100× Monolingual Data Augmentation Using LLMs to Build a Parallel Corpus for Machine Translation

Hitoshi Ito*, Naoto Shirai*, Kazutaka Kinugawa*, Hideya Mino*, and Yoshihiko Kawai*
* Japan Broadcasting Corp., Japan
E-mail: {itou.h-ce, shirai.n-hk, kinugawa.k-jg, mino.h-gq, kawai.y-lk}@nhk.or.jp Tel: +81-3-5494-3377

***Abstract*—The aim of this study is to address the generation of high-quality synthetic data for domain adaptation in news machine translation. Effective machine translation requires training data that includes domain-relevant characteristics such as topical focus and writing style. Large language models (LLMs) can generate both bilingual and monolingual data; however, producing diverse monolingual data that closely aligns with the target domain remains a challenge. In this work, we explore the feasibility of using LLMs to generate synthetic data by attempting to create 100 times parallel sentence pairs with the original monolingual sentences. We leverage domain-specific keywords and combine multiple LLMs in a generation pipeline. Experimental results on Thai–Japanese news translation confirmed that the synthetic data generated by the proposed method significantly improved the translation performance of general-purpose models. These results also revealed that the use of appropriate keywords and the strategic switching between multiple LLMs play a crucial role in outperforming conventional methods.**

I. Introduction

Although large language models (LLMs) have demonstrated human-like performance across a wide range of tasks [1]–[3], there is considerable scope for the introduction of additional approaches to effectively enhance their capabilities in domain-specific contexts. Instruction tuning is a key approach for adapting LLMs to perform specific tasks effectively [4]. It has been reported that instruction tuning improves LLMs' ability to target tasks and domains [5], [6]. However, instruction tuning for translation tasks requires high-quality bilingual data, which can be difficult to obtain, particularly in low-resource domains.

Methods that address the lack of training data in translation tasks include using lexical information [7], augmenting data via word paraphrasing [8], and synthetic bilingual data generated from monolingual data in the target domain. Pivot translation [9] and back translation [10], [11] are used techniques for generating bilingual data from monolingual corpora. In pivot translation, data are prepared in the source language and translated into the target language via an intermediate language (pivot) in a high-resource language such as English. In back translation, data are prepared in the target language and translated into the source language. However, both approaches rely on the availability of substantial monolingual data, which is not always guaranteed.

LLMs offer the potential to generate large volumes of monolingual data. However, existing approaches often focus on broad knowledge acquisition [12]–[14] and do not effectively generate domain-specific data with limited topical scope, such as "regional news in Thailand." In translation tasks, some methods aim to expand monolingual data based on the original text [15], [16]. However, the extent of expansion achieved by these methods is typically limited to only a few times the size of the original dataset. We address the challenge of scaling up source language data by a factor of 100 by generating 10,000 bilingual sentence pairs from only 100 original source sentences, which far exceeds the expansion scale of conventional methods.

The proposed method leverages LLMs to generate domain-adaptive synthetic bilingual data from a small amount of monolingual data in the source language. In our method, we augment source monolingual data, generate target language data, and filter bilingual data. When we augment source monolingual data, we process each sentence individually and extract two keywords that are crucial to its content—typically a news genre and a named entity such as a place or organization. We then create new news texts by randomly combining these sentence-level keywords with the original source sentence. When we generate target language data, we use pivot translation to generate translations of the augmented source language data. This approach is particularly useful for handling non-English translation directions, where direct translation resources may be limited or less reliable. When we filter the bilingual quality, we compare vector representations of the generated bilingual data and eliminate data with low matching results from the training data. To increase the diversity and fluency of source language data and the quality of bilingual data, we prepare separate LLMs with high capabilities for source and target language.

The experimental results demonstrated that a general-purpose LLM adapted using instruction tuning with data from the proposed flow performed 20.23 BLEU points better in the Thai–Japanese sentence-to-sentence news translation task than before domain adaptation. The proposed method improved BLEU by 0.82 points through keyword-based source data generation and 2.09 points through the use of multiple LLMs. Each of these improvements showed a statistically significant difference from conventional methods.

