低資源言語のニュース機械翻訳のための LLM を用いた合成対訳データの生成

伊藤 均　白井 尚登　衣川 和尭　美野 秀弥　河合 吉彦
NHK 放送技術研究所
{itou.h-ce, shirai.n-hk, kinugawa.k-jg, mino.h-gq, kawai.y-lk}@nhk.or.jp

概要

低資源言語の機械翻訳モデル構築における課題として，十分な量の対訳データの確保が挙げられる．さらに，対象ドメインが限定されている文の翻訳などの場合，対訳データだけでなく原言語側の単言語データを確保することさえも困難な場合がある．本研究では，対象ドメインの原言語側の単言語データが少量しかない状況下で，学習データ量不足の課題を解決するための大規模言語モデル（LLM）を活用した多様で高品質な合成データ生成手法を提案する．提案手法の効果を確認するために，タイ語から日本語へのニュース翻訳の実験を行った．提案手法で合成したデータを用いて LLM をファインチューニングしたニュース機械翻訳器は，タイ語から日本語へのニュース翻訳タスクにおいて従来手法を上回る性能を達成し，BLEU 値を 19.9 ポイント改善した．

1　はじめに

近年，大規模言語モデル（LLM）は汎用的なタスクで人間に近い性能を達成するなど大きな注目を集めている [1][2]．LLM の性能を高めている大きな要因の一つは大量の学習データであり，英語以外の低資源言語を対象としたタスクについては依然課題が残る [3]．

汎用的な LLM を一つのタスクに特化する方法として，対象のタスクのデータで事前学習済みの LLM を学習するファインチューニングがある [4][5]．翻訳タスクへのファインチューニングにおいては原言語と目的言語の対訳データが必要であり，低資源言語において十分な量の対訳データ収集は課題である．学習データ不足の対処の１つにデータ拡張があり，対訳データの拡張として，ピボット翻訳を用いた手法 [6][7] や逆翻訳を用いた手法 [8] がある．ピボット翻訳は原言語側のデータを用意し英語など高資源言語の中間言語（ピボット）への翻訳を介して目的言語へ翻訳する手法であり，逆翻訳は目的言語側のデータを用意し原言語へ翻訳する手法である．両手法は，ともに，十分な量の単言語データがあることが前提となっている．

ニュース翻訳など特定のドメインの翻訳精度向上には，いかに対象とするドメインに近い対訳データでファインチューニングできるかが鍵となる．例えば，「特定の報道機関のスタイルで書かれたタイの固有名詞を数多く含むタイ語のニュース」の機械翻訳器の構築には，「タイ語のニュース」ドメインの対訳データでは十分でなく，より詳細な限定化が必要となる．このように，ドメインを限定すればするほど対象ドメインの単言語データすらも収集することが難しくなる．

本研究では，低資源言語の機械翻訳タスクにおいて，対象とするドメインの対訳データと単言語データ双方の不足に対処するため，LLM を用いて少量の原言語側の単言語データから大量の合成対訳データを生成する手法を提案する．本稿における貢献は以下のとおりである．

- LLM を用いて対象ドメインの多様な原言語データを生成し，かつ，生成した原言語データを日本語に翻訳するための，新しいプロンプト設計と生成フローを提案した．
- タイ語から日本語へのニュース機械翻訳実験により，提案手法による翻訳性能の向上を確認した．

2　合成対訳データの生成

Pengpun ら [9] は，低資源言語において LLM の能力を高めるための効率的な合成データ生成には，データに多様性，流暢性，文化的背景の３つの特性

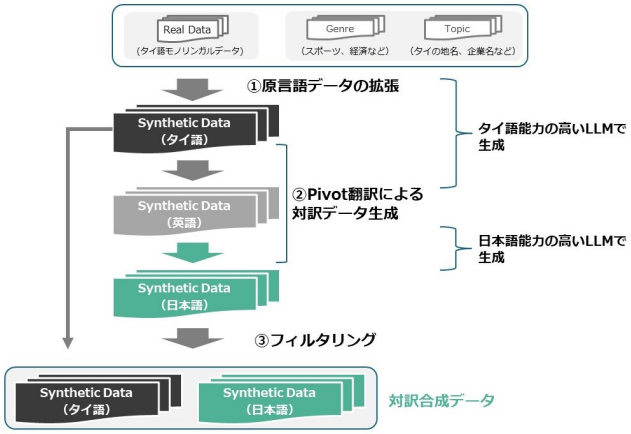

図 1: 合成対訳データ生成フロー

が必要であるとしている．本研究では，上記の研究で触れられていない翻訳タスクに焦点を当て，多様性，対訳品質，翻訳対象とのドメイン一致，の３つの要素を満たす合成対訳データの生成手法を提案する．

