Latest Research in Data Augmentation for Low Resource Language Text Translation: A Review

1st Andi Djalal Latief
Research Center for Data and Information Sciences
National Agency for Research and Innovation
Jakarta, Indonesia
andi002@brin.go.id

2nd Asril Jarin
Research Center for Data and Information Sciences
National Agency for Research and Innovation
Jakarta, Indonesia
asri003@brin.go.id

3rd Yaniasih Yaniasih
Research Center for Data and Information Sciences
National Agency for Research and Innovation
Jakarta, Indonesia
yaniasih@brin.go.id

4th Dian Isnaeni Nurul Afra
Research Center for Data and Information Sciences
National Agency for Research and Innovation
Jakarta, Indonesia
dian059@brin.go.id

5th Elvira Nurfadhilah
Research Center for Data and Information Sciences
National Agency for Research and Innovation
Jakarta, Indonesia
elvi003@brin.go.id

6th Siska Pebiana
Research Center for Data and Information Sciences
National Agency for Research and Innovation
Jakarta, Indonesia
sisk006@brin.go.id

7th Nuraisa Novia Hidayati
Research Center for Data and Information Sciences
National Agency for Research and Innovation
Jakarta, Indonesia
nura017@brin.go.id

8th Radhiyatul Fajri
Research Center for Data and Information Sciences
National Agency for Research and Innovation
Jakarta, Indonesia
radh001@brin.go.id

2024 International Conference on Computer, Control, Informatics and its Applications (IC3INA) | 979-8-3315-4231-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/IC3INA64086.2024.10732042

***Abstract*—The translation of low-resource languages remains a significant challenge in Natural Language Processing (NLP) due to the scarcity of high-quality parallel data for training machine translation models. Data augmentation techniques, which artificially expand the size and diversity of datasets, offer a promising solution to this problem. This study comprehensively reviews data augmentation techniques in low-resource language text translation. Sixteen recent articles met the specified eligibility criteria employing the systematic literature review protocol. The review categorized the various data augmentation methods into eight groups: translation-based augmentation, synthetic data generation, sentence manipulation, grammar and error correction, multi-task and advanced transformations, miscellaneous transformations, consistency and self-training, and embedding and contextual methods. The findings highlight the significant impact of data augmentation on improving translation quality, addressing data scarcity, and enhancing model robustness. Despite these advancements, challenges such as grammatical errors, semantic inconsistencies, and the quality of synthetic data remain. This review also provides insights into the effectiveness of different language pairs and dataset sizes, emphasizing the need for further research to refine these methods and address challenges in low-resource languages.**

***Keywords*—data augmentation, synthetic data, low-resource language, neural machine translation**

I. Introduction

Text translation is one of the crucial areas in natural language processing (NLP) because it allows communication across languages and cultures. Despite significant advances in machine translation models, low-resource languages continue to face major challenges. These languages lack sufficient high-quality parallel data to train effective translation models. In this context, data augmentation becomes a promising technique to overcome these limitations [1]. However, recent data augmentation methods for low-resource machine translation have rarely been systematically reviewed.

Previous systematic literature reviews [6], [7] have addressed the challenges and approaches of neural machine translation (NMT) for low-resource languages in general, there remains a gap in the literature specifically focusing on data augmentation techniques. This study aims to fill that gap by providing a comprehensive review of the latest research on data augmentation strategies applied to low-resource language text translation. The novelty of this research lies in its focus on these data augmentation techniques, which have emerged as a critical area of innovation for overcoming the limitations of low-resource NMT.

Data augmentation is the process of expanding an existing dataset by modifying or combining existing data to create new data. This technique has proven effective in a variety of NLP applications, including text translation. By leveraging data augmentation, researchers and practitioners can increase the amount and diversity of training data, which in turn can improve the performance of translation models. Many studies show that data augmentation is an effective approach in improving the performance of current machine translation models. These studies introduce various data augmentation techniques, as well as examine their positive impact on translation quality for low-resource languages [2], [3]. Research by Fadaee et al. [4] show that data augmentation can significantly improve translation quality for languages with a limited corpus by adding new, artificially generated sentence examples. This finding is in line with research by Edunov et al. [5] which revealed that data augmentation techniques can increase the accuracy of NMT models.

