



Language Understanding Laboratory, Myanmar

Summer Internship (2025) Report

mySQuAD: Myanmar SQuAD

Submitted by

SQuAD Team

Under the supervision of

Dr. Ye Kyaw Thu

National Electronics and Computer Technology Center (NECTEC)
LU. Laboratory

With mentorship from

Hay Man Htun

Department of Electrical Engineering

Kasetsart University
LU. Laboratory

Shwe Sin Moe

Faculty of Computer Engineering and
Information Technology

Yangon Technological University
LU. Laboratory

Team Members

Chit Ko Ko Lwin
Kasem Bundit University

Kaung Myat Htet
Assumption University

Khaing Zin Theint
Rangsit University

Thadoe Hein
Assumption University

Abstract

This report presents the outcomes of a summer internship at the Language Understanding Laboratory (LULab), focusing on the development of *mySQuAD*, a Myanmar adaptation of the Stanford Question Answering Dataset (SQuAD v1.1). The project involved translating large-scale English QA data into Burmese using both the NLLB (No Language Left Behind) model and Gemini Flash 2.5. Preprocessing techniques, such as context deduplication and sentence-level splitting, were applied to improve translation efficiency and accuracy. In addition, we fine-tuned an mT5 model on the translated dataset to evaluate its performance on Burmese QA tasks. The model demonstrated strengths in producing fluent, contextually relevant answers while revealing limitations such as hallucination, inconsistent transliteration, and language mixing. Overall, the work contributes a foundational dataset and initial benchmark models for low-resource QA research in Burmese, laying the groundwork for future improvements in translation quality, dataset expansion, and QA system deployment.

Contents

1	Objective	1
2	Introduction	2
3	Literature Review	3
3.1	Spanish: Carrino et al. (2019)	3
3.2	Korean: Lee et al. (2018)	3
3.3	German: Kaiser & Falenska (2024)	3
3.4	Marathi: Ghatage et al. (2024)	4
3.5	Synthesis and Insights	4
4	Translating SQuAD v1.1 into Myanmar	5
4.1	Methodology	5
4.2	Dataset Overview	5
4.3	NLLB Translation Approach	6
4.4	Gemini Flash 2.5 Translation Approach	7
4.5	Translation Progress Summary	8
4.6	Challenges and Limitations	8
4.6.1	NLLB Translation Issues	8
4.6.2	Gemini Translation Challenges	9
4.7	Conclusion	10
5	Burmese mT5 Question-Answering Model	11
5.1	Introduction	11
5.2	Training Setup	11
5.3	Analysis of Model Outputs	12
5.4	Representative Sample Outputs	13
5.5	Observations and Discussion	17
6	Future Work	19
6.1	Expanding and Improving the Dataset	19

6.2	Model Development and Fine-Tuning	19
6.3	Evaluation and Benchmarking	20
6.4	Applications and Deployment	20
7	Conclusion	21
	Acknowledgements	22
	References	23

Chapter 1

Objective

The primary objective of this project is to develop *mySQuAD*, a large-scale Question Answering (QA) dataset for the Myanmar language, adapted from the Stanford Question Answering Dataset (SQuAD v1.1). In doing so, the project seeks to:

- **Enable research in low-resource languages:** Provide the first publicly available QA dataset for Burmese, creating opportunities for natural language understanding research in an underrepresented language.
- **Preserve structural alignment:** Ensure that translated answers align correctly with their corresponding contexts, maintaining compatibility with the original SQuAD v1.1 format.
- **Establish a benchmark:** Create a reliable resource for training and evaluating multilingual and Burmese-specific models such as mT5.
- **Promote practical applications:** Encourage the integration of Burmese QA systems into real-world tools such as digital assistants, educational platforms, and information retrieval systems.

In summary, the objective is not only to adapt SQuAD v1.1 into Burmese but also to build a foundation for future research, benchmarking, and applications in low-resource language technologies.

Chapter 2

Introduction

Question Answering (QA) is a crucial task in natural language processing (NLP), driving applications such as intelligent assistants, educational platforms, and information retrieval systems. While significant progress has been made for English, many low-resource languages, including Burmese, lack large-scale QA datasets. This scarcity hampers the development of multilingual systems and limits access to language technologies for underrepresented communities.

To address this gap, our internship project focused on creating *mySQuAD*, a Myanmar-language adaptation of SQuAD v1.1. The objectives were threefold: (1) to enable research on Burmese QA through a publicly available dataset, (2) to facilitate cross-lingual transfer by aligning Burmese data with the SQuAD format, and (3) to benchmark multilingual models such as mBERT, XLM-R, and mT5 on Burmese tasks. The work combined automatic translation using NLLB and Gemini Flash 2.5 with preprocessing strategies that reduced redundancy and ensured alignment between answers and contexts.

Beyond dataset construction, we fine-tuned the mT5 model to assess its ability to generate accurate answers in Burmese. This evaluation provided both quantitative measures (e.g., training and evaluation loss) and qualitative insights (e.g., paraphrasing ability, factual accuracy, and handling of named entities). Together, these efforts not only produced the first large-scale Burmese QA resource but also established a foundation for future work in low-resource NLP.

Chapter 3

Literature Review

This chapter reviews recent efforts to construct multilingual Question Answering (QA) datasets by leveraging machine translation of the English SQuAD corpus, covering Spanish, Korean, German, Slovak, and Marathi.

3.1 Spanish: Carrino et al. (2019)

Carrino et al. introduce the *Translate-Align-Retrieve (TAR)* method to automatically translate SQuAD v1.1 into Spanish, yielding SQuAD-es v1.1. By fine-tuning a Multilingual-BERT model on this data, they achieve state-of-the-art performance: 68.1 F1 on Spanish MLQA and 77.6 F1, 61.8 Exact Match on Spanish XQuAD, outperforming previous baselines. [1]

3.2 Korean: Lee et al. (2018)

Lee et al. tackle QA dataset creation for Korean by combining a small manually annotated seed set (approximately 4K QA pairs) with a large-scale machine-translated SQuAD (MT). Pure MT yields just 52.49 F1 due to translation errors and span misalignment; however, integrating the seed data with a translation certainty filtering mechanism improves performance to 71.50 F1. [2]

3.3 German: Kaiser & Falenska (2024)

Kaiser and Falenska compare different answer-span retrieval strategies for translating SQuAD into German, including RAW (exact match), QUOTE (using markers), and TAR (alignment-based). When only translated data is available, TAR

yields the best results; in hybrid setups with small manually annotated German data, the optimal method depends on the evaluation scenario. [3]

3.4 Marathi: Ghatage et al. (2024)

Ghatage et al. present MahaSQuAD, the first full-scale SQuAD dataset in Marathi, with 118,516 training, 11,873 validation, and 11,803 test samples, plus a manually verified gold test set of 500 examples. They also propose a generic span alignment method for low-resource languages when translating SQuAD. [4]

3.5 Synthesis and Insights

A comparison across languages reveals consistent themes:

- **Need for span alignment:** Translation often distorts answer spans, making methods like TAR or approximate search essential to preserve QA integrity.
- **Hybrid approach efficacy:** Small-scale manual annotations significantly mitigate translation noise when combined with MT-derived data (e.g., Korean and German).

Thus, future multilingual QA dataset efforts should combine robust translation, effective span retrieval, and minimal high-quality human annotation to maximize both efficiency and model performance.

Chapter 4

Translating SQuAD v1.1 into Myanmar

4.1 Methodology

In the data preparation phase of our project, we aimed to create a Myanmar language question-answering dataset by translating the Stanford Question Answering Dataset (SQuAD) v1.1. The primary objective was to translate the source English text into Myanmar using two different approaches: NLLB (No Language Left Behind) neural machine translation model and Gemini Flash 2.5 large language model.

4.2 Dataset Overview

The original SQuAD v1.1 dataset consisted of:

- **Training set:** 87,599 rows
- **Validation set:** 10,570 rows
- **Total:** 98,169 rows

A critical optimization step in our data preparation involved identifying and removing redundant contexts. Since the same context paragraph typically contained two to five questions, translating each context multiple times would be computationally inefficient. By extracting unique contexts, we significantly reduced the translation workload:

- **Training unique contexts:** 18,892 (reduced from 87,599)

- **Validation unique contexts:** 2,067 (reduced from 10,570)

This deduplication process resulted in approximately 78% reduction in translation overhead while maintaining all necessary information.

4.3 NLLB Translation Approach

Our initial translation effort utilized the NLLB model to process the entire SQuAD dataset, covering both training and validation splits. However, we encountered a significant challenge with long context paragraphs being truncated during translation. The NLLB model appeared to have inherent length constraints that resulted in incomplete translations when processing particularly lengthy passages. Below is an example of context not being translated and cut off.

English

After the death of the president Benito Juárez in 1872, the first magistracy of the country was occupied by the vice-president Sebastián Lerdo de Tejada, who called for new elections. Two candidates were registered; Lerdo de Tejada and General Porfirio Díaz, one of the heroes of the Battle of Puebla which had taken place on May 5, 1862. Lerdo de Tejada won the election, but lost popularity after he announced his intent to run for re-election. On March 21, 1876, Don Porfirio Díaz rebelled against President Sebastian Lerdo de Tejada. The Plan of Tuxtepec defended the "No Re-election" principle. On June 2, 1876 the garrisons in the state of Chihuahua surrendered to the authority of General Porfirio Díaz; Governor Antonio Ochoa was arrested until all the Lerdista forces were suppressed throughout the state. Porfirio Díaz then helped Tiras regain the governorship of the state of Chihuahua allowing for the Plan of Tuxtepec to be implemented. The victory of the Plan of Tuxtepec, gave the interim presidency to Jose Maria Iglesias and later, as the only candidate, the General Porfirio Díaz assumed the presidency on May 5, 1877. During the first years of the Porfiriato (Porfirio Díaz Era), the Díaz administration had to combat several attacks from the Lerdista forces and the Apache. A new rebellion led by the Lerdista party was orchestrated from exile in the United States. The Lerdista forces were able to temporarily occupy the city of El Paso del Norte until mid-1877. During 1877 the northern parts of the state suffered through a spell of extreme drought which were responsible for many deaths in El Paso del Norte.

Burmese (Translated)

၁၈၇၂ ခုနှစ်၊ မေလ ၅ ရက်နေ့က ဖြစ်ပွားခဲ့သော Puebla တိုက်ပွဲ၏ သူရဲကောင်းများထဲမှ တစ်ဦးဖြစ်သူ Lerdo de Tejada နှင့် General Porfirio Díaz တို့ကို မှတ်ပုံတင်ခဲ့သည်။ Lerdo

de Tejada သည်ရွေးကောက်ပွဲတွင် နိုင်ခဲ့သော်လည်း ပြန်လည်ရွေးချယ်ရန် ရည်ရွယ်ကြောင်း ကြေညာပြီးနောက် လူကြိုက်များမှုကို ဆုံးရှုံးခဲ့သည်။ ၁၈၇၆ ခုနှစ်၊ မတ်လ ၂၁ ရက်တွင် Donfirio Díaz သည် သမ္မတ Sebastián Lerdo de Tejada ကိုပုန်ကန်ခဲ့သည်။ Tuxtepec အစီအစဉ်သည် "Nopec ပြန်လည်ရွေးချယ်ခြင်း" မူဝါဒကို ကာကွယ်ခဲ့သည်။ ၁၈၇၆ ခုနှစ်၊ ဇွန်လ ၂ ရက်တွင် Chihuahua ပြည်နယ်ရှိ စစ်သည်များသည် Porfirio Porfirio Porfirio El Paso ၏ အာဏာကို အပ်နှံခဲ့သည်။ Porfirio Díaz ၏ အုပ်ချုပ်ရေးမှူး Antonio Díaz သည် ပြည်နှင့်ဒဏ်ကို ခံယူခဲ့ပြီးနောက်ပိုင်းတွင် Porfirio Porfirio Porfirio သည် မေလ ၅ ရက်နေ့အထိ မြောက်ပိုင်းပြည်နယ်တွင် တိုက်ခိုက်နိုင်ခဲ့သည်။

Solution to Length Limitation: To overcome the truncation issue, we implemented a sentence-level splitting strategy. Instead of processing entire context paragraphs as single units, we segmented them into individual sentences before translation. This approach ensured complete translation coverage while maintaining semantic coherence when the sentences were recombined.

4.4 Gemini Flash 2.5 Translation Approach

In parallel to the NLLB effort, we employed the Gemini Flash 2.5 model for translation. Due to API limitations of 250 requests per day, we were constrained to translating only a subset of the training data. We processed 32,617 rows from the training set, which encompassed diverse domains including business, sports, entertainment, science, technology, politics, and world news. Below is one sample showing Gemini model can translate the full paragraph and can also handle name entities though it failed to do so in some cases.

English

A major battleground, Burma was devastated during World War II. By March 1942, within months after they entered the war, Japanese troops had advanced on Rangoon and the British administration had collapsed. A Burmese Executive Administration headed by **Ba Maw** was established by the Japanese in August 1942. Wingate's British Chindits were formed into long-range penetration groups trained to operate deep behind Japanese lines. A similar American unit, Merrill's Marauders, followed the Chindits into the Burmese jungle in 1943. Beginning in late 1944, allied troops launched a series of offensives that led to the end of Japanese rule in July 1945. The battles were intense with much of Burma laid waste by the fighting. Overall, the Japanese lost some 150,000 men in Burma. Only 1,700 prisoners were taken.

Burmese (Translated)

ဒုတိယကမ္ဘာစစ်အတွင်း မြန်မာနိုင်ငံသည် အဓိကစစ်မြေပြင်တစ်ခုဖြစ်ခဲ့ပြီး ပြင်းထန်စွာ

ဖျက်ဆီးခံခဲ့ရသည်။ ဂျပန်တပ်များသည် စစ်ထဲဝင်ရောက်ပြီး မကြာမီ လအနည်းငယ်အတွင်းဖြစ်သော ၁၉၄၂ ခုနှစ်၊ မတ်လတွင် ရန်ကုန်သို့ ချီတက်လာပြီး ဗြိတိသျှအုပ်ချုပ်ရေးလည်း ပြိုလဲသွားခဲ့သည်။ ၁၉၄၂ ခုနှစ်၊ ဩဂုတ်လတွင် ဂျပန်တို့က **ဒေါက်တာဘမော်** ဦးဆောင်သော ဗမာ စီမံအုပ်ချုပ်ရေးအဖွဲ့ကို ဖွဲ့စည်းပေးခဲ့သည်။ ဝင်းဂိတ်၏ ဗြိတိသျှ ချင်းတွင်းတပ်ဖွဲ့များကို ဂျပန်တပ်များ၏ နောက်ပိုင်းနယ်မြေများတွင် နက်နက်ရှိုင်းရှိုင်း ထိုးဖောက်ဝင်ရောက် လှုပ်ရှားနိုင်ရန် လေ့ကျင့်ပေးထားသည့် ကာလရှည် စစ်ဆင်ရေးအဖွဲ့များအဖြစ် ဖွဲ့စည်းခဲ့သည်။ အလားတူ အမေရိကန်တပ်ဖွဲ့တစ်ခုဖြစ်သော မာရီလ်၏ မာရေဒါးတပ်ဖွဲ့သည် ၁၉၄၃ ခုနှစ်တွင် ချင်းတွင်းတပ်ဖွဲ့များနောက်သို့ လိုက်ပါလျက် မြန်မာ့တောနက်ထဲသို့ ဝင်ရောက်ခဲ့သည်။ ၁၉၄၄ ခုနှစ် နှောင်းပိုင်းမှစ၍ မဟာမိတ်တပ်ဖွဲ့များသည် ထိုးစစ်ကြီးများ စတင်ဆင်နွှဲခဲ့ ကြရာ ၁၉၄၅ ခုနှစ်၊ ဇူလိုင်လတွင် ဂျပန်တို့၏ အုပ်ချုပ်မှုကို အဆုံးသတ်စေခဲ့သည်။ တိုက်ပွဲများသည် ပြင်းထန်ခဲ့ပြီး တိုက်ပွဲများကြောင့် မြန်မာနိုင်ငံ၏ ဒေသအများအပြားမှာ ပျက်စီးဆုံးရှုံးခဲ့ရသည်။ အလုံးစုံအားဖြင့် ဂျပန်တို့သည် မြန်မာနိုင်ငံတွင် လူပေါင်း ၁၅၀,၀၀၀ ခန့် ဆုံးရှုံးခဲ့ရသည်။ အကျဉ်းသား ၁,၇၀၀ ဦးသာ ဖမ်းမိခဲ့သည်။

4.5 Translation Progress Summary

Dataset	Train	Validation	Total
SQuAD v1.1	87,599	10,570	98,169
NLLB Translation	87,599	10,570	98,169
Gemini Translation	32,617	—	32,617

Table 4.1: Sizes of SQuAD v1.1 and its Myanmar translations (NLLB and Gemini)

4.6 Challenges and Limitations

4.6.1 NLLB Translation Issues

Despite the successful implementation of sentence-level splitting, several challenges remained with the NLLB translation:

1. **Length Constraints:** Even with sentence splitting, extremely long sentences occasionally suffered from incomplete translation, though this issue was significantly reduced compared to paragraph-level processing.
2. **Named Entity Transliteration:** The model failed to properly transliterate named entities into Myanmar script. For example:

English: “The assassination was led by Gaius Cassius and Marcus Brutus.”

Translation: “လူသတ်မှုကို Gaius Cassius နဲ့ Marcus Brutus တို့က ဦးဆောင်ခဲ့ကြတယ်။”

While the names were preserved, they were not transliterated into Myanmar script as would be expected in native Myanmar text.

3. **Inconsistent Script Usage:** Occasional instances of inappropriate script mixing occurred, such as Malayalam characters appearing in the Myanmar translation (e.g., “လျှပ်စစ်အင်ဂျင်ကီယာရှစ်ခု”).
4. **Context-Independent Answer Translation:** When translating answer fields in isolation without their surrounding context, the model produced awkward translations. For instance:

Answer: “Old College” → “အဟောင်း ကောလိပ်” (literal translation)

Context: “Old College, the oldest building on campus and located near the shore of St. Mary lake, houses undergraduate seminarians.”

Translation: “Old College သည် ကျောင်းဝင်းအတွင်းရှိအရင်ဆုံးအဆောက်အအုံဖြစ်ပြီး St. Mary ကန်ကမ်းနားအနီးတွင် တည်ရှိပြီး ဘွဲ့ရသင်္ချာကျောင်းသားများ နေထိုင်သည်။”

In context, the proper noun “Old College” was appropriately preserved, but when translated independently, it was incorrectly rendered as a literal translation.

4.6.2 Gemini Translation Challenges

The Gemini translation approach revealed distinct patterns of data loss:

1. **Missing Answer Fields:** A significant number of answer fields were missing in the translated output, particularly for specific entity types including numerical values (e.g., *one*, *two*, *three*), years (e.g., *1946*, *1989*, *2001*), country names (e.g., *France*, *Russia*, *Germany*, *Japan*), religious terms (e.g., *Christianity*), and other proper nouns that pose inherent translation challenges.

4.7 Conclusion

Both translation approaches successfully generated substantial portions of the Myanmar SQuAD dataset, each with distinct strengths and limitations. The NLLB approach provided complete dataset coverage with consistent translation quality, while the Gemini approach offered potentially higher translation quality for the processed subset, albeit with notable gaps in specific entity types. The sentence-level splitting strategy proved effective in addressing the length limitation issues encountered with NLLB, demonstrating the importance of preprocessing strategies in neural machine translation tasks.

Chapter 5

Burmese mT5 Question-Answering Model

5.1 Introduction

This study investigates the performance of a fine-tuned mT5 model for question-answering tasks in Burmese. Unlike straightforward classification tasks, QA requires the model to process a context passage, understand semantic and factual relationships within it, and generate accurate answers. The goal of this research was to assess how effectively a pre-trained multilingual T5 model could adapt to a low-resource, monolingual Burmese dataset and to evaluate its ability to generate fluent, contextually correct answers for various types of questions. This report presents both quantitative metrics from model training and qualitative observations of representative outputs to provide a comprehensive understanding of the model's strengths and limitations.

5.2 Training Setup

The google/mt5-base model was fine-tuned on a Burmese QA subset consisting of approximately 7,000 training examples and 1,700 evaluation examples. Training spanned five epochs with a batch size of 16, gradient accumulation of two, and bf16 precision. A data collator was used to pad sequences, and `predict_with_generate=True` enabled the model to generate answers directly rather than just predict token IDs. No intermediate evaluation was performed during training.

The model's training loss decreased steadily over the course of training. For example, at step 500 the loss was 2.7798, which dropped to 2.3206 by step 1000. The final training loss after 1,095 steps was approximately 2.524. This steady de-

cline indicates that the model was successfully learning from the dataset, although the loss suggests there remains room for improvement, reflecting the inherent difficulty of the task and the limited size of the dataset. Evaluation on the held-out test set yielded an evaluation loss of 1.7456, but no automatic exact-match (EM) or F1 scores were reported. Quantitative assessment thus requires manual comparison of model outputs against reference answers.

5.3 Analysis of Model Outputs

To assess the model’s behavior, fifteen representative test questions were manually evaluated, providing insight into both its capabilities and weaknesses. In many instances, the model produced outputs that exactly matched the reference answers. For example, in a question asking about the field in which Professor Mark Long earned his degree, the model correctly generated “စီးပွားရေး,” matching the reference answer exactly. Similarly, when asked about Wesley Clark’s initial weaknesses in his presidential campaign, the model correctly responded with “အသေးစိတ် မူဝါဒ အဆိုပြုချက် အနည်းငယ်,” demonstrating its ability to capture temporal and sequential reasoning from context. These examples suggest that the model is capable of memorizing and accurately reproducing factual information when the context closely aligns with patterns observed during training.

In other cases, the model produced answers that were semantically correct but phrased differently from the reference. For instance, when asked for the term “airport,” the model generated a more descriptive Burmese phrase translating roughly to “airstrip for aircraft.” Similarly, for a question about opposition to national and regional laws, the model reworded the phrase without changing the underlying meaning. These instances show that the model can generalize from the context, sometimes producing longer or paraphrased answers while retaining semantic correctness.

Despite these successes, the model occasionally produced hallucinatory or incorrect outputs. In one example, a question about “Albert Frederick Arthur George” had the reference answer as the Burmese name of St. Mary Magdalene Church, yet the model repeated the English name from the question itself. Another question about Ptolemaic Egypt resulted in a plausible-sounding but factually incorrect answer, blending English and Burmese phrases. Additionally, when asked about the film *Awakenings*, the model incorrectly produced a description of a 1973 television documentary instead of the expected person’s name. These examples illustrate that the model sometimes “guesses” additional facts when uncertain, leading to confident but incorrect outputs.

Language mixing also emerged as a notable issue. The model occasionally used English terms for names or technical concepts even when the question was

entirely in Burmese. For example, certain questions produced outputs such as “Albert Frederick Arthur George” or “Ptolemaic အုပ်ချုပ်ရေး,” which can reduce the model’s reliability in a monolingual Burmese QA setting. Minor lexical variations, such as the substitution of synonyms or small alterations in entity names, were observed, suggesting that while the model preserves meaning, it does not always produce the exact reference phrasing.

Overall, the model generated fluent and grammatically correct Burmese sentences. Most outputs were concise, typically one sentence or phrase, and there was no evidence of repetition or nonsensical constructions. Even when answers were hallucinated or partially incorrect, they remained syntactically coherent, indicating that the decoder is capable of producing readable and natural Burmese text.

5.4 Representative Sample Outputs

The following representative examples illustrate the model’s behavior across correct, paraphrased, and incorrect outputs:

Sample 1

- Question : ပါမောက္ခ Mark Long က ဘယ်ဘာသာရပ်မှာ ဘွဲ့ရခဲ့လဲ။
- Context : ၂၀၀၇ ခုနှစ်က ဝါရှင်တန်တက္ကသိုလ်မှ စီးပွားရေးပါမောက္ခ Mark Long ကပြုလုပ်သော လေ့လာမှုတစ်ခုတွင် အပြုသဘောဆောင်သော လုပ်ဆောင်မှု၏ အခြားရွေးချယ်မှုများသည် ကယ်လီဖိုးနီးယား၊ တီဟာစံစ်နှင့် ဝါရှင်သန်ရှိ အများပိုင်အမှတ်တံဆိပ်တက္ကသိုလ်များနှင့်လူနည်းစုဝင်ခွင့်ကို ပြန်လည်ထူထောင်ရာတွင် ထိရောက်မှုမရှိကြောင်း ပြသခဲ့သည်။ ပိုတိုကျာစတရောရရင် လူနည်းစုဝင်ခွင့် တိုးလာတာဟာ အဲဒီပြည်နယ်တွေရဲ့ အထက်တန်းကျောင်းတွေမှာ လူနည်းစုဝင်ခွင့်တိုးလာတာကြောင့် ရှင်းပြနိုင်ပြီး အတန်းအလိုက် (လူမျိုးမဟုတ်ဘဲ) အပြုသဘောဆောင်တဲ့ လုပ်ဆောင်မှုတွေရဲ့ အကျိုးခံစားသူတွေက လူဖြူကျောင်းသားတွေ ဖြစ်လိမ့်မယ်။ တစ်ချိန်တည်းမှာ အပြုသဘောဆောင်တဲ့ လုပ်ဆောင်မှုကိုယ်တိုင်ဟာ ကိုယ်ကျင့်တရားနဲ့ ရုပ်ပိုင်းအရ စရိတ်ကြီးမားပါတယ်။ လူဖြူလူထုရဲ့ ၅၂ ရာခိုင်နှုန်း (လူမည်း ၁၄ ရာခိုင်နှုန်းနဲ့ ယှဉ်ရင်) က ဒါကို ဖျက်သိမ်းသင့်တယ်လို့ ထင်တယ်။ လူဖြူတွေဟာ လူမျိုးရေး လက္ခဏာကို သုံးတာကို မုန်းတယ်လို့ ဆိုလိုပြီး အပြည့်အဝ စစ်ဆေးမှုက တက္ကသိုလ်တွေကို တစ်နှစ်ကို ဒေါ်လာ သန်း ၁.၅ ကနေ ဒေါ်လာ ၂ သန်းအထိ ကုန်ကျစေမယ်လို့ မျှော်လင့်ပါတယ်။ တရားစွဲဆိုမှုအတွက် ဖြစ်နိုင်ခြေရှိတဲ့ ကုန်ကျစရိတ်ကို မကွင်းပဲပေါ့။

- Original Answer : စီးပွားရေး
- Generated Answer: စီးပွားရေး

Sample 2

- Question : Wesley Clark အတွက် အစကတည်းက ဘယ်လို အားနည်းချက် ရှိခဲ့လဲ။
- Context : ၂၀၀၃ စက်တင်ဘာလတွင် အငြိမ်းစား လေးကြယ်မှီလ်ချုပ်ကြီး ဝက်စ်လီကလပ်က ဒီမိုကရက်တစ်ပါတီ၏ ကိုယ်စားလှယ်လောင်းအဖြစ် သမ္မတရွေးကောက်ပွဲတွင် ဝင်ရောက်ယှဉ်ပြိုင်ရန် ရည်ရွယ်ကြောင်း ကြေညာခဲ့သည်။ သူ့မဲဆွယ်မှုက ခေါင်းဆောင်မှုနဲ့ အမျိုးသားရေးဝါဒအကြောင်းအရာတွေကို အာရုံစိုက်ခဲ့တယ်။ အစောပိုင်း မဲဆန္ဒနယ် ကြော်ငြာတွေဟာ ဇီဝဇာတ်လမ်းကို အားကိုးခဲ့တယ်။ နောက်ကျပြီး စလုပ်ခဲ့လို့ အသေးစိတ် မူဝါဒ အဆိုပြုချက် အနည်းငယ်သာ ရှိခဲ့ပါတယ်။ ဒီအားနည်းချက်ဟာ သူ့ရဲ့ ပထမဆုံး ဆွေးနွေးပွဲတွေမှာ ထင်ရှားခဲ့ပေမဲ့ မကြာခင်မှာ ကြီးမားတဲ့ အခွန်အလျှော့အစီး အစီအစဉ်အပါအဝင် သဘောထားစာရွက်စာတမ်းတစ်သိကြီး တင်ပြခဲ့တယ်။ သို့သော်လည်း ဒီမိုကရက်များက သူ့မဲဆွယ်မှုကို ထောက်ခံရန် မလာခဲ့ကြပါ။
- Original Answer : အသေးစိတ် မူဝါဒ အဆိုပြုချက် အနည်းငယ်
- Generated Answer: အသေးစိတ် မူဝါဒ အဆိုပြုချက် အနည်းငယ်

Sample 3

- Question : ခရစ်ယာန်များ၏ ညီညွတ်မှု အတွင်းရေးမှူးချုပ် ဘယ်သူလဲ?
- Context : ဗာတီကန် ၂ နှစ်အတွင်းတွင် သင်းအုပ်ချုပ်ရေးအဖေများက အခြားဘာသာဝင် ခရစ်ယာန်များကို ဒေါသဖြစ်စေနိုင်သော ကြေညာချက်များကို ရှောင်ရှားခဲ့သည်။ ခရစ်ယာန်များ၏ ညီညွတ်မှု အတွင်းရေးမှူးချုပ်ဖြစ်သူ ကာဒီနယ် Augustin Bea သည် ပါမောက္ခ ပေါလု (၆) ၏ ကြိုးပမ်းမှုများကို အမြဲတမ်း အပြည့်အဝ ထောက်ခံခဲ့သည်။ Bea သည် ဂျူးဘာသာနှင့် အခြားဘာသာဝင်များနှင့် ခရစ်ယာန်သင်း၏ ဆက်ဆံရေးကို စည်းမျဉ်းစည်းကမ်းပေးသော Nostra aetate ၏ အမိန့်ကို အတည်ပြုရန်လည်း အားကောင်းစွာ ပါဝင်ခဲ့သည်။ [d]
- Original Answer : ကာဒီနယ် Augustin Bea

- Generated Answer: အာဒီနယ် Augustin Bea

Sample 4

- Question : လေယာဉ်ပျံများ ပျံသန်းရန်နှင့် ဆင်းသက်ရန်အတွက် အဆောက်အအုံများ ပါရှိသည့် လေဆိပ်ဆိုသည်မှာ ဘာနည်း။
- Context : လေဆိပ်ဆိုတာ လေယာဉ်ပျံသန်းမှုအတွက် လေယာဉ်ကွင်းတစ်ခုဖြစ်ပြီး လေယာဉ်တွေ ပျံသန်းဖို့နဲ့ ဆင်းသက်ဖို့ အဆောက်အအုံတွေ ရှိပါတယ်။ လေဆိပ်များတွင် လေယာဉ်များကို သိုလှောင်ထိန်းသိမ်းရန်နှင့် ထိန်းချုပ်ရေး မျှော်စင်ရှိသည်။ လေဆိပ်တစ်ခုမှာ လေယာဉ်ပျံသန်းဖို့ ပြေးလမ်း (သို့) ရဟတ်ယာဉ်ပျံပျံဖို့ လေယာဉ်ကွင်းလို အနည်းဆုံး လုပ်ဆောင်မှုဆိုင်ရာ တက်ကြွတဲ့ မျက်နှာပြင်တစ်ခုပါဝင်တဲ့ လေကြောင်းမှ ဝင်ရောက်နိုင်တဲ့ ပွင့်လင်းတဲ့ နေရာတစ်ခုပါဝင်ပြီး မကြာခဏ ထိန်းချုပ်ရေး မျှော်စင်တွေ၊ ဟန်ဂါများနဲ့ ကုန်တင်ယာဉ် terminal တို့လို ဘေးချင်းဆိုင်တဲ့ အသုံးအဆောင် အဆောက်အအုံတွေ ပါဝင်ပါတယ်။ အကြီးမားဆုံး လေဆိပ်များတွင် အခြေစိုက် လေကြောင်းလိုင်းများ၊ လေဆိပ်များ၊ လေကြောင်း ယာဉ်ထိန်းချုပ်ရေးစင်တာများ၊ စားသောက်ဆိုင်များနှင့် ဧည့်ခန်းများကဲ့သို့ ခရီးသည်တင်ဆောင်ရေး အဆောက်အအုံများနှင့် အရေးပေါ် ဝန်ဆောင်မှုများ ရှိနိုင်သည်။
- Original Answer : လေဆိပ်
- Generated Answer: လေယာဉ်ပျံသန်းမှုအတွက် လေယာဉ်ကွင်း

Sample 5

- Question : MiG ကို ဘာအတွက် ဒီဇိုင်းထုတ်ခဲ့တာလဲ။
- Context : USAF က MiG-15 ကို တုံ့ပြန်ဖို့ F-86 Sabre လို့ခေါ်တဲ့ ၎င်းရဲ့အစွမ်းအထက်ဆုံး တိုက်လေယာဉ် သုံးစင်းကို ပို့ပေးခဲ့တယ်။ [စာမျက်နှာ ၂၇ ပါ ရုပ်ပုံ] MiG ကို ဗုံးခွဲတိုက်ခိုက်ရေး လေယာဉ်အဖြစ် ဒီဇိုင်းထုတ်ထားတာပါ။ ၎င်းသည် အမြင့်ဆုံး ၅၀၀၀၀ ပေ (၁၅၀၀၀ မီတာ) ရှိပြီး အလွန်အမင်း လေးလံသော လက်နက်များပါဝင်သည်။ ၃၇ မီလီမီတာသေနတ်တစ်လက်နှင့် ၂၃ မီလီမီတာ သေနတ်နှစ်လက်။ P-80 Shooting Stars နဲ့ F9F Panthers တို့ရဲ့ တိုက်လေယာဉ် လိုက်ပါမှုကို ကျော်လွှားနိုင်လောက်အောင် မြန်ပြီး အမေရိကန်ရဲ့ လေးလံဗုံးကြဲရေး လေယာဉ်တွေကို ထိတွေ့ဖျက်ဆီးနိုင်ခဲ့တယ်။ B-29 ပျက်စီးမှုများကို ရှောင်လွှဲရန်

မဖြစ်နိုင်ခဲ့သဖြင့် လေတပ်သည် နေအချိန် ဗုံးကြဲမှုမှ ညအချိန် ဗုံးကြဲမှုများသို့ ပြောင်းရန် အတင်းအဓမ္မ လိုအပ်ခဲ့သည်။ MiG တွေကို F-86 Sabres တွေက တိုက်ခိုက်ခဲ့တယ်။ ၎င်းတို့၏ အမြင့်ဆုံးအကွာအဝေးသည် ပေ ၁၃၀၀၀ ရှိပြီး ရဒါပစ်မှတ်များဖြင့် အလျားသတ်မှတ်ထားသော ၅၀ လက်မ မောင်းစက်သေနတ် ၆ လုံးဖြင့် လက်နက်ကိုင်ထားသည်။ ပိုမိုတုံ့ အမြင့်မှာ ဝင်လာရင် တိုက်ခိုက်တာ (သို့) မတိုက်ခိုက်တာရဲ့ အကျိုးကျေးဇူးက MiG ကို သွားခဲ့တယ်။ လေယာဉ်တစ်စင်းလုံးမှာ အမြန်ဆုံး အမြန်နှုန်း ၁၁၀၀ ကီလိုမီတာအထိ ရောက်ရှိလာခဲ့ပါတယ်။ MiG က ပိုမြန်အောင် တက်ပေမဲ့ Sabre က ပိုကောင်းအောင် လှည့်ပြီး ရေငုပ်ခဲ့တယ်။

- Original Answer : ဗုံးကြဲတိုက်ခိုက်ရေး လေယာဉ်
- Generated Answer: ဗုံးခွဲတိုက်ခိုက်ရေး လေယာဉ်

Sample 6

- Question : မျိုးသုဉ်းမှု အန္တရာယ်ရှိ မျိုးစိတ်များ ပြန်လည်ထူထောင်ရေး အစီအစဉ်ကို ပြီးစီးဖို့ ပျမ်းမျှ အချိန်က ဘယ်လောက်လဲ။
- Context : ငါးနှင့် တောရိုင်းတိရစ္ဆာန်များဆိုင်ရာ ဝန်ဆောင်မှု (FWS) နှင့် အမျိုးသားပင်လယ်ငါးလုပ်ငန်း ဝန်ဆောင်ရေး (NMFS) တို့သည် မျိုးသုဉ်းမှုအန္တရာယ်ရှိသည့် မျိုးစိတ်များ ပြန်လည်ထူထောင်ရေး ရည်မှန်းချက်များ၊ လိုအပ်သော တာဝန်များ၊ ဖြစ်နိုင်ခြေရှိကုန်ကျစရိတ်များနှင့် ခန့်မှန်းထားသော အချိန်ဇယားကို ဖော်ပြသည့် မျိုးသုဉ်းမှုအန္တရာယ်ရှိသည့်မျိုးစိတ်များအတွက် ပြန်လည် ထူထောင်ရေး အစီအစဉ်ကို ဖန်တီးရန် လိုအပ်သည် (အဓိကအားဖြင့် မျိုးသုသေသနွယ်စုများ တိုးပွားစေရန်နှင့် ၎င်းတို့၏ စီမံခန့်ခွဲမှုကို တိုးတက်စေရန်အတွက်)။ မျိုးသုသုဉ်းခြင်းအန္တရာယ်ရှိသော စာရင်းမှ ဖယ်ရှားနိုင်သည့် အထိ)။ ESA သည် ပြန်လည်ထူထောင်ရေး အစီအစဉ်ကို မည်သည့်အချိန်တွင် ပြီးစီးရမည်ကို မသတ်မှတ်ပါ။ FWS မှာ မျိုးစိတ်တွေကို စာရင်းသွင်းပြီး သုံးနှစ်အတွင်း ပြီးမြောက်ဖို့ သတ်မှတ်ထားတဲ့ မူဝါဒတစ်ခုရှိပေမဲ့ ပျမ်းမျှအချိန်က ခြောက်နှစ်လောက်ပါ။ ပြန်လည်ထူထောင်ရေး အစီအစဉ် ပြီးမြောက်မှု နှစ်စဉ်နှုန်းသည် Ford အစိုးရမှ ၄) မှ Carter (၉) ၊ Reagan (၃၀) ၊ Bush I (၄၄) နှင့် Clinton (၇၂) အထိ တချိန်လုံး တိုးတက်လာသော်လည်း Bush II (၁၆) နှစ်တွင် ကျဆင်းခဲ့သည်။
- Original Answer : ၆ နှစ်ခန့်
- Generated Answer: ခြောက်နှစ်

Sample 7

- Question : ဘယ်ဥပဒေတွေကို ပြည်ထောင်စု ဥပဒေက လွှမ်းမိုးလဲ။
- Context : ဖက်ဒရယ်ဥပဒေနဲ့ စာချုပ်တွေဟာ အခြေခံဥပဒေနှင့် ကိုက်ညီသရွေ့ အမေရိကန်ပြည်ထောင်စု ပြည်နယ် ၅၀ နဲ့ နယ်မြေတွေမှာ ဆန့်ကျင်ဘက် နိုင်ငံတော်နဲ့ နယ်ခြားဥပဒေတွေကို ကြိုတင်ကာကွယ်ပါတယ်။ သို့သော်လည်း ဖက်ဒရယ်အာဏာ၏ ကန့်သတ်ချက်သည် ကန့်လန့်နေသည်မှာ ဖက်ဒရယ်အာရုံ၏ ကန့်လန့်ချက်သည် တစ်ကမ္ဘာလုံးတွင် မတည်ရှိသောကြောင့် ဖြစ်သည်။ အမေရိကန် ဖက်ဒရယ်စနစ်၏ နှစ်ဖက်စိုးမိုးမှုစနစ်တွင် (အင်းဒီယပိုင်မြေများရှိသောကြောင့် တကယ်တမ်းတွင် သုံးဖက်စိုးရ) ပြည်နယ်များသည် ပင်လုံစိုးအုပ်သူများဖြစ်ပြီး ၎င်းတို့၏ ကိုယ်ပိုင် အခြေခံဥပဒေရှိပြီး ဖက် ဒရယ်စိုးစံသည် အခြေခံဥပဒေပုဒ်မတွင် ဖော်ပြထားသော ကန့်သတ်ထားသော အချုပ်အခြာအာဏာကိုသာ ပိုင်ဆိုင်သည်။ တကယ်တမ်းတွင် ပြည်နယ်များသည် ပြည်ထောင်စု အခြေခံဥပဒေထက် ပိုကျယ်ပြန့်သော အခွင့်အရေးများကို ပြည်သူများကို ပေးနိုင်သည်ဆိုပါက ၎င်းတို့သည် ပြည်ထောင်စုချုပ်၏ အခြေခံဥပဒေပုဒ်မများအား ချိုးဖောက်ခြင်းမရှိပါ။ ထို့ကြောင့် အမေရိကန်ဥပဒေအများစု (အထူးသဖြင့် နိုင်ငံသားအများစု နေ့စဉ်တွေ့ကြုံခံစားရသော စာချုပ်၊ ပြစ်မှုကျူးလွန်မှု၊ ပိုင်ဆိုင်မှု၊ ရာဇဝတ်မှုနှင့် မိသားစုဥပဒေ၏ ”အသက်ရှင်သောဥပဒေ” များ) သည် အဓိကအားဖြင့် ပြည်နယ်ဥပဒေများဖြင့် ဖွဲ့စည်းထားပြီး ပြည်နယ်တစ်ခုမှတစ်ခုသို့ အလွန်ကွဲပြားနိုင်သည်။
- Original Answer : နိုင်ငံတော်နှင့် နယ်မြေဆိုင်ရာ ဥပဒေများ ဆန့်ကျင်နေခြင်း
- Generated Answer: ဆန့်ကျင်ဘက် နိုင်ငံတော်နဲ့ နယ်ခြားဥပဒေ

5.5 Observations and Discussion

The model exhibits a combination of strengths and weaknesses. It generates fluent Burmese sentences and can provide exact answers for clearly stated factoid questions. It can also paraphrase answers and generalize context for semantically correct responses. However, it sometimes hallucinates facts, produces incorrect entities, mixes languages, or slightly alters lexical choices, especially for names and technical terms. These tendencies indicate that the model relies on memorized patterns and may guess when insufficient contextual evidence exists.

Future improvements could include expanding the dataset, implementing grounding mechanisms to reduce hallucinations, enforcing Burmese-only outputs, and

computing EM/F1 metrics on a held-out set to obtain robust quantitative evaluation. A larger, more diverse training corpus could also help the model better generalize to less frequent entities and rare contexts.

Chapter 6

Future Work

6.1 Expanding and Improving the Dataset

While the mySQuAD dataset provides a solid foundation for Burmese question-answering research, several avenues exist for further enhancement:

- **Dataset Expansion:** Increase the size of the training and evaluation sets by translating additional English QA datasets or collecting new Burmese contexts and questions. This would help reduce data sparsity and improve model generalization.
- **Quality Enhancement:** Address translation errors and missing information observed in the current dataset. Incorporating manual post-editing or semi-automated quality checks could significantly improve the accuracy and reliability of the dataset.
- **Diverse Domains:** Include more specialized domains, such as healthcare, law, and literature, to ensure broader coverage of topics and improve the model's versatility.

6.2 Model Development and Fine-Tuning

Future efforts can focus on enhancing model performance and addressing current limitations:

- **Larger and Multilingual Models:** Experiment with larger pretrained models or additional multilingual architectures (e.g., XLM-R, mBART) to leverage cross-lingual transfer more effectively.

- **Domain-Specific Fine-Tuning:** Fine-tune models on specific domains or question types to improve accuracy for specialized applications.
- **Handling Hallucinations:** Implement techniques such as constrained decoding, answer verification modules, or retrieval-augmented generation to reduce instances of incorrect or hallucinated answers.

6.3 Evaluation and Benchmarking

- **Automatic Metrics:** Establish automated evaluation protocols, including exact-match (EM), F1, and BLEU scores, to allow consistent and reproducible assessment of model performance.
- **Human Evaluation:** Conduct comprehensive human assessments of model outputs to measure fluency, correctness, and cultural appropriateness, especially for low-resource languages like Burmese.
- **Cross-Lingual Comparisons:** Compare Burmese QA models with counterparts in other low-resource languages to benchmark progress and identify best practices.

6.4 Applications and Deployment

- **Practical QA Systems:** Integrate trained models into real-world applications such as digital assistants, educational tools, or information retrieval systems for Burmese speakers.
- **Interactive Feedback Loops:** Enable user feedback mechanisms to continuously improve model outputs and adapt to evolving language use.

Chapter 7

Conclusion

The summer internship at the Language Understanding Laboratory provided a comprehensive opportunity to explore natural language processing and machine learning techniques, with a particular focus on low-resource languages such as Burmese. The development of mySQuAD, a Myanmar adaptation of the SQuAD v1.1 dataset, successfully created a benchmark resource for training and evaluating question-answering models. Through the translation and pre-processing of the data using NLLB and Gemini 2.5 Flash models, the team addressed challenges such as missing sentences, entity translation errors, and redundant contexts, resulting in a cleaner, structured data set suitable for QA research.

Fine-tuning the mT5 model on the Burmese QA subset demonstrated the model's capability to generate fluent, contextually correct answers while also highlighting areas for improvement, including hallucinated outputs and language mixing. The analysis of representative outputs confirmed the model's strengths in memorizing factual information and generalizing across contexts, while also revealing the limitations inherent to low-resource data and multilingual transfer.

Overall, the internship achieved its objectives by providing practical experience in data collection, translation, model training, and evaluation. The resulting dataset and model analysis lay a solid foundation for future work, including expanding the dataset, improving translation quality, exploring larger multilingual architectures, and developing practical Burmese QA applications. These efforts collectively contribute to advancing natural language understanding technologies for low-resource languages and promoting the practical use of QA systems in Myanmar.

Acknowledgment

We extend our sincere gratitude to Dr. Ye Kyaw Thu for his invaluable guidance throughout our summer internship at the LU Lab. We also deeply appreciate the insightful support of our senior mentors, Hay Man Htun and Shwe Sin Moe, whose encouragement and advice greatly enriched our learning experience.

References

- [1] Casimiro Pio Carrino, Marta R. Costa-jussà, José A. R. Fonollosa. Automatic Spanish Translation of the SQuAD Dataset for Multilingual Question Answering. arXiv:1912.05200, 2019.
- [2] Kyungjae Lee, Kyoungcho Yoon, Sunghyun Park, Seung-won Hwang. Semi-supervised Training Data Generation for Multilingual Question Answering. ACL Anthology (LREC 2018), 2018.
- [3] Julia Kaiser, Anna Falenska. How to Translate SQuAD to German? A Comparative Study of Answer Span Retrieval Methods for Question Answering Dataset Creation. Proceedings of KONVENS 2024, ACL Anthology, 2024.
- [4] Rohit Ghatage, Abhijit Kulkarni, Rutuja Patil, Saurabh Endait, Raviraj Joshi. MahaSQuAD: Bridging Linguistic Divides in Marathi Question-Answering. arXiv:2404.13364, 2024.