提案手法の合成対訳データの生成フローを図 1 に示す．生成手順は以下である．

1. 翻訳対象とドメインが一致する少量の原言語データから合成原言語データを増強する．
2. 合成原言語データから，英語を中間言語としたピボット翻訳により合成対訳データを生成する．
3. 重複や低品質な合成対訳データを削除する．

2.1 原言語データの拡張

翻訳対象とドメインが一致する少量の原言語データ（リアルデータ）を基に，LLM を用いてドメインが等しくなるように原言語であるデータを増強する．データの増強には翻訳性能を高めるための３要素のうちの「多様性」を満たすデータとなるよう，対象ドメインのリアルデータと，ニュースに関するジャンルとトピックを付録 A.1 のプロンプトへ様々な組み合わせで代入し，多様なデータを生成する．ジャンルとトピックは対象ドメインに関するものを抽出するため，リアルデータを付録 A.2 のプロンプトに代入することで各文のニュースジャンルとトピックを取得する．

本節で用いる LLM は原言語に特化したものを用いる．

2.2 合成対訳データの生成

低資源言語の対訳データを LLM を用いて直接生成せずに，英語のような高資源言語を中間言語としてピボット翻訳し対訳データを合成することで翻訳品質を向上させる．合成対訳データの生成には，対訳品質を高めるために 2 種類の LLM を用いる．2.1 節で生成した合成原言語データ各文を原言語ー中間言語の翻訳精度の高い LLM を用いて高資源言語に翻訳し，その次に，中間言語ー目的言語の翻訳精度の高い LLM を用いて翻訳することで対訳データを生成する．

2.3 フィルタリング

ジャンルとトピックの組み合わせによっては適切な対訳データを合成することができず，リアルデータのドメインを無視してトピックの単語の対訳のみが生成されるなど，生成結果の一部に重複が生じることがある．そこで，原言語と目的言語の両方がそれぞれ完全一致している対訳データについては一方を削除し，重複を除く．加えて，対訳データの中で英単語が 10 単語以上連続している箇所については不要な湧き出しであると判断し，その箇所を削除したものを対訳として採用する．

3 実験

3.1 実験設定

タイ語ー日本語のニュース翻訳タスクで翻訳実験を行った．提案手法で生成した合成データを用いて Llama-3-8B-Instruct[2] をファインチューニングし，タイ語のニュース記事を日本語へ翻訳するタスクの翻訳性能を比較した．本稿では，原言語データが不足している状況を再現するため，AI for Thai プラットフォーム[1)]で公開されている，場所や組織名などの固有表現を含んだタイ語モノリンガルデータ（NE コーパス）の一部を実験に用いることとし，学習データ 1,400 文，開発データ 550 文，評価データ 550 文をランダムに抽出した．学習データ 1,400 文のうち 1,200 文はジャンル・トピック生成のための入力文として用いた．各文をジャンル・トピック生成プロンプトへ代入し，各出力結果の重複を削除することで 333 ジャンル・411 トピックを取得した．

1） https://aiforthai.in.th

残る 200 文はリアルデータとして原言語データの増強に用いた．リアルデータ 1 文ごとにジャンルとトピックの組み合わせ 100 組を用いて 100 文の擬似タイ語データを生成し，リアルデータ 5 文ごとにジャンルとトピックの組み合わせを重複のないよう入れ替え，計 200 文と 4,000 組から 2 万文へ拡張した．

