界の人々の生産性を上げ人類を飛躍させると今回改めて確信しました。

AAMTジャーナル「機械翻訳」No. 81

【 発 行 日 】2024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