II. Generation of Synthetic Bilingual Data for Domain Adaptation

Fig. 1 shows the synthetic bilingual data generation flow of the proposed method. The generation procedure is as follows:

979-8-3315-7206-8/25/$31.00

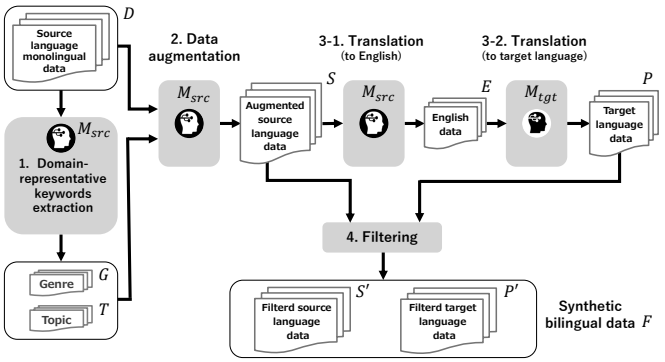

Fig. 1: Flow of synthetic data generation.

1) Extract domain-representative keywords from source language data. (Section II-A)
2) Augment the source language data with the extracted domain-representative keywords. (Section II-B)
3) Generate translations of the augmented source language data using pivot translation with English as the intermediate language. (Section II-C)
4) Eliminate low-quality bilingual data by evaluating the quality of the bilingual data. (Section II-D)

We use different LLMs for each step to improve data quality. The LLM M_{src}, with high source language capability, is used for keyword extraction, source data augmentation, and translation into the pivot language. The LLM M_{tgt}, with high target language capability, is used for translation from the pivot to the target language.

A. Extraction of Domain-representative Keywords

Before we augment the source data, we extract domain-representative keywords from the target domain data. In this study, we hypothesize that, in news articles, three elements are important for defining the domain: news genre, topic, and sentence structure. To extract genres and topics, we input source-language monolingual data $D = \{D_1, D_2, D_3, ..., D_d\}$ (seed data), which consists of d sentences, into the prompt and obtain the output from M_{src}. The genre prompt asks the model to assign a concise label (up to three English words) that best describes the news type, whereas the topic prompt requests a single proper noun, such as a location, person, or organization. We use M_{src} to extract domain information because it effectively distills knowledge such as specific terms used in the source language. For each sentence contained in D, we generate one genre and one topic. After collecting all the generated genres and topics, duplicates are removed to construct the final genre set $G = \{G_1, G_2, G_3, ..., G_g\}$ and topic set $T = \{T_1, T_2, T_3, ..., T_t\}$, where g and t represent the number of unique genres and topics extracted from the original d sentences, respectively. Table I shows examples of G and T extracted from the seed data. For G, we extract not only general news genres such as financial and art news, but also domain-specific news genres, including news about Buddhism and the royal family. For T, we extract place names, company names, and personal names.

TABLE I: Examples of the genres and topics extracted

	Examples
Genre	Financial News, Art News, Buddhist News, Infrastructure Development, Royal News
Topic	Phuket, Chiang Rai, Trat, King Rama IX, Pae Hingam National Park, Krung Thai Bank, Provincial Electricity Authority

B. Expansion of Source Language Data

We use the domain-representative keywords extracted in Section II-A to augment source language data. For each sentence D_k, we generate multiple new sentences by randomly selecting a genre G_r and topic T_q for each generation, where k indexes the sentence, and r and q index the randomly sampled elements from the predefined sets. These selected elements are then assigned to the prompt of M_{src}. This prompt instructs the model to rewrite the sentence D_k as content related to G_r and T_q, while preserving the sentence' s length and structure. We generate s extended source language data $S = \{S_1, S_2, S_3, ..., S_s\}$ by assigning multiple combinations of genres and topics to each D_k, applied across all elements in D. Table II shows examples generated by applying different genre and topic combinations to the same source sentence, following the settings in Section III.

C. Generation of Synthetic Parallel Translation Data

After augmenting D, we generate a target language translation P for S using the LLM. To improve translation quality, we generate $P = \{P_1, P_2, P_3, ..., P_s\}$ using the pivot translation of intermediate generated data $E = \{E_1, E_2, E_3, ..., E_s\}$ written in a high-resource language such as English. We use M_{src} for translation from the source language to the intermediate language and M_{tgt} for translation from the intermediate language to the target language.