合成原言語データ生成用プロンプトとジャンル・トピック生成用のプロンプトはそれぞれ付録 A に掲載した．LLM はタイ語データの拡張と英訳，ジャンルとトピックの生成に Typhoon-v1.5x-70b-instruct-awq[10]，日本語訳に Qwen2.5-72B-Instruct[11] を用い，タイ語データの拡張時のみ temparature を 0.9 に設定した．ファインチューニングには QLora[5]（lora rank 8，lora alpha 16）を採用し，batch size 32，learning rate 1e-4 で 10 epoch 学習した後，翻訳精度の評価尺度として用いられる BLEU[12] で各モデルの性能を比較し，開発データの BLEU が最も高かったパラメータを採用した．最適化手法には AdamW[13] を用いた．

本稿では，まず 1 万文の合成対訳データでファインチューニングした各モデルの翻訳性能を比較し，最も性能が高かったハイパーパラメータと手法の組み合わせについて，ファインチューニングに用いる合成対訳データを 2 万文に増やした際の性能を調査した．比較手法はそれぞれ，ファインチューニング前の Llama-3-8B-Instruct，NE コーパスから人手で生成した対訳 100 文でファインチューニングしたモデル (NE100)，English Wikinews[2] の人手多言語対訳コーパスである ALT コーパス [14] のタイ語と日本語の対訳データのうち 1 万文でファインチューニングしたモデル (ALT10k)，GPT-3.5-Turbo [15] でタイ語から日本語へ直接推論した翻訳結果 (GPT-3.5-Turbo) と，ファインチューニング前の Llama-3-8B-Instruct による中間言語を英語としたタイ語から日本語へのピボット翻訳の結果 (Llama3-8B-Instruct Pivot) とした．

3.2 実験結果

実験結果を表 1 に示す．正しい翻訳結果をほとんど出力することができなかった Llama3-8B-Instruct モデルを提案手法で生成したデータでファインチューニングすることにより，BLEU 値が向上することを確認した．リアルデータとして用いた NE コーパス 100 文の人手対訳データによるファインチューニングでは性能向上がほぼ見られなかったことからも，提案手法の合成データの有効性が示された．また，提案手法は人手翻訳の ALT コーパス 1 万文でファインチューニングしたモデルよりも性能が高いという結果になった．この性能の差はデータのドメインの異なりが起因していると考えられる．NE コーパスはタイ現地のニュースを中心としたコーパスであるのに対し，ALT コーパスはアメリカを中心とした世界のニュースを扱ったものであり，ニュースという点では共通しているものの，扱うニュースの話題や内容については翻訳対象のドメインとは異なる．付録の表 5 に翻訳結果の比較を示す．ALT コーパスで学習したモデルの翻訳結果は流暢ではあるが，「県知事」という用語や、仏歴（原言語文は西暦 2021 年を仏歴 2564 年で表記しており，日本語へ翻訳する際には仏歴から西暦への変換が必要）など，タイに関するニュースで用いられる用語やタイ特有の表現については学習データが不足しているため，翻訳に失敗している．提案手法は，タイのニュース記事を増強して学習しているため，基にしたリアルデータが少量であってもタイに関する用語を翻訳することができている．このことから，翻訳性能向上に必要なデータの要素として翻訳対象とのドメインの一致が必要であることが示された．

表 1: 実験結果

手法	BLEU
Llama3-8B-Instruct	0.55
Llama3-8B-Instruct Pivot	1.10
Fine-Tuning (NE100)	1.53
Fine-Tuning (ALT10k)	15.85
GPT-3.5-Turbo	16.45
Fine-Tuning (Synthetic data 10K)	**19.82**
Fine-Tuning (Synthetic data 20K)	**20.48**

3.3 アブレーションスタディ

本稿ではリアルデータ 5 文ごとにジャンルとトピックの組み合わせを変えて生成したが，組み合わせを変える単位を変化させた場合の精度を比較した．100 文のリアルデータから 1 万文の合成データを生成する際，何文ごとにジャンルとトピックの組み合わせ (GTpair) を変えるのが最適であるか，BLEU を比較した結果を表 2 に示す．

実験結果から，1GTpair あたり 5 例ずつ生成，つ

2） https://en.wikinews.org

表 2: ハイパーパラメータの調整

生成数/GTpair	GTpair 数	BLEU
1	10,000	19.10
2	5,000	19.32
5	2,000	**19.82**
10	1,000	19.13
100	100	18.37

