Cross-Lingual Data Augmentation For Thai Question-Answering

Parinthapat Pengpun[‡], **Can Udomcharoenchaikit**[†],
Weerayut Buaphet[†], **Peerat Limkonchotiwat**[†]
[‡]Bangkok Christian International School, Thailand
[†]School of Information Science and Technology, VISTEC, Thailand
parinzee@protonmail.com
{canu_pro,weerayut.b_s20,peerat.l_s19}@vistec.ac.th

Abstract

This paper presents an innovative data augmentation framework with data quality control designed to enhance the robustness of Question Answering (QA) models in low-resource languages, particularly Thai. Recognizing the challenges posed by the scarcity and quality of training data, we leverage data augmentation techniques in both monolingual and cross-lingual settings. Our approach augments and enriches the original dataset, thereby increasing its linguistic diversity and robustness. We evaluate the robustness of our framework on Machine Reading Comprehension and the experimental results illustrate the potential of data augmentation to effectively increase training data and improve model generalization in low-resource language settings, offering a promising direction for the data augmentation manner.

1 Introduction

Question Answering (QA) systems are algorithms developed to answer questions posed in a natural language format accurately. A primary task in QA is Machine Reading Comprehension (MRC), which aims to extract answers from text passages given a question. Previous works demonstrate that improving the performance of MRC increases the accuracy in real-world applications, i.e., conversational chatbots (Yang et al., 2023; Jin and Lee, 2022; Hardalov et al., 2019).

While the performance of QA systems in English is largely considered to be a solved problem, it is still an open problem in low-resource languages, i.e., Thai. Recent works in QA demonstrate that the performance gap in QA systems for Thai and English is wide. For example, the baseline performance of English on the XQuAD (Artetxe et al., 2019) dataset has an F1 score of 83.5, while only 42.7 on Thai. Moreover, only five of the research works (Noraset et al., 2021; Decha and Patanukhom, 2017; Hochreiter and Schmidhuber, 1997; Wongpraomas et al., 2022; Limkonchotiwat et al., 2022b) has been published on QA for the Thai language within the last five years. Such disparity has led to a significant gap in the capabilities of NLP systems between high-resource languages and low-resource ones (Artetxe et al., 2019; Lewis et al., 2019; Wongpraomas et al., 2022). The robustness of NLP applications in low-resource languages leaves much to be desired due to the lack of extensive and diverse language data that covers all aspects of the language.

Existing literature offers several methods to mitigate the problem of robustness specifically for low data in QA, such as transfer learning (Pandya et al., 2021), back translation (Riabi et al., 2021), and the use of multilingual language models (Kumar et al., 2022). However, these techniques present drawbacks, e.g., transfer learning's success hinges largely on the relatedness of the source and target languages. In addition, multi-lingual models are impacted by the imbalanced data distribution and often do not perform as well as monolingual models (Artetxe et al., 2019; Lewis et al., 2019). For example, WangchanBERTa (Lowphansirikul et al., 2021), a monolingual RoBERTa-based (Liu et al., 2019) model trained explicitly on the Thai Language, outperforms multilingual models (XLM-R (Lample and Conneau, 2019) and mBERT (Devlin et al., 2018)). Additionally, there is a noticeable lack of studies focusing on these methods in the context of the Thai language, rendering the extension of such strategies unclear. The issue of robustness and generalization also remains largely unaddressed in recent literature, thereby leaving a significant gap in the field.

To improve performance in a low-resource setting, we propose an automatic framework for improving out-of-distribution robustness in QA. Our framework integrates back translation, word replacement, large language model (LLM) automated paraphrase generation, and LLM automated gram-

Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP, pages 193–203

mar correction to construct a 10-way parallel corpus. This corpus features multiple varied sentence formulations to encourage robustness and generalization. While our framework heavily leverages machine translation (MT), which allows our augmentations to leverage the vast library of English augmentations thereby improving the robustness of Thai QA, it is our quality control system that sets our work apart. By rigorously removing noisy samples from the data—filtering the distances between the semantic representations of the augmented and the original data—the system ensures that the dataset obtained contains an optimal signal-to-noise ratio.

As shown in Figure 1, the proposed framework works as follows. Firstly, we aggregate, clean and normalize our datasets: TyDiQA (Clark et al., 2020), XQuAD (Artetxe et al., 2019), Iapp Wiki QA (Viriyayudhakorn and Polpanumas, 2021), and Thai QA (Trakultaweekoon et al., 2019). Then, we translate all questions into English and back-translate to Thai using Google Translate. Next, the gpt-3.5-turbo-0301 model is used in the third stage for grammar correction and paraphrasing of these translated questions. In the fourth stage, the Quality Controlled Paraphrase Generation (QCPG) (Bandel et al., 2022) model generates additional paraphrases for the translated questions, which are then translated back to Thai. In addition, we leverage WordNet, Thai2Fit, Thai2Transformers, and Large Thai Word2Vec (LTW2Vec) for word-replacement on the Thai questions without back-translation. Lastly, we utilize our quality control mechanism to filter noisy augmented samples from our corpora. The end corpora is a versatile, 10-way parallel corpus, ready for use in the MRC task.

We evaluate the effectiveness of our framework compared to common augmentations without any cross-lingual augmentations on standard QA datasets for the MRC task. Specifically we test for out-of-distribution by using completely different MRC Datasets for evaluation. For more information please refer to our evaluation card in the Appendix. The experimental results demonstrate that our framework significantly enhances the robustness and generalization of Thai QA systems. For example, we improve the performance (Exact Match/F1) of WangchanBERTa on TyDiQA and XQuAD datasets from 39.46/54.87 and 34.92/48.80, to 42.76/56.51 and 35.25/49.43, respectively. Our design analysis of the quality control mechanism demonstrates that our mechanism

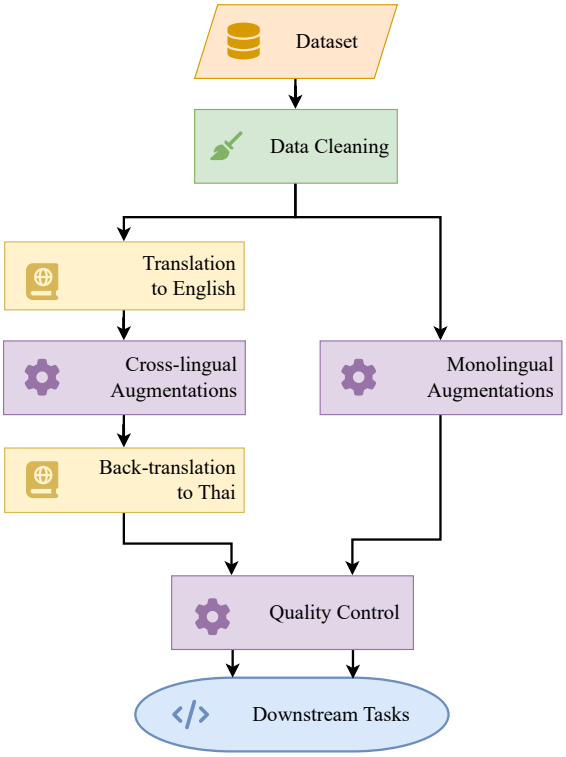

Figure 1: Our proposed automatic framework for improving the robustness of Thai QA systems.

removes noisy data from augmentation schemes and decreases the training time of MRC from 80 to 50 minutes (∼38.25% faster).

We summarize the contribution of our work as follows:

- We propose a unified framework leveraging cross-lingual augmentations to improve Thai QA performance. The framework consists of 10 augmentations, including monolingual and cross-lingual settings.
- We release the first extensively cleaned, 10-way parallel QA corpus which unifies all publicly available QA datasets for the Thai language.
- We conduct an extensive study on large-scale experiments: 10 augmentation schemes, two benchmark datasets, and two ablation studies.
- We release all the datasets, code, and trained models at this GitHub Repo.

2 Related Work

2.1 Thai Question Answering

Numerous studies have attempted to address Thai QA systems in monolingual and multilingual settings. Noraset et al. (2021) introduced Wabi QA, a Wikipedia-based QA system that integrated both a retrieval model and an MRC model. Their method used a bidirectional Long Short-Term

Memory (LSTM) (Hochreiter and Schmidhuber, 1997) model as the MRC model and surpassed several extant methodologies in Thai QA. Similarly, Decha and Patanukhom (2017) proposed an open-domain Thai QA system that employed a keyword extraction system in combination with rule-based word segmentation and neural network-based sentence segmentation. Furthermore, an alternative QA system was put forward by Wongpraomas et al. (2022), which used a rule-based pattern-matching method relying on regular expressions and cosine similarity with a SQL database. Although these methods exhibited strong performance within their respective benchmarks, they fall short in providing results on any standardized or widely-recognized datasets, thereby rendering an accurate performance evaluation challenging. Furthermore, it is apparent that there is a lack of integration of Transformer-based models (Vaswani et al., 2017), and the concerns pertaining to system generalization and robustness are not addressed.

There are also multilingual approaches to tackling Thai QA. A common technique is to train a transformer model on multilingual QA datasets, such as XQuAD (Artetxe et al., 2019) and TyDiQA (Clark et al., 2020). Recently, there have been multilingual QA works focused on improving multilingual performance. Asai et al. (2021) proposed CORA, a unified multilingual many-to-many QA model. This work consisted of two QA models: (i) a multilingual dense passage retriever (mDPR); and (ii) an autoregressive answer generation model (mGEN) trained on multilingual Wikpedia. Notably, it did not use translation and can generalize to languages without annotated data sources but with large-scale training data. Limkonchotiwat et al. (2022b) proposed CL-ReLKT, a cross-lingual ReQA learning approach using knowledge transfer. In particular, this work distilled the performance of high-resource to low-resource languages, resulting in increased performance in a wide range of low-resource languages, including Thai.

Despite these advances, it becomes apparent that there are still significant gaps in the literature concerning implementing multilingual models in Thai QA. Specifically, these works do not address the robustness aspect of the model nor the specifics of the Thai language. Our research recognizes and addresses these gaps within the literature. Our approach employs a Transformer-based model, introduces a novel data augmentation technique to enhance robustness and generalization, and benchmarks results on well-established datasets.

2.2 Improving Robustness in QA

Robustness plays a crucial role in QA systems to perform efficiently on unseen data with varying distributions (Hupkes et al., 2023). There are numerous aspects to robustness in QA such as resistance to adversarial questions, ability to generalize to unseen domains, and capability to understand different phrasings or expressions of the same sentence. Various strategies have been proposed to improve this aspect of QA models, applicable across languages.

Bartolo et al. (2021) proposed an adversarial QA data generation pipeline that automatically generates question-answer pairs to bolster the QA model's robustness against adversarial questions. This pipeline improved both F1 and exact match (EM) scores of the models trained on adversarial synthetic data. Khashabi et al. (2020) proposed using natural human-driven perturbations on a small dataset to improve model performance instead of constructing new large-scale datasets. Each perturbation is independently verified to ensure minimal yet meaningful changes and that the questions are answerable. The experimental results demonstrated that the proposed model is more generalized and robust when dealing with small variations in a question.

The aforementioned studies have introduced remarkable advancements in the field of QA robustness. However, there are notable gaps in the literature regarding the application of these techniques in low-resource languages, specifically Thai. Despite the proven effectiveness of these strategies in languages with abundant resources, their applicability to Thai remains under-explored. This is particularly significant given the unique challenges presented by the Thai language, such as word and sentence boundary issues and scarcity of resources. Moreover, these works do not provide an established best practice for adapting these techniques to Thai or similar languages, presenting an additional obstacle to their application.

Despite this, we observe that many of the main ideas in the literature apply to the Thai language. Thus, our research aims to bridge these gaps by creating a framework that applies robustness-improvement techniques to Thai QA. We aim to develop and implement an efficient cross-lingual

data augmentation system that can help overcome the specific challenges associated with the Thai language. Although our framework was designed for Thai specifically, it can easily be modified for usage in other languages by adjusting the translation part and the monolingual embedding-based augmentations.

3 Methodology

3.1 Overview

To increase the out-of-distribution robustness of QA models, we introduce a data augmentation framework (Figure 1), which employs back translation, word replacement, and the application of Large Language Models (LLMs) for automated paraphrase generation and grammar correction. This also includes cross-lingual data augmentation strategies that allow us to use high-performance NLP tools that only perform well in high-resource languages, such as paraphrase generation models. Moreover, our framework also incorporates a quality control step that can remove noisy samples to ensure the quality of the generated data.

Our framework enhances the quality of training data by synthetically adding linguistic diversity. This allows QA models to handle a diverse array of queries more effectively and improve their generalization. We detail how we formulate and augment our training data for enhanced robustness as follows.

3.2 Data Selection and Preprocessing

We select two standard multilingual QA datasets, such as TyDiQA (Clark et al., 2020) and XQuAD (Artetxe et al., 2019) since both datasets had Thai in their datasets. However, we found that there is no Thai training data for XQuAD. Thus, we add standard Thai QA datasets: Iapp Wiki QA (Viriyayudhakorn and Polpanumas, 2021) and Thai QA (Trakultaweekoon et al., 2019) into our framework. Then, we perform data cleaning by first stripping out HTML, XML, and other markup syntaxes inside the questions, answers, and contexts. We also drop rows that contain invalid answers and markup syntax that could not be removed automatically. Such invalid answers include incomplete answers and answers that start with a Thai tone mark or syllable. Then, we dropped all the duplicated questions and context sets to prevent data leakage into the test set. We then format all the questions to have a question mark at the end for extra clarity. Lastly, we realign all answer start positions (labels) to match the cleaned context with the cleaned answers.

3.3 Data Augmentation

As discussed in Section 2.1, Thai QA systems lack the generalization ability to handle unseen questions. The studies from previous work demonstrated promising results to improve the generalization by applying data augmentation schemes. However, previous augmentation schemes do not apply to Thai, it is unclear how to extend those beyond a monolingual setting. Thus, we apply these augmentation schemes using the back-translation (TH→EN and EN→TH) method [1] on questions of the training data.

We present 10 data augmentation schemes to enhance linguistic diversity and improve the generalization of our model. We split the data augmentation into two groups: *cross-lingual data augmentation* and *monolingual data augmentation*. The first group used the back-translation process (TH→EN→TH) with English augmentation schemes. The second group used monolingual data augmentation without the back-translation process.

Cross-lingual data augmentation. We employ four off-the-shelf data augmentation schemes from English QA to improve the robustness of Thai QA as follows:

- *Back-translation*: We utilize Google Translate to translate our Thai questions to English, then translate those questions back to Thai again. The Thai texts before and after the translation will be changed, which helps enhance the model's robustness, as it is trained to understand and respond to a broader set of phrasings. Moreover, it boosts the model's generalization capacity by enabling it to recognize and respond appropriately to imperfect translations and improper grammar (Zhang et al., 2021; Limkonchotiwat et al., 2022a).
- *LLM Grammar Error Correction (GEC)*: We utilized the gpt-3.5-turbo-0301 model [2] to perform GEC on English-translated data from the previous step, then used Google Translate to translate the corrected dataset back to Thai. This approach not only allows the model to understand

[1] We employ Google NMT as the back-translation method where the version date is 19 May 2023

[2] We use the version of 03.01.23. If an updated version is used, exact results may be hard to replicate. However, results obtained should be similar or better.

the core meaning of questions amidst grammatical inaccuracies but also enhances the model's robustness by expanding its exposure to a variety of corrected syntax. Furthermore, it improves the model's capacity to handle diverse and complex grammatical structures, thereby fostering better generalization.

- *LLM Paraphrase*: We also utilized the gpt-3.5-turbo-0301 model to generate a paraphrase of each question on the English translation of our questions, then used Google Translate to translate the paraphrased questions back to Thai. Prompting was done to allow GPT to assume the meaning of ambigious grammatically incorrect translations. For example, "You are a highly skilled language model AI that returns only one line of grammatically perfect text. Your task is to evaluate the text below and correct its grammar. Even if the text is incomplete or unintelligible, YOU MUST make a grammatical correction, you can make assumptions about the intended meaning. If the text is grammatically correct, do not change it. Your output should be presented WITH ONLY the corrected text IN ONE LINE and without any extra dialogue from you." This strategy improves the QA model's robustness by exposing it to a wider range of expressions and phrasing.
- *QCPG Paraphrase*: We utilize QCPG (Bandel et al., 2022), specifically the qcpg-questions model, to generate a paraphrase of each question on the English translation and the LLM GEC set of our questions, then used Google Translate to translate the paraphrased questions back to Thai. In total, three distinct paraphrase sets were generated by experimenting with three different values of settings applied to each dataset. These variations result from adjusting three QCPG settings, namely lexical, syntactic, and semantic, across the range of 0 to 1. Specifically, the chosen values for these settings were 0.2, 0.5, and 0.8, leading to the creation of three separate datasets. We select the best-performing dataset to be included in our main corpora (the 0.8 setting).
- *LLM GEC + QCPG Paraphrase*: We apply the QCPG Paraphrase atop the LLM GEC augmentation. This decision was driven by the hypothesis that the noisy translation might influence the QCPG model's performance and that rectifying this issue would yield an improved performance in paraphrasing. By employing this method, the model's robustness is enhanced, as it is exposed to modified yet semantically congruent expressions. Furthermore, this dataset variation promotes the model's generalization capabilities, facilitating its comprehension of the diverse manners in which a question might be framed.

Monolingual data augmentation. All the methods below are based on synonym replacement using pre-trained word embeddings. This method enriches the dataset and trains the model to recognize and understand equivalent words, resulting in improving the model's generalization ability and semantic understanding. These augmentations were selected since they represent monolingual methods commonly found in text augmentation and have been widely used. One such commonly used method is word or token replacement.

We leverage five monolingual pre-trained models as follows: (i) the **Thai WordNet** (Thoongsup et al., 2009) is employed for embedding-based word replacement; (ii) **Thai2Fit** (Polpanumas and Phatthiyaphaibun, 2021), a Thai adaptation of ULMFit (Howard and Ruder, 2018), which we used in the same manner as Thai WordNet; (iii) **Thai2Transformers** (Lowphansirikul et al., 2021) which utilizes the embeddings from the WangchanBERTa model to perform word-replacement; (iv) **LTW2Vec** (Phatthiyaphaibun, 2022), a comprehensive Thai Word2Vec model; and (v) **FastText** (Bojanowski et al., 2017) is also utilized for embedding-based word substitution.

3.4 Quality Control of Data Augmentation

As discussed in the data augmentation schemes, we use back-translation as the backbone of the cross-lingual augmentation. As the translation process is imperfect, this will lead to a variation of sentences with a similar meaning being produced. Despite this, the produced translation can be nearly perfect to almost unintelligible. Thus, we need to control the quality of our augmentation training data to not let low-quality paraphrases into our dataset.

One such consideration is that a good paraphrase should have a similar meaning between the original and augmented texts (Bandel et al., 2022). Thus, we chose to control the quality by evaluating the semantic score of augmentation datasets and selecting the best semantic score threshold (calculated on the development dataset) to find high-quality

data for training QA models. We calculate the semantic score using the cosine distance between the embedding of augment and original questions obtained from Multilingual Universal Sentence Encoder (mUSE) (Yang et al., 2020). For monolingual augmentations, we can simply use the cosine distance between it and the original questions. In contrast, for cross-lingual augmentations, we use the harmonic distance (HD) between two measures: (i) the distance between the original questions and the English augmentation and (ii) the between the original questions and the translated English augmentation. We calculate the HD score as follows:

$$\text{HD} = \frac{2(1-\cos(t_{\text{org}}, t_{\text{en}}))(1-\cos(t_{\text{org}}, t_{\text{th}}))}{(1-\cos(t_{\text{org}}, t_{\text{en}})) + (1-\cos(t_{\text{org}}, t_{\text{th}}))} \quad (1)$$

where $\cos(\cdot)$ is cosine similarity, t_{org} is the original text before back-translation, t_{en} is the English text obtain from back-translation, t_{th} is the Thai text obtain from back-translation. While the arithmetic mean could oversimplify and skew results by providing equal weight to both measures, we opted for the harmonic mean, which is less sensitive to large discrepancies and offers a more balanced representation of the data. This approach ensures that neither the translation nor back-translation distances dominate the final score, yielding a more representative score that appreciates the intricacies of the multi-faceted nature of translation tasks. For more information and examples about how we calculate and select the augmentation ratio with the HD score, please see Appendix A.1.

4 Experimental Setting

4.1 Downstream tasks: MRC

We train the MRC model based on Wangchan-BERTa model (Lowphansirikul et al., 2021) on iAPP QA and ThaiQA datasets with a single V100 GPU for 140 hours with hyperparameters, as shown in Table 1. In addition, we report F1 and extract match (EM) scores for this task. For the hypothesis test, due to resource constraints, we were only able to test our results with one seed. Thus, we chose McNeMar due to its ability to work using one seed.

To incorporate our data augmentation schemes, we supplement the original training set with the augmented set of questions while maintaining the original context. To perform quality control, we benchmark each top-k% sample of the augmented data, systematically examining increments of 10% using harmonic distance. We select the best ratio on the validation set and show only the best score.

Hyperparameter	Value
Learning rate	$2e^{-5}$
Per-device train batch size	32
Per-device evaluation batch size	128
Gradient accumulation steps	2
Number of training epochs	20
Warmup ratio	0.2
Weight decay	0.01
Seed	42
Use fp16 precision	True

Table 1: Hyperparameters for the MRC task.

4.2 Datasets

- XQuAD (Artetxe et al., 2019). A cross-lingual QA dataset consists of 240 paragraphs and 1190 question-answer pairs from the development set of SQuAD v1.1, translated by professionals into 11 different languages, including Thai.
- iAPP QA (Viriyayudhakorn and Polpanumas, 2021). A Thai QA dataset from Thai Wikipedia consists of 9,170 question-answer pairs across 1,961 documents.
- ThaiQA (Trakultaweekoon et al., 2019). A Thai QA dataset from various domains of Wikipedia consists of 4051 question-answer pairs.
- TyDiQA (Clark et al., 2020). A multilingual QA dataset includes around 4500 questions in the Thai language.

For the evaluation setting, we use iAPP and ThaiQA as the training data for the MRC task while testing our model on XQuAD and TyDiQA datasets. Note that we use only Thai questions and context for multilingual datasets.

5 Experimental Results

We demonstrate the experimental results of the machine reading comprehension task in out-of-domain settings in Section 5.1. Section 5.2 demonstrates the analysis of the harmonic distance method on the performance and training speed efficiency. In addition, we present error analysis by comparing various augmentation texts with the original text in Section 5.3.

5.1 Machine Reading Comprehension

To identify the most efficacious augmentation strategy, we evaluate our MRC models with various

Augmentation	Ratio	Val EM/F1	TyDiQA EM/F1	XQuAD EM/F1
Original	N/A	50.75 / 62.40	39.46 / 54.87	34.92 / 48.80
Cross-lingual Augmentations				
Back translation	0.4	⬆ 0.05 / 0.27	⬆ 2.55 / 1.56 †	⬇ -1.02 / -0.76 †
LLM GEC	0.4	⬆ 0.31 / ⬇ -0.16	⬆ 2.66 / 1.23 †	⬆ 0.33 / 0.63 †
LLM Paraphrase	0.7	⬇ -0.88 / -0.47	⬆ 3.28 / 1.62 †	⬇ -2.55 / -1.37 †
QCPG	1.0	⬆ 0.36 / 0.22	⬆ 2.86 / 0.92 †	⬆ 0.16 / 0.21 †
LLM GEC + QCPG	0.7	⬆ 0.23 / 0.22	⬆ 3.30 / 1.64 †	⬇ -0.60 / ⬆ 0.16 †
Monolingual Augmentations				
WordNet	0.4	⬆ 1.60 / 1.32	⬆ 2.86 / 1.34 †	⬇ -1.28 / ⬆ 0.17 †
Thai2Fit	0.7	⬇ -0.48 / ⬆ 0.06	⬆ 2.26 / 1.13 †	⬇ -0.77† / -0.05
Thai2Transformers	0.4	⬆ 0.58 / 0.44	⬆ 2.70 / 1.39 †	⬇ -0.51 / ⬆ 1.40 †
LTW2Vec	1.0	⬆ 1.38 / 1.43	⬆ 1.84 / 1.75 †	⬇ -0.09 / ⬆ 1.23†
FastText	0.9	⬆ 0.27 / 0.33	⬆ 1.93 / 1.29 †	⬇ -2.21 / -1.13 †

Table 2: Optimal ratio for best performance in each augmentation— selected from validation scores. The ratio is obtained by performing the quality control method. † represents a significant result calculated from McNeMar's test.

ratios in out-of-domain settings. We discussed the experiment setup in Section 4.1.

Results. Table 2 exhibits the performance of the most effective models for each augmentation strategy, presenting both F1 and EM scores. The table also elucidates the optimal ratio of augmented to real data (calculated from the harmonic distance method). The experimental results demonstrate that using cross-lingual augmentations improves the performance of the original model in all test datasets except for Back translation and LLM Paraphrase. Moreover, we found that the performance of cross-lingual augmentations is higher than monolingual augmentations in the XQuAD dataset. For example, the QCPG method outperforms the FastText method by 0.93 EM score and 1.34 F1 score. While cross-lingual augmentations somewhat show reduced performance here, the overall results still surpass those of monolingual augmentations.

Discussion. The results in Table 2 substantiate the nuanced benefits of data augmentation techniques in MRC. While monolingual augmentations show promise of improvement in validation and TyDiQA datasets, their impact on the performance of the XQuAD dataset. For example, WordNet and Thai2Transformers increase the performance of validation and TyDiQA datasets while decreasing the performance on XQuAD compared to the original model. It is possible that these augmentations are too noisy, perhaps performing word replacement on a crucial section of the question, thus changing the semantic meaning altogether. In contrast, "LLM GEC" and "QCPG" are the most effective, delivering statistically significant improvements. This also implies that cross-lingual augmentations can improve the MRC's generalization better than the monolingual strategy.

5.2 Harmonic Distance (HD)

As stated in Section 3.4, we benchmark the performance of each dataset performing top-k% sampling in increments of 10% (see more information about top-k% in Appendix A.1). To investigate the efficacy of sampling using HD score, we choose to examine the best-performing augmentations: "LLM GEC" and "QCPG" on the TyDiQA test set.

Results. As shown in Figures 3 and 4, employing the quality control mechanism to filter noise data improves the MRC performance compared to the original model. We found that the best threshold of LLM GEC is 0.9, which improves the F1 score from the original model by 1.41 points. In addition, at the best ratio of QCPG, we observe the performance improvement of 1.69 points by the F1 score. Moreover, when comparing the whole dataset and only the 0.3 dataset, the performance remains identical with the training time decrease from ∼80 to ∼50 minutes (∼38.25% reduction).

Discussion. The HD score demonstrates its effectiveness by the compelling performance gains with lower data ratios—in this scenario, ratios of 0.2 and 0.3 outperform even a full dataset. Generally, using a lower ratio than 0.1 still results in equal or better performance, however, the ratio has to be searched

Original	LLM GEC	FastText	Ratio
โปแลนด์บอล เป็นการ์ตูนที่สร้างขึ้นโดยใคร? (Poland Ball is a cartoon created by who?	ใครเป็นผู้สร้างการ์ตูนโปลันด์บอล? (Who created Poland Ball?)	โปแลนด์โลก2026 เป็นการ์ตูน 4ช่องที่สร้างมากกว่าโดยแล้ว? (Poland World2026 is a cartoon 4 channel which created more than?)	0.20
แม่น้ำไนล์ตั้งอยู่ในทวีปใด? (In which continent is the Nile River located in?)	แม่น้ำไนล์ตั้งอยู่ที่ไหน? (Where is the Nile River located?)	บริเวฯกีปตั้งอยู่นำทวีปใด? (Boriwaekip is located in front of which continent?)	0.40
ประเทศเลบานอนอยู่ในภูมิภาคใด? (The country of Lebanon is located in which area?)	เลบานอนอยู่ที่ไหน? (Where is Lebanon?)	ประเทศ Lebanonข้อ อยู่ในแอฟริกาหรือ? (The country of Lebanon number is in Africa?)	0.60
ประพันธ์โดยใคร? (Composed by who?)	แต่งโดยใคร? (Produced by who?)	ป.ชื่นเช่นใคร? P. Cheen such who?	0.80

Figure 2: Dataset examples from the robustness augmentation (LLM GEC) compared to the original texts and poor augmentation (FastText) with varying ratios.

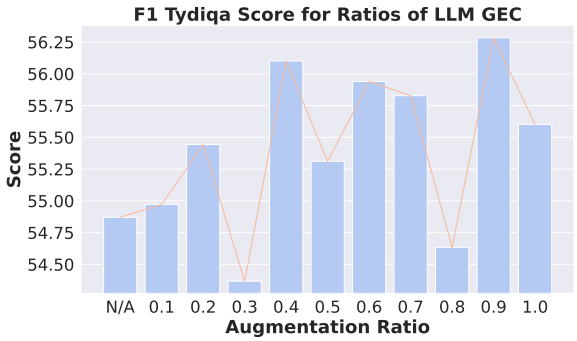

Figure 3: F1 Score of the LLM GEC dataset when fine-tuned with different augmentation ratios on TyDiQA.

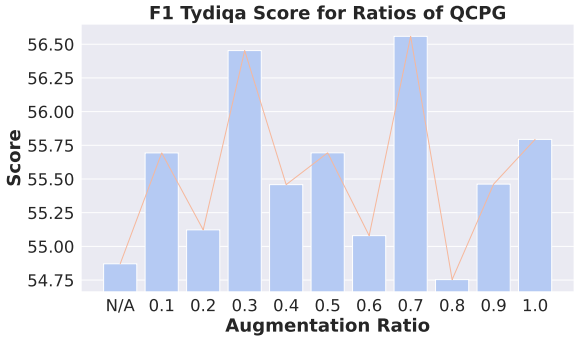

Figure 4: F1 Score of the QCPG dataset when finetuned with different augmentation ratios on the TyDiQA.

(see Appendix A.2). This underlines HD's value in reducing computational load while optimizing performance, signifying its potential as a cornerstone in future data augmentation strategies.

5.3 Error Analysis

In this study, we demonstrate error analysis from different datasets and augmentation ratios to decipher why certain augmentations and ratios perform better and to identify the characteristics of augmentations at these specific ratios. In addition, we use the augmented datasets from the LLM GEC and FastText augmentation schemes.

Figure 2 presents sentences from varying ratios from the best and worst performing augmentation. The LLM GEC augment scheme maintains better semantic meaning than the FastText augmentation. Sentences at around 0.1-0.2 ratios tend to produce more conservative paraphrases, preserving the semantic integrity of the original text. For the ratio between 0.4 and 0.6, sentences lean toward more liberal paraphrasing strategies, often omitting some keywords and introducing higher noise levels into the data. In addition, we observe that sentences at 0.8 replace key-specific words with more ambiguous synonyms. However, the augmentation results from FastText demonstrate that it fails to produce robust augmentation in all cases, resulting in performance degradation (Table 2). Additionally, the results from Figures 3 and 4 also support our findings, indicating that top-performing scores can be achieved without utilizing the entire dataset.

6 Conclusion and Future Work

We present an automatic framework for improving robustness in QA tasks. The experimental results demonstrate that our cross-lingual augmentation improved the performance of Thai QA more consistently than monolingual augmentations. Moreover, we present the quality control for data augmentations, which decrease the training time and maintain the performance with only 20% of total augmented data. For future work, we would like to explore the application and task of our augmentation and quality control approaches to improve the generalization, such as conversational chatbots (Yang et al., 2023; Jin and Lee, 2022; Hardalov et al., 2019) and retrieval QA (Asai et al., 2021; Limkonchotiwat et al., 2022b).

References

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. On the cross-lingual transferability of monolingual representations. *CoRR*, abs/1910.11856.

Akari Asai, Xinyan Yu, Jungo Kasai, and Hanna Hajishirzi. 2021. One question answering model for many languages with cross-lingual dense passage retrieval. In *Advances in Neural Information Processing Systems*, volume 34, pages 7547–7560. Curran Associates, Inc.

Elron Bandel, Ranit Aharonov, Michal Shmueli-Scheuer, Ilya Shnayderman, Noam Slonim, and Liat Ein-Dor. 2022. Quality controlled paraphrase generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 596–609, Dublin, Ireland. Association for Computational Linguistics.

Max Bartolo, Tristan Thrush, Robin Jia, Sebastian Riedel, Pontus Stenetorp, and Douwe Kiela. 2021. Improving question answering model robustness with synthetic adversarial data generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8830–8848, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*.

Hatsanai Decha and Karn Patanukhom. 2017. Development of thai question answering system. In *Proceedings of the 3rd International Conference on Communication and Information Processing, ICCIP 2017, Tokyo, Japan, November 24-26, 2017*, pages 124–128. ACM.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Momchil Hardalov, Ivan Koychev, and Preslav Nakov. 2019. Machine reading comprehension for answer reranking in customer support chatbots. *Information*, 10(3).

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Dieuwke Hupkes, Mario Giulianelli, Verna Dankers, Mikel Artetxe, Yanai Elazar, Tiago Pimentel, Christos Christodoulopoulos, Karim Lasri, Naomi Saphra, Arabella Sinclair, Dennis Ulmer, Florian Schottmann, Khuyagbaatar Batsuren, Kaiser Sun, Koustuv Sinha, Leila Khalatbari, Maria Ryskina, Rita Frieske, Ryan Cotterell, and Zhijing Jin. 2023. State-of-the-art generalisation research in nlp: A taxonomy and review.

Nayoung Jin and Hana Lee. 2022. StuBot: Learning by teaching a conversational agent through machine reading comprehension. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3008–3020, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Daniel Khashabi, Tushar Khot, and Ashish Sabharwal. 2020. More bang for your buck: Natural perturbation for robust question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 163–170, Online. Association for Computational Linguistics.

Gokul Karthik Kumar, Abhishek Gehlot, Sahal Shaji Mullappilly, and Karthik Nandakumar. 2022. MuCoT: Multilingual contrastive training for question-answering in low-resource languages. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 15–24, Dublin, Ireland. Association for Computational Linguistics.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *CoRR*, abs/1901.07291.

Patrick S. H. Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2019. MLQA: evaluating cross-lingual extractive question answering. *CoRR*, abs/1910.07475.

Peerat Limkonchotiwat, Wuttikorn Ponwitayarat, Lalita Lowphansirikul, Can Udomcharoenchaikit, Ekapol Chuangsuwanich, and Sarana Nutanong. 2022a. ConGen: Unsupervised control and generalization distillation for sentence representation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6467–6480, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Peerat Limkonchotiwat, Wuttikorn Ponwitayarat, Can Udomcharoenchaikit, Ekapol Chuangsuwanich, and Sarana Nutanong. 2022b. CL-ReLKT: Cross-lingual language knowledge transfer for multilingual retrieval question answering. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2141–2155, Seattle, United States. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis,

Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Lalita Lowphansirikul, Charin Polpanumas, Nawat Jantrakulchai, and Sarana Nutanong. 2021. Wangchanberta: Pretraining transformer-based thai language models. *CoRR*, abs/2101.09635.

Thanapon Noraset, Lalita Lowphansirikul, and Suppawong Tuarob. 2021. Wabiqa: A wikipedia-based thai question-answering system. *Information processing & management*, 58(1):102431.

Hariom Pandya, Bhavik Ardeshna, and Brijesh Bhatt. 2021. Cascading adaptors to leverage English data to improve performance of question answering for low-resource languages. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pages 544–549, National Institute of Technology Silchar, Silchar, India. NLP Association of India (NLPAI).

Wannaphong Phatthiyaphaibun. 2022. Ltw2v: The large thai word2vec.

Charin Polpanumas and Wannaphong Phatthiyaphaibun. 2021. thai2fit: Thai language implementation of ulmfit.

Arij Riabi, Thomas Scialom, Rachel Keraron, Benoît Sagot, Djamé Seddah, and Jacopo Staiano. 2021. Synthetic data augmentation for zero-shot cross-lingual question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7016–7030, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sareewan Thoongsup, Thatsanee Charoenporn, Kergrit Robkop, Tan Sinthurahat, Chumpol Mokarat, Virach Sornlertlamvanich, and Hitoshi Isahara. 2009. Thai WordNet construction. In *Proceedings of the 7th Workshop on Asian Language Resources (ALR7)*, pages 139–144, Suntec, Singapore. Association for Computational Linguistics.

Kanokorn Trakultaweekoon, Santipong Thaiprayoon, Pornpimon Palingoon, and Anocha Rugchatjaroen. 2019. The first wikipedia questions and factoid answers corpus in the thai language. In *2019 14th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, pages 1–4.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Kobkrit Viriyayudhakorn and Charin Polpanumas. 2021. iapp_wiki_qa_squad.

Pongsathorn Wongpraomas, Chitsutha Soomlek, Wanna Sirisangtragul, and Pusadee Seresangtakul. 2022. Thai question-answering system using pattern-matching approach. In *2022 1st International Conference on Technology Innovation and Its Applications (ICTIIA)*, pages 1–5.

Changlin Yang, Siye Liu, Sen Hu, Wangshu Zhang, Teng Xu, and Jing Zheng. 2023. Improving knowledge production efficiency with question answering on conversation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 225–234, Toronto, Canada. Association for Computational Linguistics.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2020. Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 87–94, Online. Association for Computational Linguistics.

Yan Zhang, Ruidan He, Zuozhu Liu, Lidong Bing, and Haizhou Li. 2021. Bootstrapped unsupervised sentence representation learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5168–5180, Online. Association for Computational Linguistics.

A Appendix

A.1 How Ratio Was Selected

As shown in Figure 5, we demonstrate how the ratio was selected. We calculate the HD from the LLM GEC method and arrange the semantic distance from lowest to highest to formulate the distance distribution. We then select the top-k% of the distribution as the training data, the k value can be between 0.0 (using only the original training data) to 1.0 (using entirely augmentation and training corpora). This technique can control the quality of the final dataset by maintaining a good signal-to-noise ratio of the dataset, maintaining a similar meaning between the original and augmented texts.

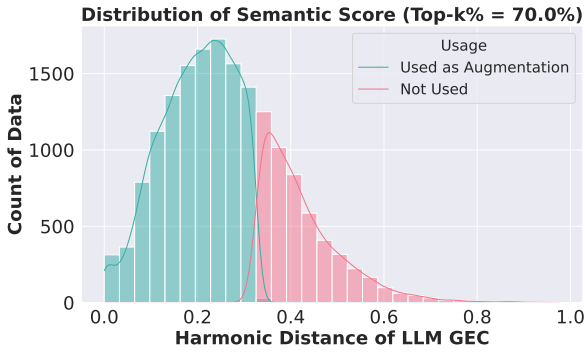

Figure 5: Example distribution of semantic score on the LLM GEC augmentation. With top-k% = 0.7, we will select the top 70% of the dataset that has the closest semantic similarity to the original data. The blue distribution indicates that data was selected for augmentation, while the red line indicates unselected data for this top-k% value.

Motivation			
Practical	*Cognitive*	*Intrinsic*	*Fairness*
☐			
Generalisation type			
Compositional	*Structural*	*Cross Task*	*Cross Language*	*Cross Domain*	*Robustness*
					☐
Shift type			
Covariate	*Label*	*Full*	*Assumed*
☐			
Shift source			
Naturally occuring	*Partitioned natural*	*Generated shift*	*Fully generated*
☐			
Shift locus			
Train–test	*Finetune train–test*	*Pretrain–train*	*Pretrain–test*
	☐		

A.2 Effectiveness of Harmonic Distance Scores in Reducing Augmentation Ratio While Maintaining Performance

As shown in Figure 6, for all of our augmentation sets, an higher or similar score can be obtained by using a smaller ratio when compared to using the whole augmentation dataset. While there may be no clear pattern in what the best ratio might be for each augmentation, it is evident that for the large majority of the augmentations, using less than a 1.0 ratio leads to a better if not equal score— while reducing the computational resources needed.

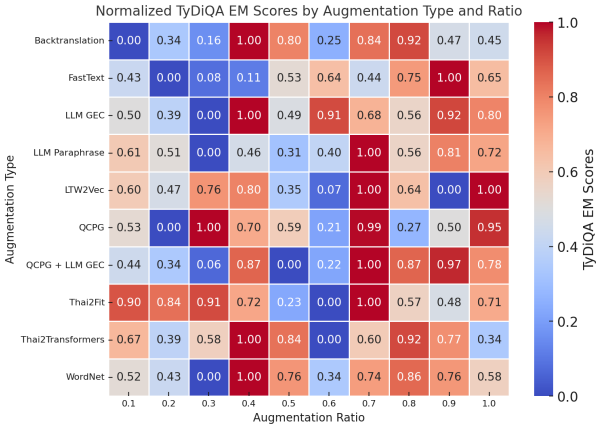

Figure 6: Heatmap illustrating the normalized TyDiQA Exact Match (EM) scores across various data augmentation techniques and ratios. The color scale represents the normalized EM score, with blue indicating lower performance and red indicating higher performance. Each cell displays the average normalized EM score for a specific combination of augmentation type and ratio.