D. Filtering

From the generated bilingual data $B = \{(S_i, P_i)\}_{i=1}^{s}$, we remove data unsuitable for training. First, we remove duplicates. Depending on the combination of G_r and T_q, it may not be possible to synthesize appropriate bilingual data, and only translations of keywords may be generated while ignoring the domain information or some of the generated results may contain duplications. If a sentence pair (S_n, P_n) exactly matches another pair (S_m, P_m) in B, we remove one of them, where m and n are indices of sentence pairs in B. Additionally, we remove low-quality bilingual training data. We use Sentence-BERT [17] to calculate the cosine similarity between S_l and P_l in B, where ℓ is the index of the pair, and then remove pairs with low similarity to B. We regard these low-similarity pairs as low-quality bilingual training data and exclude them from the training data. After we remove duplicates and low-quality pairs, we combine the source data $S' = \{S'_1, S'_2, S'_3, ..., S'_{s'}\}$ and target data $P' = \{P'_1, P'_2, P'_3, ..., P'_{s'}\}$ to obtain the final training set $F = \{(S'_j, P'_j)\}_{j=1}^{s'}$.

TABLE II: Examples of sentences generated by applying different genre and topic combinations to a single source sentence

Original (translated into English)		นายกรัฐมนตรี เตรียมลงพื้นที่ตรวจราชการจังหวัดนครศรีธรรมราชและจังหวัดสุราษฎร์ธานี ติดตามโครงการบรรเทาอุทกภัย (The Prime Minister is scheduled to visit Nakhon Si Thammarat and Surat Thani provinces to inspect flood mitigation projects.)
Genre	**Topic**	**Generated text (translated into English)**
Government Policy	Social Security Office	รัฐบาลเตรียมเสนอกระทรวงสาธารณสุขและสำนักงานประกันสังคมเพื่อแก้ไขปัญหาความเหลื่อมล้ำทางสุขภาพ ในโครงการส่งเสริมสุขภาพชุมชน (The government is preparing to propose that the Ministry of Public Health and the Social Security Office address health disparities in the community health promotion project.)
Local Food News	Songkhla	ผู้ว่าราชการจังหวัดสงขลา เตรียมลงพื้นที่ตรวจราชการตลาดน้ำสงขลาและตลาดสดเทศบาลนครสงขลา ติดตามโครงการบรรเทาความเดือดร้อนเรื่องอาหารแพง (The governor of Songkhla province is preparing to inspect the Songkhla floating market and the fresh market of Songkhla Municipality to monitor the project to alleviate the hardship of expensive food.)

III. Experiments

A. Experimental Settings

We conducted translation experiments on the Thai–Japanese news translation task. We performed instruction tuning (hereafter referred to as fine-tuning) on Llama-3-8B-Instruct[1] using synthetic data generated by the proposed method. We compared translation performance with that of the models described in Section III-A3.

1) Data: We used only a portion of the NE-Corpus, which is monolingual Thai data included in the AI for Thai platform[2]. This corpus is mainly composed of news articles about Thailand and includes tagged named entities. We removed named entity tags from the data, and then randomly extracted 500 sentences as seed data and 550 sentences as evaluation data. We obtained genres and topics for each sentence in the seed data by substituting each sentence into the prompts, and then removed duplicates from the output results to obtain 179 genres and 232 topics. We generated 100 Thai sentences using 100 genre-topic combinations for each seed sentence. To avoid excessive overlap of content and words in the generated sentences, we replaced all 100 genre and topic combinations every time we generated 1,000 sentences. We expanded the maximum number of sentences from 500 sentences and 5,000 genre-topic combinations to 50,000 sentences. After we generated the synthetic data, we applied the filtering process. This involved removing duplicates and excluding data with a cosine similarity of less than 0.4 between the vector representations of the source and target language data from the training data, which we calculated using Sentence-BERT with the *paraphrase-xlm-r-multilingual-v1*[3] model.