表 3: 合成対訳データの生成に用いた LLM の違いによる性能比較

model	BLEU
only Typhoon	13.67
only Qwen	18.93
Typhoon+Qwen (提案手法)	**19.82**

まりリアルデータ 100 文から 1 万文を生成する際に 2,000GTpair を用いて生成すると最も性能が高くなることがわかった．以降の実験では各モデルにこのハイパーパラメータを採用し，合成対訳データを生成する．

次に，複数の LLM を用いたことによる効果を確認するため，LLM を切り替えない場合の翻訳性能との比較を行った．比較手法は原言語データ拡張と Pivot 生成をともに Typhoon-v1.5x-70b-instruct-awq で生成したデータで Llama-3-8B-Instruct をファインチューニングしたモデル (only Typhoon)，ともに Qwen2.5-72B-Instruct で生成したデータで Llama-3-8B-Instruct をファインチューニングしたモデル (only Qwen) である．結果を表 3 に示す． 実験結果から，モデルを使い分けることによる効果を確認した．提案手法では目的に応じてモデルを使い分けることで，対訳品質の高い合成データを生成することが可能となり，翻訳性能が向上することが示された．

最後に，タイ語データを拡張する際のプロンプトに，リアルデータ，ジャンル，トピックを代入したことによる効果を確認するための実験を実施した．比較手法は，(1) プロンプトに固定の 1 例のみリアルデータを例示し，ジャンルとトピックを代入しない手法，(2) リアルデータを代入し，ジャンル，トピックについては使用しない手法，(3) リアルデータとジャンルとトピックを代入するが，ジャンルとトピックの生成にリアルデータを参照しない手法，である．(3) の手法では，ジャンルとトピックの生成時にリアルデータなどの参照文を与えることなく，ニュースのジャンルを生成したものをジャンル，タイの固有名詞を生成したものをトピックとしてプロンプトへ代入し，合成対訳データを生成した．(3) のジャンルとトピックの生成には Qwen2.5-72B-Instruct を用い，それぞれ 248 ジャンル，227 トピックを取得した．結果を表 4 に示す． 実験結果から，リアルデータと，2.1 節の要領で生成したジャンル・トピックを変数として代入することでそれぞれ翻訳性能の向上を確認した．また，ジャンルとトピックについてはリアルデータの情報を参照せずに生成したものを用いると性能が低下することがわかった．これは，適切でないジャンルとトピックを設定したことにより翻訳対象のドメインからずれてしまったことが理由であると考えられ，翻訳性能を高めるためには学習データと翻訳対象のドメインの一致が重要であることが示唆された．

表 4: データ拡張時のプロンプトの違いによる性能比較

リアルデータ	ジャンル・トピック	BLEU
✓(1 文のみ)	-	16.80
✓	-	19.06
✓	✓(リアルデータ非参照)	18.66
✓	✓(提案手法)	**19.82**

4 結論

本稿では，低資源言語におけるニュース機械翻訳のための合成データ生成について，翻訳対象のドメインの原言語データを拡張し，対訳データを生成する手法を提案した．提案手法は，原言語データを増強するプロンプトと，複数 LLM を用いた対訳データ生成フローで構成され，少量の原言語モノリンガルデータから多様で高品質な合成対訳データの生成を可能とする．生成された合成データを用いてファインチューニングすることにより汎用的なモデルを翻訳対象のドメインに適合させることが可能となり，翻訳性能を高める効果があることが示された．今後はより大規模なモデルや大規模なデータを用いた場合の効果の検証を進めていく．

謝辞

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

本研究は National Electronics and Computer Technology Center (NECTEC) が作成した「NE-Corpus」を使用しています（CC BY-NC-SA 3.0 TH）．詳細は次のリンクをご参照ください (https://creativecommons.org/licenses/by-nc-sa/3.0/th/).

参考文献

[1] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, and et al. Anna-Luisa Brakman. Gpt-4 technical report, 2024.

[2] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, and Bobbie Chern et al. The llama 3 herd of models, 2024.

[3] Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Maxamed Axmed, Kalika Bali, and Sunayana Sitaram. Mega: Multilingual evaluation of generative ai, 2023.