Recognizing the importance of this technique, it is essential to summarize and analyze the state of the art in these methods, focusing on the available techniques, their performance, challenges, and future directions. The following research questions guide this review:
RQ 1: What are the recent data augmentation techniques?
RQ 2: What are the objectives of these techniques?

RQ 3: How is the performance comparison?
RQ 4: What are the challenges and future research direction?

The end of this research provides recommendations for the direction and focus of further research to improve the effectiveness of data augmentation in low-resource language text translation.

II. Methodology

This study used a rigorous and systematic approach following the guidelines outlined by Kitchenham et al. [8] to conduct a Systematic Literature Review (SLR). We use Parsifal as the primary tool for managing and analyzing data, due to its ability to integrate with key databases as well as its efficient data management features [9].

We performed a title and abstract search on the Scopus database using the following search strings and filtering (document type article journal, published in the last 3 years, written in English): TITLE-ABS-KEY(("machine translation" OR "translation model*") AND (neural OR transformer OR "deep learning") AND ("data augmentation") AND (low-resource OR "low resource" OR limited-resource OR "limited resource" OR under-resource OR "under resource" OR less-resource OR "less resource" OR resource-scarce OR "resource scarce" OR scarcity) AND (language* OR data)) AND PUBYEAR > 2021 AND PUBYEAR < 2025 AND (LIMIT-TO (DOCTYPE,"cp") OR LIMIT-TO (DOCTYPE,"ar")) AND (LIMIT-TO (LANGUAGE,"English")). This search yielded 84 relevant papers.

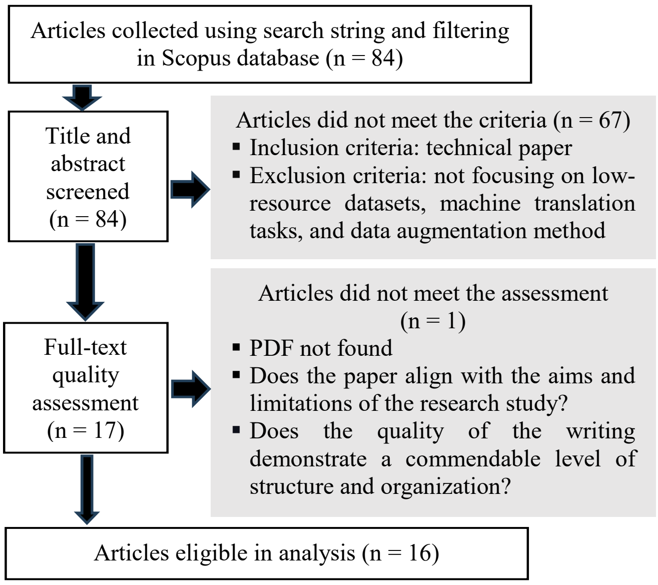
Articles collected using search string and filtering in Scopus database (n = 84)
Title and abstract screened (n = 84)
Articles did not meet the criteria (n = 67)
▪ Inclusion criteria: technical paper
▪ Exclusion criteria: not focusing on low-resource datasets, machine translation tasks, and data augmentation method
Full-text quality assessment (n = 17)
Articles did not meet the assessment (n = 1)
▪ PDF not found
▪ Does the paper align with the aims and limitations of the research study?
▪ Does the quality of the writing demonstrate a commendable level of structure and organization?
Articles eligible in analysis (n = 16)

Fig. 1. The protocol diagram

Then, we applied inclusion and exclusion criteria to screen relevant papers. By applying these criteria, the number of relevant papers is reduced to 17. Of the 17 papers screened, we searched for the full text and succeeded in finding 16 papers (1 paper had no full text found). We then confirmed two quality assessment criteria of these 16 full-text articles. For the 16 papers that met the requirements, we conducted an in-depth analysis to answer the research questions. The filtering, screening, and extraction stages were conducted by three individuals for each article to ensure the absence of bias, following the established criteria. The detailed steps are shown in the protocol diagram in Fig. 1.

III. Results and Discussion

In recent years, the field of highlighting and Machine Translation (MT) has witnessed significant advancements, particularly in addressing the challenges associated with low-resource languages. Data augmentation techniques have emerged as a pivotal strategy to enhance the quality and robustness of translation models by artificially expanding the training datasets. The objective of this research is to provide a comprehensive review of data augmentation techniques in low-resource language text translation.

A. *State of the Art of Data Augmentation Techniques on Neural Machine Translation*

This section categorizes and describes the various data augmentation techniques employed over the past five years, based on an SLR of 16 research articles. The identified techniques can be broadly classified into eight distinct categories (Table I), each leveraging unique approaches to data augmentation, as follows:

TABLE I. Classification of Data Augmentation Techniques

Articles	Data Augmentation Techniques							
	TBA	***SDG***	***SM***	***GEC***	***MAT***	***MT***	***CST***	***ECM***
[10]	✓							
[11]		✓						
[12]			✓					
[13]					✓			
[14]	✓						✓	
[15]	✓						✓	
[16]	✓							
[17]	✓					✓		✓
[18]			✓					
[19]				✓				
[20]				✓				
[21]					✓			
[22]		✓						
[23]	✓	✓						
[24]	✓	✓						
[25]						✓		

1) Translation-Based Augmentation (TBA)

Translation-based augmentation includes techniques like back-translation, tagged back-translation, and self-training. These methods create more training data by translating sentences into different languages and then back to the original. Automated translation tools are used to increase dataset size and diversity, addressing data scarcity in low-resource language pairs. Several studies [10], [14], [15] show these techniques improve the robustness and generalization of machine translation models by incorporating diverse linguistic structures and more training examples.

2) Synthetic Data Generation (SDG)

Synthetic Data Generation techniques create artificial data that mimics the statistical properties of real datasets, specifically by generating synthetic parallel sentences or phrase pairs. Methods like Synthetic Parallel-Sentence Augmentation (SPA) and Phrase Pairs Augmentation (PPA) are used to replicate authentic parallel sentence structures, addressing data scarcity and enriching training datasets with artificial examples. Researchers such as Khenglawt et al. [11], Laskar et al. [23], [24], and Ngo et al. [22] have shown that these techniques enhance translation quality by providing models with a wider range of examples, reducing overfitting, and improving generalization in various linguistic contexts.

3) Sentence Manipulation (SM)

Sentence Manipulation techniques involve modifying the structure or content of sentences to create diverse training instances. This category includes methods such as Sentence Trunks Augmentation (STA) and Semantic-Context Data Augmentation, which aim to generate varied sentence forms to enhance the model's ability to handle diverse linguistic structures and contexts. Researchers like Li et al. [18] and Zhang et al. [12] have utilized these techniques to augment the training dataset by reordering, splitting, or otherwise altering sentences. This approach enriches the dataset with a wider range of sentence structures, improving the model's robustness and its capability to achieve higher translation quality across different linguistic variations and contexts.

4) Grammar and Error Correction (GEC)

Grammar and Error Correction techniques enhance training data quality by addressing grammatical errors and improving coherence. This includes methods like Grammar Error Correction Data Augmentation (GECDA), which uses aggressive transformations for grammatical correctness and semantic consistency. Researchers Liu et al. [19] and Solyman et al. [20] have shown these techniques generate high-quality training data, improving translation model robustness and effectiveness in handling varied grammatical inputs, thus enhancing overall translation performance.

5) Multi-task and Advanced Transformations (MAT)

Multi-task and Advanced Transformations encompass sophisticated techniques that leverage multi task learning and complex data transformations to enrich the training dataset. This category includes methods such as Multi-task Learning Data Augmentation (MaTiLDA), SwitchOut, and Ciphertext Based Data Augmentation (CipherDAug), which expose the model to diverse linguistic patterns and tasks. Researchers such as San et al. [13] and Sánchez-Cartagena et al. [21] have demonstrated the efficacy of these techniques in enhancing model performance across various translation tasks by leveraging shared knowledge and exposing the model to a broad spectrum of transformations. By incorporating multi-task and advanced data manipulation, these approaches improve the model's ability to generalize and achieve higher translation quality in different linguistic contexts.

6) Miscellaneous Transformations (MT)

Miscellaneous Transformations encompass a diverse range of augmentation techniques that introduce unique variations into the training data, enhancing the robustness and adaptability of translation models. This category includes methods such as Constrained Sampling, Delete Adjective Augmenter (AdjAug), SwitchOut, and Cipher transformations [13], [25]. These techniques introduce randomness and variability into the dataset, aiming to prevent overfitting and improve the model's generalization capabilities across different linguistic contexts. By incorporating innovative transformations that may not fit into other specific categories, these approaches contribute to the overall enhancement of translation quality and model performance.

7) Consistency and Self-Training (CST)

Consistency and Self-Training techniques aim to enhance model robustness and performance by ensuring data and model output consistency. Strategies like Data-level and Model-level Consistency Training (DM-CT) and self-training are used to achieve this. Researchers [14], [15] have shown that self-training, where models use their high-confidence predictions as additional training data, effectively refines predictions and improves generalization across data variations, thereby boosting overall translation performance and reliability.

8) Embedding and Contextual Methods (ECM)

Embedding and Contextual Methods use advanced embedding techniques and contextual information to enhance training data for machine translation models. Techniques like the Contextual Word Embedding Augmenter (EmbedAug) generate or modify sentences to include semantic nuances, improving contextual understanding. Research by Kchaou et al. [17] shows that embedding-based augmentation methods effectively improve translation quality by enriching datasets with contextual information, thereby helping models capture language complexities and nuances for more accurate and contextually relevant translations.

B. Thematic Analysis of Research Objectives

The thematic analysis of research objectives extracted from the 16 reviewed articles reveals several key trends that underpin the use of data augmentation techniques in low-resource language text translation. These themes demonstrate the primary motivations guiding researchers in this domain, reflecting the growing need to address data scarcity and improve the overall performance of translation systems. As shown in Table II, the themes encompass objectives such as enhancing translation quality, improving model robustness, and exploring advanced methodologies, all of which have been central to research efforts in the past three years.

TABLE II. THEMATIC ANALYSIS OF RESEARCH OBJECTIVES

Objective	Articles
Addressing Data Scarcity	[10], [25]
Improving Translation Quality	[11], [13], [24]
Enhancing Model Robustness and Generalization	[11]
Leveraging Advanced Techniques	[13]
Utilizing Synthetic Data	[10]
Fine-tuning Pre-trained Models	[20]

1) Addressing Data Scarcity

Many studies focus on addressing the fundamental challenge of data scarcity in low-resource languages. For example, Khenglawt et al. [11] tackle the data scarcity problem in English-Mizo by using data augmentation techniques and language models. Similarly, Yu [10] proposes a filtered data augmentation method based on model competence evaluation to improve the quality of training data and performance in low-resource NMT. The primary objective of these efforts is to enhance the availability and diversity of training data, thereby improving translation quality and model robustness.

2) Improving Translation Quality

A significant number of studies aim to improve the overall quality of translations produced by NMT models by utilizing data augmentation techniques. For example, Li et al. [18] propose an efficient data augmentation method named STA for low-resource NMT, which utilizes sentence trunks to generate pseudo-parallel sentence pairs and improve translation quality. Similarly, Laskar et al. [23], [24] aim to enhance translation accuracy for the low-resource English-Assamese language pair by creating a domain-specific parallel corpus and employing various data augmentation techniques. The primary objective of these efforts is to utilize data augmentation to enhance the accuracy, fluency, and coherence of translations, particularly for challenging language pairs.

3) Enhancing Model Robustness and Generalization

Several studies focus on making NMT more robust and capable of generalizing better across different datasets and scenarios. For instance, Liang et al. [15] aim to enhance the robustness and generalization of models in low-resource NLP tasks by incorporating both data-level and model-level consistency training through the DM-CT framework. Additionally, Sánchez Cartagena et al. [21] demonstrate that non-fluent synthetic target sentences can improve translation performance when used in a multilingual MT framework. The overarching objective of these efforts is to develop models that maintain high performance across various contexts and input types, even with limited training data.

4) Leveraging Advanced Techniques

Research objectives often involve leveraging advanced techniques such as semantic context augmentation, constrained sampling, and hybrid methods. For example, Zhang et al. [12] aim to address issues of data sparsity and semantic inconsistency in Mongolian–Chinese neural machine translation (MNMT) by proposing a new semantic-context data augmentation (DA) method. Similarly, Maimaiti et al. [25] focus on augmenting the corpus for low-resource languages NMT by introducing a constrained sampling method and evaluating its effectiveness. The overarching objective is to explore and implement sophisticated data augmentation methods to push the boundaries of what is achievable in low-resource NMT.

5) Utilizing Synthetic Data

Creating and using synthetic data is a key strategy to enhance real-world data and address resource constraints. Ngo et al. [22] developed an effective method for generating synthetic bilingual data without external resources, improving NMT for Chinese and Japanese to Vietnamese. Similarly, Vu et al. [14] examined how transformer-based NMT models perform with more synthetic data for low-resource language pairs, specifically Chinese-to-Vietnamese. The goal is to produce high-quality synthetic data to augment training datasets and enhance model performance.

6) Fine-tuning Pre-trained Models

Fine-tuning pre-trained models with augmented data is a common goal in NMT. For instance, San et al. [13] seeks to improve the performance of low-resource NMT for Thai-Myanmar-English pairs using data augmentation and fine-tuning pre-trained models. This approach aims to enhance existing models with additional augmented training data, boosting their effectiveness and accuracy in translation tasks.

C. Performance Comparison Between Different Techniques

Evaluating the performance of machine translation (MT) systems involves several metrics, each with its strengths and limitations. Traditional metrics like BLEU and TER have been widely used due to their simplicity and ease of implementation, but they often fall short of capturing semantic nuances and the overall quality of translations [26]. More recent metrics, such as BERTScore and COMET, leverage neural network models to better capture semantic similarity and have shown improved correlation with human judgments [27]. However, these neural metrics can be less interpretable and may lack defined score ranges, complicating their use in practical settings. Additionally, confidence intervals for metrics like BLEU and NIST can help determine the statistical significance of differences between MT systems, providing a more robust evaluation framework [28].

TABLE III. METRICS EMPLOYED TO MEASURE PERFORMANCE

Metrics	Articles	Count of Articles	Percentage
BLEU	[10]–[25]	16	100%
TER	[11], [13], [24]	3	18%
METEOR	[11]	1	6%
ChrF	[13]	1	6%
RIBES	[10]	1	6%
Precision	[20]	1	6%
Recall	[20]	1	6%
F1	[11], [20]	2	12%

The metrics utilized in the chosen articles, as shown in the survey results (Table III), indicate that BLEU remains the predominant metric across all articles. Following BLEU, the second most used metric is TER, with F1 score coming in third, and several newer metrics following suit. As all articles employ the BLEU metric, the comparative performance evaluation is centered around its value, specifically focusing on the technique's performance within the same language pair. Table III presents seven language pairs which are used as datasets for more than one article so can be compared. One key metric used to evaluate translation quality is the BLEU score.

In all results, BLEU scores range from 8.06 to 38.01, indicating variability in translation quality depending on the technique and dataset size used. For example in language pairs that most studied (English-German) translation performance, the Scenario-Generic NMT dataset, which includes 4 million datasets, achieved a BLEU score of 18.91. In contrast, techniques such as Constrained Sampling, applied to a smaller dataset of 22,000 sentence pairs, resulted in a higher BLEU score of 27.94. This suggests that specific techniques can significantly impact translation performance, even with varying dataset sizes.

Table IV reveals that various augmentation and sampling techniques can dramatically influence machine translation performance. For instance, the MaTiLDA (Multi-task Learning Data Augmentation) technique yielded a BLEU score of 25.1 with a dataset of 172,000 pairs, highlighting its efficacy in improving translation quality. On the other hand, constrained sampling techniques, even with smaller datasets, demonstrated high performance, as evidenced by the BLEU score of 27.94. These findings suggest that while larger datasets generally support better translation models, the choice of technique can play an equally crucial role in determining the final performance.

The table also sheds light on the impact of dataset size and validation on translation performance. Larger datasets do not necessarily guarantee the highest BLEU scores. It happened in almost all the language pairs. Instead, the combination of data size and sophisticated augmentation techniques appears to be the key to achieving superior translation results. The relatively modest dataset sizes used in conjunction with advanced techniques like Constrained Sampling and STA underscore the potential for efficient data utilization in training effective machine translation models. The analysis reveals that translation performance is influenced by a combination of dataset size, techniques employed, and the inherent difficulty of the language pairs.

TABLE IV. PERFORMANCE COMPARISON

Language Pairs	BLEU Score	Size of Original Dataset	Techniques	Articles
English-German (en-de)	18.91	4M	Scenario-Generic NMT data augmentation method	[19]
	27.94	22K	Constrained Sampling	[25]
	25.1	172K	MaTiLDA (Multi-task Learning Data Augmentation)	[21]
	28.37	175K	Filtered back-translation	[10]
German-English (de-en)	35.01	173K	STA (Sentence Trunks Augmentation)	[18]
	35.14	800K	Constrained Sampling	[25]
	38.01	174K	Data-level and Model-level Consistency Training (DM-CT)	[15]
	31.39	173K	Filtered back-translation	[10]
Assamese-English (as-en)	20.04	57K	Phrase pairs and synthetic parallel data	[23]
	13.02	639K	Phrase pairs injection and Synthetic parallel data augmentation using the back-translation (BT) technique	[24]
English-Assamese (en-as)	16.02	57K	Phrase pairs and synthetic parallel data	[23]
	8.06	639K	Phrase pairs injection and Synthetic parallel data augmentation using the back-translation (BT) technique	[24]
Chinese-Vietnamese (zh-vi)	27.0	192K	Combination of back-translation, sentence concatenation, and self-training with fine-grained tagging	[14]
	15.70	266K	Artificial Translation Units (ATUs)	[22]
Vietnamese-English (vi-en)	29.82	135K	STA (Sentence Trunks Augmentation)	[18]
	29.79	702K	Constrained Sampling	[25]
Chinese-English (zh-en)	15.83	215K	STA using concatenation	[18]
	22.84	211K	Filtered back-translation	[10]

D. *Challenges and Opportunities for Future Research*

The systematic literature evaluation demonstrates that the translation process in multilingual contexts presents significant obstacles due to variances in word order, grammatical structure, and data scarcity. Several research has identified recurring difficulties that reduce the effectiveness of NMT systems. First, data scarcity and insufficient diversity in training data pose substantial challenges. Liu et al. [19] and Yu and Zhang [10] underscore these challenges when translating from English to German and other languages, highlighting the need for more diversified and rich data sources. Similarly, Quoc et al. [16], Sanchez-Cartagena et al. [21], and Ngo et al. [22] highlight the challenges posed by limited data sources and low-quality synthetic data in their respective translation assignments. Grammatical and semantic alignment provide further obstacles. Zhang et al. [12] and Vu et al. [14] observe that distinctive grammatical patterns in languages such as Mongolian and Vietnamese result in grammatical errors in synthetic data. Furthermore, Solyman et al. [20] and Laskar, Manna, et al. [23] cite grammatical faults and poor contextual meaning in augmented data as critical challenges in Arabic and Assamese translations. Phonological and cryptographic discrepancies exacerbate the translating procedure. Kchaou et al. [17] and Khenglawt et al. [11] explore how phonological and cryptographic changes in Arabic dialects and Mizo affect translation quality, requiring fine-tuned pre-processing and data augmentation procedures.

Handling out-of-vocabulary words is also an important topic. Laskar et al. [24] emphasize the difficulties of managing out-of-vocabulary words, named entities, and technical terminology in English-to-Assamese translations, highlighting the necessity for enlarged lexicons and better handling procedures. The quality of synthetic data remains an ongoing issue. Li et al. [18] examined translation between English and several languages, including German, Spanish, Chinese, Vietnamese, and Turkish. This study also faced challenges associated with limited parallel data, which necessitated high-quality data augmentation. However, the augmentation process was impeded by considerable granularity issues. Maimaiti et al. [25] and Ngo et al. [22] emphasize the difficulty of creating high-quality synthetic data and the significance of effective augmentation measurement in improving translation performance. Furthermore, hallucinations within NMT systems, indicate better integration approaches, such as MaTilDA, for dealing with large data sources [21]. Liang et al. [15] address computational problems such as data scarcity, randomness, robust model convergence, hyperparameter sensitivity, and computing costs, emphasizing the importance of efficient frameworks such as DM-CT. Finally, structural disparities and unequal data distribution pose considerable obstacles in multilingual translations. San et al. [13] discuss issues in Thai-Myanmar-English translations caused by structural inequality in writing systems and imbalanced cross-language data distribution, suggesting better fine-tuning and prompting tactics.

These studies highlight the complex issues of multilingual translation. Enhanced data augmentation approaches, improved pre-processing procedures, and novel ways for improving translation accuracy and contextual meaning are needed to address these challenges.

IV. CONCLUSIONS AND RECOMMENDATIONS

Data augmentation techniques, including back translation, synthetic data generation, sentence manipulation, and advanced transformations, consistently demonstrate improved translation quality for low-resource languages. This technique increases the diversity and size of the training dataset, thereby contributing to better model performance. However, there are still several challenges, such as grammar alignment issues, semantics, and synthetic data quality, as well as dealing with words outside the vocabulary.

The theoretical contributions of this review, particularly in classifying augmentation techniques, provide a structured foundation for future research. On the practical side, the insights from this review are directly applicable to real-world machine translation projects. The choice of augmentation techniques plays a crucial role in determining performance, especially across different language pairs and dataset sizes.

Some recommendations from the review include: (1) making efforts to collect more diverse and rich data sets, including leveraging community contributions and exploring new parallel data sources; (2) focusing future research on refining augmentation techniques and combining multiple methods to address grammatical and semantic problems; (3) creating an efficient framework to manage computational costs and hyperparameter sensitivity while exploring techniques like consistency training and iterative self-training to optimize model performance; and (4) although BLEU is becoming the primary metric for translation quality assessment, future research should emphasize developing and standardizing metrics for more accurate evaluation, such as combining BERTScore and COMET.

The combination of theoretical insights and practical applications presented in this review offers a comprehensive approach to addressing the challenges faced in low-resource language translation. This balanced approach provides a solid foundation for future advancements in the field, encouraging further refinement of both data augmentation techniques and evaluation metrics.

References

[1] Y. Yuyun *et al.*, "Enhancing Neural Machine Translation Model for Low-Resource Languages: A Case Study of Indonesian to Mamuju," *SSRN*, 2024, doi: https://dx.doi.org/10.2139/ssrn.4872856.

[2] B. Zoph, D. Yuret, J. May, and K. Knight, "Transfer Learning for Low-Resource Neural Machine Translation," Apr. 2016.

[3] T. Kocmi and O. Bojar, "Trivial Transfer Learning for Low-Resource Neural Machine Translation," in *Proceedings of the Third Conference on Machine Translation: Research Papers*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2018, pp. 244–252. doi: 10.18653/v1/W18-6325.

[4] 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)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2017, pp. 567–573. doi: 10.18653/v1/P17-2090.

[5] S. Edunov, M. Ott, M. Auli, and D. Grangier, "Understanding Back-Translation at Scale," Aug. 2018.

[6] B. K. Yazar, D. Ö. Şahın, and E. Kiliç, "Low-Resource Neural Machine Translation: A Systematic Literature Review," *IEEE Access*, vol. 11, pp. 131775–131813, 2023, doi: 10.1109/ACCESS.2023.3336019.

[7] S. M. Ul Qumar, M. Azim, and S. M. K. Quadri, "Neural Machine Translation: A Survey of Methods used for Low Resource Languages," in *2023 10th International Conference on Computing for Sustainable Global Development (INDIACom)*, 2023, pp. 1640–1647.

[8] B. Kitchenham, O. Pearl Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, "Systematic literature reviews in software engineering – A systematic literature review," *Inf Softw Technol*, vol. 51, no. 1, pp. 7–15, Jan. 2009, doi: 10.1016/j.infsof.2008.09.009.

[9] D. Stefanovic, S. Havzi, D. Nikolic, D. Dakic, and T. Lolic, "Analysis of the Tools to Support Systematic Literature Review in Software Engineering," *IOP Conf Ser Mater Sci Eng*, vol. 1163, no. 1, p. 012013, Aug. 2021, doi: 10.1088/1757-899X/1163/1/012013.

[10] Z. Yu and H. Zhang, "Filtered data augmentation approach based on model competence evaluation," *Physical Communication*, vol. 62, p. 102253, Feb. 2024, doi: 10.1016/j.phycom.2023.102253.

[11] V. Khenglawt, S. R. Laskar, P. Pakray, and A. K. Khan, "Addressing data scarcity issue for English–Mizo neural machine translation using data augmentation and language model," *Journal of Intelligent & Fuzzy Systems*, vol. 46, no. 3, pp. 6313–6323, Mar. 2024, doi: 10.3233/JIFS-235740.

[12] H. Zhang, Y. Ji, N. Wu, and M. Lu, "A Mongolian–Chinese Neural Machine Translation Method Based on Semantic-Context Data Augmentation," *Applied Sciences*, vol. 14, no. 8, p. 3442, Apr. 2024, doi: 10.3390/app14083442.

[13] M. E. San, S. Usanavasin, Y. K. Thu, and M. Okumura, "A Study for Enhancing Low-resource Thai-Myanmar-English Neural Machine Translation," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 23, no. 4, pp. 1–24, Apr. 2024, doi: 10.1145/3645111.

[14] H. Vu and N. D. Bui, "On the scalability of data augmentation techniques for low-resource machine translation between Chinese and Vietnamese," *Journal of Information and Telecommunication*, vol. 7, no. 2, pp. 241–253, Apr. 2023, doi: 10.1080/24751839.2023.2186625.

[15] X. Liang, R. Mao, L. Wu, J. Li, M. Zhang, and Q. Li, "Enhancing Low-Resource NLP by Consistency Training With Data and Model Perturbations," *IEEE/ACM Trans Audio Speech Lang Process*, vol. 32, pp. 189–199, 2024, doi: 10.1109/TASLP.2023.3325970.

[16] T. N. Quoc, H. Le Thanh, and H. P. Van, "Low-Resource Neural Machine Translation Improvement Using Data Augmentation Strategies," *Informatica*, vol. 47, no. 3, Aug. 2023, doi: 10.31449/inf.v47i3.4761.

[17] S. Kchaou, R. Boujelbane, and L. Hadrich, "Hybrid Pipeline for Building Arabic Tunisian Dialect-standard Arabic Neural Machine Translation Model from Scratch," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 22, no. 3, pp. 1–21, Mar. 2023, doi: 10.1145/3568674.

[18] F. Li, C. Chi, H. Yan, B. Liu, and M. Shao, "STA: An efficient data augmentation method for low-resource neural machine translation," *Journal of Intelligent & Fuzzy Systems*, vol. 45, no. 1, pp. 121–132, Jul. 2023, doi: 10.3233/JIFS-230682.

[19] X. Liu, J. He, M. Liu, Z. Yin, L. Yin, and W. Zheng, "A Scenario-Generic Neural Machine Translation Data Augmentation Method," *Electronics (Basel)*, vol. 12, no. 10, p. 2320, May 2023, doi: 10.3390/electronics12102320.

[20] A. Solyman *et al.*, "Optimizing the impact of data augmentation for low-resource grammatical error correction," *Journal of King Saud University - Computer and Information Sciences*, vol. 35, no. 6, p. 101572, Jun. 2023, doi: 10.1016/j.jksuci.2023.101572.

[21] V. M. Sánchez-Cartagena, M. Esplà-Gomis, J. A. Pérez-Ortiz, and F. Sánchez-Martínez, "Non-Fluent Synthetic Target-Language Data Improve Neural Machine Translation," *IEEE Trans Pattern Anal Mach Intell*, vol. 46, no. 2, pp. 837–850, Feb. 2024, doi: 10.1109/TPAMI.2023.3333949.

[22] T.-V. Ngo, P.-T. Nguyen, V. V. Nguyen, T.-L. Ha, and L.-M. Nguyen, "An Efficient Method for Generating Synthetic Data for Low-Resource Machine Translation," *Applied Artificial Intelligence*, vol. 36, no. 1, Dec. 2022, doi: 10.1080/08839514.2022.2101755.

[23] S. Rahman-Laskar, R. Manna, P. Pakray, and S. Bandyopadhyay, "A Domain Specific Parallel Corpus and Enhanced English-Assamese Neural Machine Translation," *Computación y Sistemas*, vol. 26, no. 4, Dec. 2022, doi: 10.13053/cys-26-4-4423.

[24] S. R. Laskar, A. F. U. R. Khilji, P. Pakray, and S. Bandyopadhyay, "Improved neural machine translation for low-resource English–Assamese pair," *Journal of Intelligent & Fuzzy Systems*, vol. 42, no. 5, pp. 4727–4738, Mar. 2022, doi: 10.3233/JIFS-219260.

[25] M. Maimaiti, Y. Liu, H. Luan, and M. Sun, "Data augmentation for low-resource languages NMT guided by constrained sampling," *International Journal of Intelligent Systems*, vol. 37, no. 1, pp. 30–51, Jan. 2022, doi: 10.1002/int.22616.

[26] S. Lee *et al.*, "A Survey on Evaluation Metrics for Machine Translation," *Mathematics*, vol. 11, no. 4, p. 1006, Feb. 2023, doi: 10.3390/math11041006.

[27] N. Moghe, T. Sherborne, M. Steedman, and A. Birch, "Extrinsic Evaluation of Machine Translation Metrics," Dec. 2022.

[28] Y. Graham and Q. Liu, "Achieving Accurate Conclusions in Evaluation of Automatic Machine Translation Metrics," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2016, pp. 1–10. doi: 10.18653/v1/N16-1001.