2) Training and Evaluation: For the LLM, we used Llama-3-Typhoon-v1.5x-70b-instruct-awq[4] from the Llama-3-Typhoon series [18] to generate genres and topics, expand the Thai data, and translate it into English. Additionally, we used Qwen2.5-72B-Instruct[5] from the Qwen series for the Japanese translation. Llama-3-Typhoon is a model that improves the Thai language ability of Llama-3, which is a general-purpose LLM, through additional training with Thai data. We set the temperature to 0.9 only when expanding the Thai data. For all other generation tasks, we set the temperature to 0. For fine-tuning, we used QLoRA [19] (lora rank 8, lora alpha 16) and trained for 5 epochs with a batch size of 32 and learning rate of 2e-4. We used the optimization method AdamW [20]. For the evaluation, we translated the evaluation data into Japanese using zero-shot learning and compared the performance of each model using BLEU [21] and COMET [22]. We computed BLEU scores using SacreBLEU [23] and COMET scores using *Unbabel/wmt22-comet-da*[6]. To compare the BLEU scores, we conducted significance tests with paired approximate randomization [24] using 10,000 approximate randomization trials and a p-value threshold of 0.05. We performed the tests using the SacreBLEU Python package.

3) Compared Methods: We compared the translation performance of three methods: (a) direct translation using generic LLMs, (b) pivot translation via English as an intermediate language, and (c) fine-tuning using human-translated data.

For direct translation using generic LLMs, we investigated the results of direct translation from Thai to Japanese using Llama-3-8B-Instruct, Azure OpenAI GPT-3.5-Turbo[7], Llama-3-Typhoon-v1.5x-70b-instruct-awq, and Qwen2.5-72B-Instruct before fine-tuning. For pivot translation, we compared the results of translating Thai, English, and Japanese, in that order, with English as the intermediate language. The methods compared were Llama-3-8B-Instruct and switching between Llama-3-Typhoon-v1.5x-70b-instruct-awq and Qwen2.5-72B-Instruct. In this switching method, we used Llama-3-Typhoon for Thai-to-English translation and Qwen for English-to-Japanese translation. For fine-tuning using human-translated data, we used a model fine-tuned with 100 bilingual sentences manually

[1]https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct
[2]https://aiforthai.in.th
[3]https://huggingface.co/sentence-transformers/paraphrase-xlm-r-multilingual-v1
[4]https://huggingface.co/scb10x/llama-3-typhoon-v1.5x-70b-instruct-awq
[5]https://huggingface.co/Qwen/Qwen2.5-72B-Instruct
[6]https://huggingface.co/Unbabel/wmt22-comet-da
[7]https://learn.microsoft.com/ja-jp/azure/ai-services/openai/concepts/models

TABLE III: Overall results

Method	BLEU	COMET
Direct translation using generic LLMs		
- (A) Llama-3-8B-Instruct	0.55	0.3681
- (B) GPT-3.5-Turbo	16.45	0.8212
- (C) Llama-3-Typhoon-v1.5x-70b-instruct-awq	15.70	0.8299
- (D) Qwen2.5-72B-Instruct	19.96	0.8494
Pivot translation		
- (E) Llama-3-8B-Instruct	1.10	0.3851
- (F) Llama-3-Typhoon-v1.5x-70b-instruct-awq + Qwen2.5-72B-Instruct	20.01	0.8385
Fine-tuning using human-translated data		
- (G) Llama-3-8B-Instruct (NE100)	1.53	0.3978
- (H) Llama-3-8B-Instruct (ALT10k)	14.93	0.8007
Fine-tuning using data generated by the proposed method		
- (I) Llama-3-8B-Instruct (Synthetic data 10k)	**20.13**	**0.8622**
- (J) Llama-3-8B-Instruct (Synthetic data 50k)	**20.78**	**0.8625**

generated from the NE-Corpus (NE100)[8] and a model fine-tuned with 10,000 sentences from the Thai–Japanese bilingual data of the ALT corpus [25], which is a multilingual corpus of English Wikinews[9] (ALT10k). We fine-tuned Llama-3-8B-Instruct using 10,000-sentence synthetic bilingual data generated from a seed dataset of 100 sentences, and evaluated the performance of each method. For the proposed method, we additionally investigated performance using 50,000-sentence data generated from 500 seed sentences. We used the 8B model in both cases because of hardware constraints.

B. Experimental Results

1) Overall Results: The experimental results are shown in Table III. The results demonstrated that the models fine-tuned using the data generated by the proposed method (Model I and Model J) achieved the highest scores in both the BLEU and COMET evaluations. The BLEU score of the base model (Model A), which initially produced poor translations, increased by 20.23 points after fine-tuning using the proposed method' s data (Model J). The BLEU score of Model I was comparable with that of the pivot translation model (Model F), which uses translations from Llama-3-Typhoon-v1.5x-70b-instruct-awq and Qwen2.5-72B-Instruct as training data. The results also demonstrated the effectiveness of the synthetic data generated by the proposed method because we observed almost no performance improvement when fine-tuning using the NE-Corpus (Model G), which consists of 100 sentences of human-translated bilingual data used as seed data. Additionally, Model I performed better than the model fine-tuned on the ALT corpus (Model H), which contains 10,000 sentences of human translation. Although both the NE-Corpus and ALT corpus are categorized as news datasets, their content differs significantly. The NE-Corpus primarily contains Thai local news, while the ALT corpus focuses on U.S.-centric international news. This difference in regional focus and terminology is likely to have contributed to the domain mismatch and observed performance gap. In the proposed method, when we increased the amount of synthetic data from 10,000 sentences (Model I) to 50,000 sentences (Model J), the BLEU score improved further, whereas the COMET score remained stable. This indicates that increasing the volume of synthetic data contributed positively to translation performance.

A comparison of the translation results is shown in Table IV. The translations produced by GPT-3.5 Turbo (Model B) and the model trained on the ALT corpus (Model H) were both fluent, but they were unable to translate terms and expressions specific to Thailand, such as the Buddhist calendar and place names. Regarding the Buddhist calendar (bold in the table), in the original text, 2021 was written as 2564 in the Buddhist calendar, and when translated into Japanese, it was necessary to convert from the Buddhist calendar to the Western calendar, but Model H output 2564 without conversion. Regarding place names (underlined with a wavy line in the table), the output of Model H was correct for "Songkhla city" in terms of the meaning, but it was output in Chinese. By contrast, the proposed method (Model I) output appropriate Japanese notation, although there was some variation in the notation with correct data. Because Model I was trained on data expanded from Thai news articles, it was able to accurately translate Thailand-related terms. These results suggest that matching the domain of training data to the target domain is crucial for improving translation performance.

2) Data Augmentation Using Keywords: We conducted an experiment to confirm the effect of substituting seed data, genre, and topic in the prompt when expanding Thai data. The methods compared were: (1) using one fixed example of seed data in the prompt without genre or topic, (2) assigning seed data without genre or topic, and (3) assigning seed data, genre, and topic, where genres and topics were generated without reference texts. In method (3), for genres, we used those obtained by instructing the LLM to "generate as many news genres as possible." For topics, we used those obtained by instructing the LLM to "generate as many terms related to Thailand as possible, such as place names, facility names, and company names." We generated genres and topics in method (3) using Qwen2.5-72B-Instruct, and then removed duplicates from those generated to obtain 248 genres and 227 topics. We designed methods (2) and (3) to approximate techniques from previous work. Our method is similar in approach to those in previous studies; however, it differs in that our monolingual data expansion rate is 100 times greater. Method (2) is based on the keyword-free data expansion approach [15], whereas Method (3) is based on the keyword-based data augmentation technique [14]. Among other related methods, the storytelling approach [16] was difficult to apply under the conditions of this study. Preliminary experiments showed that it was not possible to generate 100 story sentences from a single input sentence. By contrast, the paraphrasing method presented in the same study as the storytelling approach closely resembles Method (2).

The results are shown in Table V. The results showed that the method generating data using both the seed data and the genres and topics extracted by the proposed method achieved the best performance. We observed a statistically significant

[8]We used this human-translated data only for performance comparison; we did not use target language data in the proposed generation method.

[9]https://en.wikinews.org

TABLE IV: Comparison of the translation results

Source	เทศบาลนครสงขลา เปิดศูนย์อำนวยการเฉพาะกิจการป้องกันและช่วยเหลือผู้ประสบอุทกภัย วาตภัยและโคลนถล่ม ประจำปี **2564**
Reference	ソンクラー市は、**2021** 年度の洪水、暴風雨、土砂崩れの防止と被災者支援を行う特殊任務指令センターを開設した。
Model	**Output**
(B) GPT-3.5-Turbo	市役所は、**2021** 年に洪水、台風、土砂崩れの被害者を支援するための特別任務センターを開設しました。
(H) Llama-3-8B-Instruct (ALT10k)	宋卡市は、水害、洪水、土砂崩れのための特別な救援センターを、**2564** 年に開設した。
(I) Proposed	サンクラー市は、**2021** 年の洪水、嵐、土砂災害の対策と救援のための特別対策センターを開設しました。

TABLE V: Performance comparison for different prompts when data were expanded

	Seed data	Genre & topic	BLEU	COMET
(1)	✓ (fixed sentence)	-	12.87	0.6982
(2)	✓	-	19.31	0.8531
(3)	✓	✓ (w/o seed data)	19.13	0.8539
Proposed	✓	✓	**20.13**	**0.8622**

TABLE VI: Performance comparison of different LLMs used to generate synthetic bilingual data

Model used to generate synthetic data	BLEU	COMET
Typhoon only	16.14	0.8207
Qwen only	18.04	0.8355
Typhoon+Qwen (Proposed)	**20.13**	**0.8622**

difference between the proposed method and the comparison methods. We confirmed that performance was improved by showing each sentence in the seed data as an example rather than showing only a fixed sentence, and that performance was further improved by providing the genres and topics extracted by the proposed method. These results demonstrate that using both the sentence structure of the source language data and keywords that indicate news genres and topics was effective for generating synthetic data for domain applications. Additionally, we found that even when using keywords to generate data, the BLEU score did not improve unless the genres and topics were created with reference to the seed data. The performance remained comparable to the case without keywords. This is likely to be caused by inappropriate genres and topics being set, which deviated from the domain of the target translation. This shows that using seed data for genre and topic extraction produced data more aligned with the target domain.

3) Switching LLMs: We investigated translation performance by switching LLMs during synthetic data generation. The comparison methods are a model that uses only Llama-3-Typhoon-v1.5x-70b-instruct-awq for both source data expansion and pivot generation (Typhoon only), and a model that uses only Qwen2.5-72B-Instruct (Qwen only), which does not switch LLMs for fine-tuning data generation. The genres and topics used in both methods were the same as those in the proposed method.

The results are shown in Table VI. The experimental results confirmed that using multiple LLMs improved translation performance, with statistically significant differences observed compared with methods that use a single LLM. The proposed method showed that using different models according to their purpose enabled the generation of high-quality synthetic data and improved translation performance. By contrast, in the results of Table III, there was almost no difference in performance between the results of direct translation from Thai to Japanese using Qwen and the results of pivot translation by switching between Llama-3-Typhoon and Qwen. This result suggests that Llama-3-Typhoon contributed more to the generation of data dealing with Thai terms and topics than to the improvement of translation quality.

IV. Conclusion

In this study, we proposed an effective synthetic data generation method for adapting a general-purpose LLM to the news domain when only a small amount of source language data is available. Unlike previous approaches, the proposed method enables the large-scale expansion of monolingual data: up to 100 times the original size. The experimental results demonstrated that to generate effective synthetic data for domain applications, it is essential to control the expansion of source language sentences using appropriate keywords and to select suitable LLMs based on the language of the text to be generated. Future work will evaluate the performance under different data augmentation scales and across various LLMs.

Acknowledgments

These research results were obtained from the commissioned research (No. 225) by National Institute of Information and Communications Technology (NICT), Japan.

We used “NE-Corpus” created by the National Electronics and Computer Technology Center (NECTEC) (CC BY-NC-SA 3.0 TH). For details, please refer to the following link (https://creativecommons.org/licenses/by-nc-sa/3.0/th/).

References

[1] OpenAI, J. Achiam, S. Adler, *et al.*, *Gpt-4 technical report*, 2024. arXiv: 2303.08774 [cs.CL].

[2] A. Grattafiori, A. Dubey, A. Jauhri, *et al.*, *The llama 3 herd of models*, 2024. arXiv: 2407.21783 [cs.AI].

[3] Qwen, : A. Yang, *et al.*, *Qwen2.5 technical report*, 2025. arXiv: 2412.15115 [cs.CL].

[4] J. Wei, M. Bosma, V. Y. Zhao, *et al.*, *Finetuned language models are zero-shot learners*, 2022. arXiv: 2109.01652 [cs.CL].

[5] H. W. Chung, L. Hou, S. Longpre, *et al.*, *Scaling instruction-finetuned language models*, 2022. arXiv: 2210.11416 [cs.LG].

[6] J. Li, H. Zhou, S. Huang, S. Cheng, and J. Chen, "Eliciting the translation ability of large language models via multilingual finetuning with translation instructions," *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 576–592, 2024. DOI: 10.1162/tacl_a_00655.

[7] F. Zheng, E. Marrese-Taylor, and Y. Matsuo, "Improving low-resource machine translation for formosan languages using bilingual lexical resources," in *Findings of the Association for Computational Linguistics: ACL 2024*, Association for Computational Linguistics, 2024, pp. 11 248–11 259. DOI: 10.18653/v1/2024.findings-acl.670.

[8] M. Fadaee, A. Bisazza, and C. Monz, "Data augmentation for low-resource neural machine translation," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Association for Computational Linguistics, 2017, pp. 567–573. DOI: 10.18653/v1/P17-2090.

[9] T. Cohn and M. Lapata, "Machine translation by triangulation: Making effective use of multi-parallel corpora," in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Association for Computational Linguistics, 2007, pp. 728–735.

[10] R. Sennrich, B. Haddow, and A. Birch, "Improving neural machine translation models with monolingual data," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany: Association for Computational Linguistics, 2016, pp. 86–96. DOI: 10.18653/v1/P16-1009.

[11] J. Sälevä and C. Lignos, "Language model priors and data augmentation strategies for low-resource machine translation: A case study using Finnish to Northern Sámi," in *Findings of the Association for Computational Linguistics: ACL 2024*, Association for Computational Linguistics, 2024, pp. 12 949–12 956. DOI: 10.18653/v1/2024.findings-acl.768.

[12] T. Ge, X. Chan, X. Wang, D. Yu, H. Mi, and D. Yu, *Scaling synthetic data creation with 1,000,000,000 personas*, 2025. arXiv: 2406.20094 [cs.CL].

[13] J. Santoso, P. Sutanto, B. Cahyadi, and E. Setiawan, "Pushing the limits of low-resource NER using LLM artificial data generation," in *Findings of the Association for Computational Linguistics: ACL 2024*, Association for Computational Linguistics, 2024, pp. 9652–9667. DOI: 10.18653/v1/2024.findings-acl.575.

[14] P. Pengpun, C. Udomcharoenchaikit, W. Buaphet, and P. Limkonchotiwat, "Seed-free synthetic data generation framework for instruction-tuning LLMs: A case study in Thai," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, Association for Computational Linguistics, 2024, pp. 445–464, ISBN: 979-8-89176-097-4.

[15] Y. Moslem, R. Haque, J. Kelleher, and A. Way, "Domain-specific text generation for machine translation," in *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, Association for Machine Translation in the Americas, 2022, pp. 14–30.

[16] S. Oh, S. A. Lee, and W. Jung, *Data augmentation for neural machine translation using generative language model*, 2023. arXiv: 2307.16833 [cs.CL].

[17] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Association for Computational Linguistics, 2019, pp. 3982–3992. DOI: 10.18653/v1/D19-1410.

[18] K. Pipatanakul, P. Jirabovonvisut, P. Manakul, *et al.*, *Typhoon: Thai large language models*, 2023. arXiv: 2312.13951 [cs.CL].

[19] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "Qlora: Efficient finetuning of quantized llms," in *Advances in Neural Information Processing Systems*, vol. 36, Curran Associates, Inc., 2023, pp. 10 088–10 115.

[20] I. Loshchilov and F. Hutter, *Decoupled weight decay regularization*, 2019. arXiv: 1711.05101.

[21] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, 2002, pp. 311–318. DOI: 10.3115/1073083.1073135.

[22] R. Rei, C. Stewart, A. C. Farinha, and A. Lavie, "COMET: A neural framework for MT evaluation," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, 2020, pp. 2685–2702. DOI: 10.18653/v1/2020.emnlp-main.213.

[23] M. Post, "A call for clarity in reporting BLEU scores," in *Proceedings of the Third Conference on Machine Translation: Research Papers*, Association for Computational Linguistics, 2018, pp. 186–191. DOI: 10.18653/v1/W18-6319.

[24] S. Riezler and J. T. Maxwell, "On some pitfalls in automatic evaluation and significance testing for MT," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, Association for Computational Linguistics, 2005, pp. 57–64.

[25] H. Riza, M. Purwoadi, T. Uliniansyah, *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)*, IEEE, 2016, pp. 1–6.