[4] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

[5] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms, 2023.

[6] Anna Currey and Kenneth Heafield. Zero-resource neural machine translation with monolingual pivot data. In Alexandra Birch, Andrew Finch, Hiroaki Hayashi, Ioannis Konstas, Thang Luong, Graham Neubig, Yusuke Oda, and Katsuhito Sudoh, editors, **Proceedings of the 3rd Workshop on Neural Generation and Translation**, pp. 99–107. Association for Computational Linguistics, 2019.

[7] Trevor Cohn and Mirella Lapata. Machine translation by triangulation: Making effective use of multi-parallel corpora. In Annie Zaenen and Antal van den Bosch, editors, **Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics**, pp. 728–735, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[8] Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In Katrin Erk and Noah A. Smith, editors, **Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 86–96. Association for Computational Linguistics, 2016.

[9] Parinthapat Pengpun, Can Udomcharoenchaikit, Weerayut Buaphet, and Peerat Limkonchotiwat. Seed-free synthetic data generation framework for instruction-tuning llms: A case study in thai, 2024.

[10] Kunat Pipatanakul, Phatrasek Jirabovonvisut, Potsawee Manakul, Sittipong Sripaisarnmongkol, Ruangsak Patomwong, Pathomporn Chokchainant, and Kasima Tharnpipitchai. Typhoon: Thai large language models, 2023.

[11] Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, Kai Dang, Yang Fan, and Yichang Zhang et al. Qwen2.5-coder technical report, 2024.

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

[13] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.

[14] Hammam Riza, Michael Purwoadi, Teduh Uliniansyah, Aw Ai Ti, Sharifah Mahani Aljunied, Luong Chi Mai, Vu Tat Thang, Nguyen Phuong Thai, Vichet Chea, Sethserey Sam, et al. Introduction of the asian language treebank. In **2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA)**, pp. 1–6. IEEE, 2016.

[15] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, and Tom Henighan et al. Language models are few-shot learners, 2020.

表 5: 翻訳結果の比較

model	出力 (Reference)
Reference	アムナートジャルーン県知事は、2021 年度プラパトムチェーディー奨学金の奨学金授与式を開催した。
Fine-Tuning (ALT10k)	アンジャーンの市長は、2564 年のプログラムで、プログラムのための教育基金を提供する。
GPT-3.5-Turbo	父親のアムナート・チャロエン・マハーヴィロヤーンは、2021 年にプラタムジェーディー財団の奨学金を授与するための式典を開催しました。
Fine-Tuning (Synthetic data 10K)	アムナットチャレーン県知事は、2021 年のプラパトムチェディー財団の奨学金授与式を主催しました。

付録

A　プロンプト

A.1　合成原言語データ生成プロンプト

Prompt

You are a news reporter specializing in {*GENRE*} in Thailand. Based on the following sentence structure, please write a brand new Thai {*GENRE*} news article by using the word '{*TOPIC*}' without changing the length of the sentence. You don't have to use any of the original words, just rewrite the whole sentence so it makes sense.
OUTPUT ONLY THE REWRITTEN SENTENCE, AND NOTHING ELSE. DO NOT ADD ANY ADDITIONAL NOTES OR INFORMATION.!!!
Input: {*REAL DATA*}
Output:

A.2　ジャンル・トピック生成プロンプト

Prompt (to extract Genre)

You are a news reporter. If news genres were divided into 300 categories, what genre would this news be? Please answer within 3 English words, not sentences.
OUTPUT ONLY THE GENRE, AND NOTHING ELSE. DO NOT ADD ANY ADDITIONAL NOTES OR INFORMATION.!!!
Input: {*REAL DATA*}
Output:

Prompt (to extract Genre)

Extract the one most symbolic proper noun related to Thailand (person's name, place name, company name, facility name, etc.) from the following sentence. Be sure to extract only one proper noun.
Do not extract proper nouns that are not related to Thailand or non-proper nouns (dates, times, numbers, quantities, etc.).
OUTPUT ONLY THE PROPER NOUNS, AND NOTHING ELSE. DO NOT ADD ANY ADDITIONAL NOTES OR INFORMATION.!!!
Input: {*REAL DATA*}
Output: