ENGLISH MIZO LANGUAGE PAIRS : AUTOMATIC MACHINE TRANSLATION

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

VANLALMUANSANGI KHENGLAWT

MZU REGN NO : 1906313

Ph. D REGN NO : MZU/Ph.D./1585 OF 31.07.2019

DEPARTMENT OF COMPUTER ENGINEERING

SCHOOL OF ENGINEERING AND TECHNOLOGY

JULY, 2024

ENGLISH MIZO LANGUAGE PAIRS : AUTOMATIC MACHINE TRANSLATION

BY

VANLALMUANSANGI KHENGLAWT
DEPARTMENT OF COMPUTER ENGINEERING

Supervisor : Prof. (Dr.) Ajoy Kumar Khan
Joint-Supervisor : Dr. Partha Pakray

Submitted

In partial fulfillment of the requirement of the Degree of Doctor of Philosophy in Computer Engineering of Mizoram University, Aizawl.

CERTIFICATE

This is to certify that the thesis entitled **"English Mizo Language Pairs : Automatic Machine Translation"** submitted to the Mizoram University for the award of the degree of **Doctor of Philosophy** in **Computer Engineering** by **Vanlalmuansangi Khenglawt**, Ph.D. Registration No. **MZU/Ph.D./1585 of 31.07.2019** is Ph.D. scholar in the Department of Computer Engineering, under our guidance and supervision, and has not been previously submitted for the award of any degree in any Indian or foreign university. She has fulfilled all criteria prescribed by the UGC (Minimum Standard and Procedure governing Ph.D. Regulations). She has fulfilled the mandatory publication (Publication enclosed) and completed Ph.D coursework. It is also certified that the scholar has been admitted to the department through an entrance test, followed by an interview as per UGC Regulation of 2018.

Prof. (Dr.) Ajoy Kumar Khan
Supervisor & Professor
Dept. of Computer Engineering
Mizoram University,
Aizawl-796 004

Place:
Date:

Dr. Partha Pakray
Joint Supervisor & Associate Professor
Dept. of Computer Science & Engineering
National Institute of Technology Silchar,
Assam-788 010

Place:
Date:

DECLARATION

Mizoram University

December, 2024

I, Vanlalmuansangi Khenglawt, hereby declare that the subject matter of this thesis is the record of work done by me, that the contents of this thesis did not form basis of the award of any previous degree to me or to do the best of my knowledge to anybody else, and that the thesis has not been submitted by me for any research degree in any other University/Institute.

This is being submitted to the Mizoram University for the degree of Doctor of Philosophy in Computer Engineering.

(VANLALMUANSANGI KHENGLAWT)

(Prof. (Dr.) AJOY KUMAR KHAN)

Supervisor

(Dr. V.D.AMBETH KUMAR)

Head

(Dr. PARTHA PAKRAY)

Joint Supervisor

ACKNOWLEDGEMENTS

"I will praise You, for I am fearfully and wonderfully made;...My frame was not hidden from you, When I was made in secret and skillfully wrought in the lowest parts of the earth...When I awake, I am still with You." Psalm 139

First, I would like to thank my research supervisor, Dr. Ajoy Kumar Khan, for allowing me to work under him. I am grateful for his guidance, encouragement, and support throughout my doctoral research work. His welcoming nature and constructive suggestions have guided me in meeting the requirements for my research.

I would also like to thank my co-supervisor, Dr. Partha Pakray. He is the individual who pioneered the integration of the Mizo language into the realm of Machine Translation. I am grateful to have him as my mentor and thrilled to work under his guidance. This thesis comes from his helpful discussions, supervision, and meticulous attention to detail.

I take this opportunity to express my appreciation to my family for the support they have given me. My three precious children are my motivation. Having them by my side gives me the strength to face challenges; they are my inspirations. I would also like to extend a heartfelt thanks to my parents for the love, care, and support they have given me throughout my life. My mother's prayers have shaped me into the person I am today.

Mrs. Vanlalmuansangi Khenglawt
Reg. No.: MZU/Ph.D./1585 OF 31.07.2019
Dept. of Computer Engineering,
Mizoram University.

Contents

ACKNOWLEDGEMENTS **i**

LIST OF TABLES **vii**

LIST OF FIGURES **viii**

ABBREVIATIONS **ix**

1 INTRODUCTION **1**

1.1 Natural Language Processing 1

1.2 Scope of Thesis 3

1.3 Machine Translation 4

1.3.1 Rule-based Machine Translation (RBMT) 4

1.3.2 Corpus-based Machine Translation 6

1.4 Challenges of NLP-Based Machine Translation 8

1.5 Motivation 10

1.6 Objectives of Thesis 11

1.7 Contributions 12

1.8 Organization of Thesis 14

2 Literature Survey **16**

2.1 History of Machine Translation 16

2.1.1 Rule-Based Machine Translation 16

2.1.2 Corpus Based Machine Translation 17

2.2 Related Works on Machine Translation 21

2.2.1 Low Resource Machine Translation 22

2.2.2 Machine Translation on Indian Languages 26

2.2.3 Prior Works on English↔Mizo Machine Translation 27

3 The Study of the Structure and Challenges of Mizo language 29

3.1 The Mizo Language 29

3.2 Low Resource Pair: English↔Mizo 30

3.3 Challenges of English↔Mizo Machine Translation 33

3.3.1 Tonal Words 33

3.3.2 Tonal Polysemy 34

3.3.3 Vocabulary Distinction 37

3.3.4 Symbolic Words 37

3.4 Development of Parallel Corpus: English↔Mizo 38

3.4.1 Corpus Details 38

3.4.2 Corpus Extraction Approaches 39

3.4.3 Data Cleaning and Split 41

3.4.4 Domain Coverage 42

3.5 Conclusion 44

4 Machine Translation for English↔Mizo Pair Encountering Tonal Words 45

4.1 Introduction 45

4.2 EnMzCorp1.0: English↔Mizo Corpus 45

4.3 Baseline System 47

4.3.1 Satistical Machine Translation 47

4.3.2 Neural Machine Translation 49

4.3.3 Experimental Setup 51

4.3.4 Results 51

4.4 System Description for Encountering tonalily 56

4.5 Analysis 58

4.6 Conclusion 64

5 Building Low Resource English-to-Mizo NMT Model with Post Processing 65

5.1 Introduction 65
5.2 Low Resource NMT 66
5.2.1 Traditional Approach 66
5.2.2 Knowledge-Transfer Approach 67
5.3 Approaches of Low Resources NMT 67
5.3.1 Back Translation (BT) 68
5.3.2 Word Embedding 69
5.3.3 Transfer Learning 69
5.3.4 Zero shot NMT 70
5.3.5 Multilingual Approach 71
5.3.6 Multimodal Approach 72
5.4 Corpus Preparation 73
5.5 System Description 74
5.6 Experiment and Result Analysis 77
5.7 Conclusion 81

6 Mizo Visual Genome 1.0: A Dataset for English↔Mizo Multimodal NMT 83

6.1 Introduction 83
6.2 Multimodal NMT 84
6.2.1 Feature extraction from images 84
6.2.2 Initialising Encoder 84
6.2.3 Initialising Decoder 85
6.3 Mizo Visual Genome 1.0 (MVG 1.0) 86
6.4 System Description 89
6.5 Experiment and Result Analysis 91
6.6 Conclusion 95

7 English↔Mizo NMT Using Language Model and Addressing Data Scarcity Problem 98

7.1 Introduction 98

7.2 Transformer Based NMT 98

7.3 Development of English↔Mizo NMT 100

7.4 Experimental Result and Analysis 102

7.5 Error Analysis and Discussion 106

7.6 Conclusion 113

8 CONCLUSION 114

REFERENCES 118

LIST OF PAPERS BASED ON THESIS 133

List of Tables

3.1 Alphabet of Mizo Language 30
3.2 Word-order of Mizo 32
3.3 Variation of tone (a) in Mizo 34
3.4 Examples of the tonal words in Mizo 35
3.5 Example of tonal-polysemy words in Mizo 36
3.6 Vocabulary distinction 37
3.7 Examples of symbolic words in Mizo 38
3.8 Sources and statistics of English↔Mizo Corpus 39
3.9 Examples of parallel and monolingual sentences 40

4.1 Corpus sources and statistics 46
4.2 Extracted (Mz) data using Google OCR 46
4.3 Statistics for train, valid and test set 47
4.4 BLEU scores of baseline systems 52
4.5 TER (%) scores of baseline systems 53
4.6 METEOR scores of baseline systems 55
4.7 F-measure scores of baseline systems 55
4.8 HE (Overall Rating (%)) scores of baseline systems 56
4.9 Augmented train data statistics 57
4.10 Automatic evaluation results of proposed approach 58
4.11 Human evaluation results of proposed approach 58

5.1 Example of a predicted sentence (tone markers are marked as bold) 66
5.2 Mizo Tonal Words in Corpus 73
5.3 Corpus Statistics 73
5.4 Monolingual Data on Mizo Language 74
5.5 Statistics for train, valid and test set 74

5.6 Synthetic parallel data statistics 76

5.7 BLEU scores of preliminary experiments 78

5.8 Comparative results (BLEU scores) of different models for En-to-Mz translation, M1: Transformer, M2: BERT-fused Transformer, SPC: Synthetic Parallel Corpus, PP: Post-processing, UPC: Unidirectional Parallel Corpus, BPC: Bidirectional Parallel Corpus 79

5.9 Human evaluation scores of different models for En-to-Mz translation 79

5.10 Analysis of different models for En-to-Mz translation 82

6.1 Multimodal Corpus Statistics of MVG 1.0 87

6.2 Google Translate output with correction 88

6.3 Example of Parallel Train, Valid and Test Data 89

6.4 Automatic Evaluation Scores of En-Mz Multimodal and Text-only NMT Systems 91

7.1 Statistics for phrase-pairs 102

7.2 BLEU scores of En-to-Mz and Mz-to-En translation. SPA: Synthetic Parallel-Sentence Augmentation, PPA: Phrase Pairs Augmentation, LM: Language Model 104

7.3 TER scores of En-to-Mz and Mz-to-En translation, SPA: Synthetic Parallel-Sentence Augmentation, PPA: Phrase Pairs Augmentation, LM: Language Model 104

7.5 F-measure scores of En-to-Mz and Mz-to-En translation, SPA: Synthetic Parallel-Sentence Augmentation, PPA: Phrase Pairs Augmentation, LM: Language Model 104

7.4 METEOR scores of En-to-Mz and Mz-to-En translation, SPA: Synthetic Parallel-Sentence Augmentation, PPA: Phrase Pairs Augmentation, LM: Language Model 105

7.6 Human evaluation scores of En-to-Mz and Mz-to-En translation, SPA: Synthetic Parallel-Sentence Augmentation, PPA: Phrase Pairs Augmentation, LM: Language Model 105

7.7 BLEU scores of En-to-Mz and Mz-to-En translation on FLORES-200 test data 105

List of Figures

3.1 Example of English–Mizo word order 32

3.2 Data acquisition 41

4.1 Abstract diagram of phrase-based SMT 48

4.2 English–Mizo NMT system (attention-based RNN) 51

4.3 Proposed Approach 57

5.1 English-to-Mizo NMT System 75

5.2 Comparative analysis on tonal frequency of words. Reference: Mizo target sentences (test data) 80

6.1 An example of MVG 1.0 dataset: a picture with its caption in English and Mizo. 87

6.2 English↔Mizo MNMT System 90

6.3 Examples of MNMT's Best Predicted Output with Text-only NMT 92

6.4 Examples of MNMT's Worst Predicted Output with Text-only NMT 93

6.5 Example 1 of MNMT's Average Predicted Output with Text-only NMT 96

6.6 Example 2 of MNMT's Average Predicted Output with Text-only NMT 97

7.1 Transformer Architecture Model 99

7.2 Encoder and Decoder of Transformer Model 99

7.3 English↔Mizo NMT System 101

7.4 Comparison among the previous work, present work (best model: with SPA+PPA+LM), and Google Translate in terms of BLEU and HE scores for En-to-Mz 106

7.5 Comparison among the present work (best model: with SPA+PPA+LM) and Google Translate in terms of BLEU and HE scores for Mz-to-En 107

ABBREVIATIONS

NLP	Natural Language Processing
MT	Machine Translation
AI	Artificial Intelligence
POS	Part of Speech
NER	Named Entity Recognition
RBMT	Rule-Based Machine Translation
EBMT	Example-Based Machine Translation
SMT	Statistical Machine Translation
PBSMT	Phrase-Based Statistical Machine Translation
NMT	Neural Machine Translation
RNN	Recurrent Neural Network
BRNN	Bidirectional Recurrent Neural Network
CNN	Convolutional Neural Network
TNN	Transformer Neural Network
BT	Back Translation
IBT	Iterative Back Translation
SVO	Subject-Verb-Object
OSV	Object-Subject-Verb
SOV	Subject-Object-Verb
MMT	Multimodal Machine Translation
MNMT	Multimodal Neural Machine Translation
ALPAC	Automatic Language Processing Advisory Committee
PTB	Penn Treebank
TM	Translation Model
LM	Language Model

TLM	Translation Language Modelling
MLM	Masked Language Modelling
LSTM	Long Short Term Memory
GRU	Gated Recurrent Unit
OOV	Out-of-Vocabulary
BPE	Byte Pair Encoding
MVG	Mizo Visual Genome
HE	Human Evaluation
BLEU	BiLingual Evaluation Understudy
TER	Translation Error Rate
METEOR	Metric for Evaluation of Translation with Explicit ORdering
RIBES	Rank Based Intuitive Bilingual Evaluation Scores

Chapter 1

INTRODUCTION

Natural Language Processing (NLP) plays a vital role in developing machine translation (MT) systems, comprising the capacity to process and interpret human languages in a computational context. It involves various techniques and models to understand, interpret, and generate human languages. Using NLP in MT allows for more nuanced and contextually appropriate translations, improving fluency, adequacy, and quality. It provides numerous benefits for MT, significantly enhancing translation systems' overall performance and effectiveness.

By leveraging advanced models, contextual understanding, and continuous learning, NLP has transformed MT into a powerful tool for global communication, bridging linguistic and cultural gaps more effectively than ever before. Despite ongoing challenges, the advancements in NLP continue to push the boundaries of what MT can achieve.

1.1 Natural Language Processing

NLP is a branch of artificial intelligence (AI) that enables machines to read, understand, and derive meaning from human language. Its area combines linguistics and computer science to analyze linguistic rules and structures and create models that can understand, dissect, and distinguish significant characteristics from text and speech. The significance of NLP lies in its ability to eliminate the barriers between computing systems and human language. By enabling computer systems to process and comprehend natural language, NLP enables numerous applications and opportunities in various fields. It encompasses developing and applying algorithms, models, and strategies to process, examine, and generate data in natural language.

The field of NLP covers a broad range of activities and applications, including:

- **Part of Speech Tagging:** Labeling words in a text with their appropriate parts of speech is known as part-of-speech (POS) tagging. Nouns, verbs, adjectives, and other grammatical categories may be included. POS tagging is helpful for several NLP applications, including information extraction, named entity identification, and MT. Additionally, it may be used to determine a sentence's grammatical structure and clarify words with multiple meanings. Based on the context, the POS tagging algorithm determines the appropriate POS tag for a particular word. It plays a vital role in various NLP applications by providing essential syntactic information about words in a text corpus [1, 2].

- **Sentiment Analysis:** Sentiment analysis mainly categorizes textual material according to its positive, negative, or neutral polarity. It is a decisive approach that helps determine people's opinions [3]. Customer satisfaction and product quality are the two most crucial factors in business performance. Therefore, businesses must consider consumers' needs to ensure competitiveness [4]. Using sentiment analysis tools, decision-makers may monitor changes in public or consumer perception of certain entities, activities, goods, technology, and services [5].

- **Named Entity Recognition (NER):** NER aims to identify and classify named entities in text into predefined categories such as person names, locations, organizations, dates, and numerical expressions. It involves analyzing unstructured text to extract entities of interest, providing valuable insights for various applications like information retrieval, question answering, and sentiment analysis. By accurately identifying named entities, NER enables automated information extraction, entity linking, and knowledge discovery from large text corpora, enhancing the efficiency of text-processing tasks.

- **Speech Recognition:** Speech Recognition is a technology in which a machine can recognize human spoken words and phrases, which may then be utilized to produce text. Accuracy and speed are the two metrics used to measure the performance of the machines that may use speech recognition technology. Speech is a crucial component of human communication. Speech recognition is widely applicable in computer science, medicine, and other sciences. Acoustic and linguistic modeling are two methods commonly used by speech recognition systems [6].

- **Text Classification:** Text classification is classifying unstructured text documents into predetermined categories or classes based on content. Information filtering, mail categorization, search engines, query intent prediction, topic tracking, text corpus development, and other fields have extensively used text classification technologies. It can assist users in adequately categorizing garbled data to acquire organized text information and address the issue of users' requirements for information [7].

- **Machine Translation (MT):** MT involves using computer algorithms to automatically translate text or speech from one language to another. This technology leverages linguistic and computational techniques to bridge language barriers, facilitating global communication and access to information. The complexity of human languages, nuances, idiomatic expressions, and cultural contexts adds to the difficulty of developing robust MT systems.

1.2 Scope of Thesis

With AI and NLP technology developments, MT's scope is vast and constantly growing. While challenges remain, ongoing innovation in this field promises a more inclusive and efficient translation quality. MT between language pairings comprises handling syntax and semantics, precisely translating text across various linguistic structures, and addressing specific domain requirements. Advances in neural models have significantly improved translation quality, especially for high-resource languages with abundant training data. However, challenges persist for low-resource languages due to limited corpora. As AI and computational linguistics progress, MT has become essential for promoting understanding and communication worldwide.

The scope of NLP-based MT is extensive, covering theoretical research, practical applications, and future technological advancements. NLP has profoundly influenced MT, playing a crucial role in advancing translation technologies and influencing multiple industries. The key aspects are enhanced fluency and accuracy, context sensitivity, increased language support, multilingual MT, customizability, domain-specific translation, and reduced human intervention. A thesis in this field can focus on improving model architectures, addressing low-resource languages, enhancing real-time translation capabilities, and exploring the ethical implications of translation technologies. By contributing to any of these areas, research can significantly advance the field, making translation more accurate, accessible, and equitable.

1.3 Machine Translation

MT removes human intervention from translating one natural language to another using automatic translation, thereby resolving linguistically ambiguous problems. It includes a variety of techniques, ranging from rule-based approaches to advanced neural networks, to provide accurate, fluent, and contextually relevant translations. Despite the advancements, MT still faces challenges, including handling low-resource languages, maintaining context and coherence in longer texts, and producing accurate, fluent, and natural translations. It is divided into two broad categories: rule-based and corpus-based approaches.

1.3.1 Rule-based Machine Translation (RBMT)

The knowledge-driven approach is another name for a rule-based approach based on the linguistic information of the language. The system is built using a set of grammatical rules and linguistic experts. Due to the integration of syntactic, semantic, and morphological analysis in both the source and target languages, the RBMT system generates more predictable results for grammar. Developing RBMT is expensive because all the language rules must be implemented, and a vast amount of linguistic knowledge is required. However, once constructed, its syntax and semantics can be thoroughly analyzed.

RBMT systems may be divided into three categories: direct, transfer, and interlingua, based on the level of linguistic analysis.

- In the direct approach, replacement is done word-by-word or phrase-by-phrase [8, 9, 10]. It translates the source language directly into the target language. It is a unidirectional bilingual MT that requires substantial morphological analysis but relatively little syntactic and semantic analysis.

- In the transfer approach, the translation comprises three phases: analysis, transfer, and generation. In the analysis stage, the source language is syntactically and semantically analyzed to create an abstract representation of the source language. Then, in the transfer stage, it is transferred into an abstract representation of the target language using linguistic rules. Finally, a morphological analyzer is used at the generation stage to produce the target language [11].

- In the interlingua approach, the source language is translated into an intermediate language called the interlingua of the source language. The interlingua is then converted to the target language. This translation utilizes knowledge of the source language, grammar rules, and lexicons [12]. The interlingual representation is independent of language; therefore, all possible sentences with the same meaning in various languages are represented identically. Since it does not depend on any particular pair of languages, it may be used for multilingual MT.

Although the rule-based methods have reasonable translation accuracy, it requires a considerable amount of time and effort to pre-design a set of translation rules and the languages' grammatical structures.

1.3.2 Corpus-based Machine Translation

The corpus-based approach is also known as the data-driven approach. The corpus-based approach can self-learn using bilingual corpora that require a considerable volume of bilingual content in both the source and target languages. The corpus-based approach acquires translation information using these parallel data. There has been a significant change in the translation method from rule-based to corpus-based since relying on parallel sentences is more practical than complex grammatical rules with linguistic experts and knowledge of NLP techniques. Example-Based Machine Translation, Statistical Machine Translation, and Neural Machine Translation are the three methods of corpus-based MT.

- **Example-based machine translation (EBMT):** The EBMT requires a parallel corpus, and the central concept is text similarity. It identifies the approximately matching sentences (i.e., examples) using point-to-point mapping and similarity measures such as word, syntactic, or semantic similarity. The retrieval module and the adaptation module are the two modules that make up the translation method. The retrieval module finds identical parallel sentences from the corpus for a given input sentence. The adaptation module determines the parts of the translation to be reused from the retrieval module. The relevant match concerning the source language is used if it does not match.

- **Statistical Machine Translation (SMT):** In the corpus-based approach, the main drawback of EBMT is that, in real-time scenarios, various types of sentences cannot be covered by examples alone. To encounter this issue, SMT is introduced [13, 14]. This approach uses a statistical model, and the parameters are computed from bilingual corpus analysis. The translation problem is reformulated using a mathematical reasoning problem. In SMT, there are different forms of translation: word-based translation, phrase-based translation, syntax-based translation, and hierarchical phrase-based translation. Among them, phrase-based translation is the most widely used. Before

NMT, phrase-based SMT achieves a state-of-the-art approach.

- **Neural Machine Translation (NMT):** In the MT task, the NMT approach attains state-of-the-art for both high and low resource pair translations [15, 16, 17, 18, 19]. NMT can learn the model end-to-end by mapping the source and target sentence. The main problem with SMT is that it creates a model context by considering a limited-size set of phrases. As the phrase size increases, the data sparsity will reduce the quality. Likewise, feed-forward-based NMT calculates the phrase pair score by considering the length of the fixed phrases. However, in real-time translation, the phrase length of both source and target is not fixed. Therefore, recurrent neural networks (RNN) based NMT [20, 21, 22] is introduced to tackle variable-length phrases. RNN can process each word in a sentence of arbitrary length via continuous space representations. These representations can assist the long-distance relationship among words in a sentence. The drawback of RNN is that input processing follows a strict temporal order, which means it computes context in one direction based on preceding words, not future words. RNN is impotent to look ahead into future words. BRNN (Bidirectional RNN) [15] resolves this issue by utilizing two distinct RNNs, one for the forward direction and another for the backward direction. A BRNN-based model improves translation accuracy on low-resource pairs like English–Hindi, English–Tamil [23].

  Moreover, the Convolutional Neural Network (CNN) based NMT is introduced [24, 25] by taking advantage of parallelizing operation and considering relative positions of the tokens instead of the temporal dependency among the sequence tokens. However, it lags behind RNN features that enhance the source sentence's encoding. The demerits of CNN-based approaches require many layers to hold long-term dependency, making the network large or complex without ever succeeding, which is impractical. To handle such an issue, a transformer-based NMT comes in [26]. The idea behind the transformer model is to encode each position and apply a self-attention mechanism to connect two different words, which would be parallelized to accelerate learning. Unlike the traditional attention mechanism, the self-attention mechanism calculates attention several times, known as multi-head attention.

1.4 Challenges of NLP-Based Machine Translation

MT driven by NLP has seen significant progress, but various challenges remain. Below are some of the main challenges that MT encounters:

- **Contextual Understanding:** Although NLP has advanced in comprehending context, MT systems still face difficulties with ambiguous words, idiomatic language, and cultural nuances. Correctly understanding these elements is crucial for accurate translation. While modern neural models excel at word-by-word translation, they often struggle with capturing and preserving broader contextual information. This limitation can lead to mistranslations, especially in cases where context plays a crucial role in determining meaning. Additionally, handling idiomatic expressions, cultural nuances, and ambiguous language constructs further complicates contextual understanding. Overcoming these challenges requires innovative approaches integrating contextual cues, long-range dependencies, and world knowledge into MT models, aiming to produce accurate translations at the word level that are coherent and contextually appropriate.
- **Limitation of Resources:** Producing high-quality translations demands much training data. However, resource limitations pose significant challenges in MT. Low-resource languages lack sufficient parallel corpora for effective model training, leading to poor translation quality. Additionally, domain-specific translations and dialectal variations may be scarce, hindering the adaptation of MT systems to specific contexts. Overcoming these challenges requires innovative approaches such as transfer learning, data augmentation, and crowd-sourcing to leverage available resources more effectively. Collaborative efforts to collect and curate parallel data and invest in infrastructure

development are crucial for advancing MT for all languages and domains.

- **Rare Words:** Rare words present notable challenges in MT. Traditional MT models struggle to accurately translate infrequent or unseen words, as they rely on statistical patterns learned from training data. Low-frequency terms may lack sufficient context for accurate translation, leading to errors in output. Additionally, rare words may carry specialized meanings or domain-specific knowledge, exacerbating translation difficulties. Addressing these challenges requires novel techniques such as subword tokenization, rare word handling strategies, and incorporating external knowledge sources. Moreover, leveraging contextual information and domain-specific resources can enhance the translation of rare words in MT systems.

- **Requirement of Human Evaluation (HE):** Despite advancements in NLP-based MT, HE is essential in MT to ensure translation accuracy, fluency, and contextual appropriateness. Automated metrics like BLEU scores often fail to capture nuances such as idiomatic expressions and cultural references. Human evaluators can provide qualitative insights, identify subtle errors, and assess overall translation quality from a user perspective. Their feedback helps refine MT models by highlighting issues that algorithms might overlook. Moreover, HE is crucial for low-resource languages, where automated tools may lack sufficient data for accurate assessment. Therefore, incorporating human judgment is vital for developing reliable and user-centric MT systems. However, this additional step can slow down workflows and increase costs.

- **Preserving Cultural Sensitivity:** Cultural sensitivity in MT is crucial to avoid misinterpretations and offensive translations. MT systems frequently miss cultural nuances, resulting in mistranslations or culturally inappropriate results. Therefore, MT systems must understand and respect cultural nuances, idioms, and context-specific meanings. It involves training models on diverse, culturally rich datasets and incorporating rules that account for cultural differences. This issue is especially critical in settings where understanding cultural context is essential. Developing context-aware models considering the broader situational context also helps maintain cultural sensitivity. By prioritizing these strategies, MT can produce translations that are not only accurate but also culturally respectful and appropriate.

Addressing these challenges requires continuous research and development in NLP-based MT. Improvements in model architectures, training data, and real-time processing capabilities are crucial to overcoming these hurdles and enhancing MT systems' overall quality and reliability.

1.5 Motivation

MT is the bridge for communication barriers among people with different linguistic backgrounds. Beneath every language, there is a culture involved. A language is defined by the people living in the area, their origin, traditions, customs, cuisine, and many more. Therefore, a language means communication and defines the people using it. Around the world, a large number of languages have become extinct. It is endangered as a low-resource language, and a minority uses it. Mizo is a low-resource language with limited users, so it is at risk of extinction. Since MT removes barriers between languages, it can prevent the extinction of languages with limited resources. With the advent of technology, the Mizo language can have a chance to survive and can encounter technological advances in MT. There are very limited MT works on English↔Mizo pair [16, 27], that lag in encountering tonal words of Mizo. Additionally, automatic translations like Google and Bing cover 133[1] and 105[2] languages across the globe; however, the Mizo language has recently been included in Google Translate. Due to the lack of a standard corpus, it is still challenging to produce a correct translation.

MT is crucial in today's globalized world, breaking language barriers and facilitating seamless communication across various domains. In business, MT enables multinational companies to interact with clients and partners in different lan-

[1] https://en.wikipedia.org/wiki/Google$_$Translate

[2] https://www.bing.com/

guages, enhances customer support with multilingual chatbots, and localizes marketing materials and product information. The travel and tourism industry benefits from real-time translation of signs, menus, and conversations, making international travel more accessible. In education, MT provides students and researchers access to materials and research papers in multiple languages, promoting global knowledge exchange. The healthcare sector uses MT to bridge communication gaps between medical professionals and patients, ensuring an accurate understanding of medical information. Social media and online platforms utilize MT to translate user-generated content and news articles, fostering global conversations. Governments and international organizations rely on MT for diplomatic communication and translating official documents. Advancements in NMT and AI continually improve accuracy and contextual understanding, enhancing MT's effectiveness and reliability across these sectors.

1.6 Objectives of Thesis

The objectives of this thesis serve as the foundation for the research. It clearly defines the research goals, outlines the specific questions the study aims to address, establishes the scope of the research, and identifies the methodologies used. The objectives of this research work are as follows:

- Corpus Creation: Building of Parallel Corpus of English and Mizo Language (Train, Development, and Test Data) will be done based on different sources (Bible, Social Media, Movie Subtitles, Newspaper, Mizo Song Lyrics, etc.)

- Study of Statistical Machine Translation (SMT) will be carried out to automatically train translation models for English ↔ Mizo language pair.

- Study of Neural Machine Translation (NMT) will be carried out. There are two types of NMT: sequence-to-sequence-based models and transformer-based models.

- Evaluation of generated translation: by automatic using BLEU, METEOR, etc, and Human Evaluation (Adequacy, Fluency)
- Detailed study and comparison of different approaches shall be done using the parallel corpus.

An in-depth study of both languages is required to address the above objectives. English↔Mizo MT aims to develop an automated system that accurately and fluently translates text between English and Mizo languages. Key goals include ensuring linguistic and semantic correctness, preserving contextual meaning, and producing natural and idiomatic translations in both languages. Additionally, the system aims to handle the limited parallel data available for Mizo by leveraging techniques like data augmentation, phrase pair extraction, and a pre-trained language model. Ultimately, the goal is facilitating communication, information access, and cultural exchange between English and Mizo speakers, supporting education, government, and everyday communication applications.

1.7 Contributions

Contributions to MT can span a wide range of areas, from fundamental research in algorithms and models to practical solutions addressing the challenges of the English↔Mizo language pair. It can enhance global communication and accessibility, allowing real-time, efficient translation across languages. MT advancements support low-resource languages, promoting linguistic diversity and inclusivity. Automated translation reduces costs and speeds up processes in various sectors, including healthcare and legal services. Additionally, MT fosters cultural exchange by making literature and media accessible to diverse audiences. By making meaningful contributions in any of these areas, researchers can advance the field and make MT more accurate, efficient, and accessible to users. MT advancements bridge language barriers, fostering a more interconnected and inclusive global community.

The following significant advancements are made in the field of MT to address the objectives of this thesis:

- The primary requirement of the MT system is to establish a parallel corpus for the English↔Mizo language pair. A parallel corpus is collected from various online sources and also done manually. Monolingual data on the Mizo language has also been prepared.
- An in-depth study of the low-resource Mizo language has been performed. Several linguistic challenges and research gaps between English↔Mizo language pairs are investigated. Many strategies have been used to address specific issues. A large amount of the data is also manually gathered to deal with tonal words
- Various NMT and SMT models are used to train the created parallel corpus. Several methods have been investigated to improve the effectiveness and quality of translation using NMT. Multimodal NMT for English↔Mizo language pair has also been developed.
- The predicted sentences from different models are evaluated using automatic metrics such as BLEU, TER, METEOR, and F-measures. HE is also considered to measure the fluency and accuracy of the output.
- Various NMT models have demonstrated remarkable performance in achieving high translation quality, fluency, and contextual understanding across diverse language pairs and domains. Therefore, an automatic translation for the English↔Mizo NMT System is developed. The proposed system deals with linguistic challenges, and state-of-the-art results are achieved for English to Mizo and Mizo to English translation.

1.8 Organization of Thesis

The thesis is organized into eight chapters.

Chapter 1: Introduction - This chapter introduces NLP and its applications in various areas. The scope of the thesis is explained, along with the introduction of MT and the challenges of NLP-based MT. The motivation and the objectives for this thesis are also highlighted.

Chapter 2: Literature Survey - This chapter reviews the existing literature on MT, focusing on the history and evolution of MT through the years, i.e., rule-based and corpus-based approaches. Related works on various fields of MT are featured, including the prior works of English Mizo MT.

Chapter 3: The Study of the Structure and Challenges of Mizo Language - An in-depth study of the structure and challenges of the Mizo language is featured in this chapter. Several challenges for the English↔Mizo language pair are explained. Development of parallel corpus for English Mizo, including corpus details, the extraction process, domain coverage, pre-processing data, etc., are explained.

Chapter 4: Machine Translation for English↔Mizo Pair Encountering Tonal Words - As dealing with tonal language is one of the main challenges for the English↔Mizo language pair in MT, chapter four highlights the development of the corpus baseline system using data collection and model training. Various metrics are used to evaluate the quality of the translation. A data augmentation approach is proposed to encounter tonal words, resulting in better translation accuracy than the baseline system. Analysis of the resulting outputs was explained .

Chapter 5: Building Low Resource English-to-Mizo NMT Model with Post Processing - This chapter aims to build language resources for low-resource English-Mizo language pair. An NMT System is proposed to address several challenges for the two language pairs. The approaches of low-resource language are also explained. The system is also designed for the Mizo language using BERT fused NMT, and post-processing steps are included to improve translation accuracy. Various NMT models were investigated

Chapter 6: Mizo Visual Genome 1.0: A Dataset for English↔Mizo Multimodal NMT - A multimodal NMT approach for the Mizo language is introduced in this chapter. Since the standard multimodal corpus is unavailable, Mizo Visual Genome 1.0 was created for Multimodal NMT, where the dataset comprises images paired with corresponding bilingual textual descriptions. The English↔Mizo multimodal NMT system is the pioneering work for the language pair.

Chapter 7: English↔Mizo NMT Using Language Model and Addressing Data Scarcity Problem - This chapter summarizes the essential findings and challenges of the English↔Mizo language pair, such as the data scarcity problem and linguistic divergence between the language pair. These challenges are addressed using data augmentation, phrase pair extraction, and pre-trained language models. The research is concluded by building a language model for the English↔Mizo NMT system. The state-of-the-art results on the test data are achieved for English to Mizo and Mizo to English translation.

Chapter 2

Literature Survey

2.1 History of Machine Translation

Over several decades, MT has evolved through constant advances in developing technologies, algorithms, and the availability of large amounts of data. During World War II, there was a great demand for rapid document translation. Early in the 1950s, the concept of employing computers to automatically translate between human languages initially evolved. Weaver introduced the idea of MT to the computing industry in a memorandum sent in July 1949 [8]. The approach of MT is divided into two broad categories: rule-based and corpus-based.

2.1.1 Rule-Based Machine Translation

Rule-based machine translation (RBMT) systems were developed by researchers in the 1950s and 1960s. The 'direct translation' technique was the standard method used in systems at this time until the middle of the 1960s. These systems employed dictionaries and linguistic rules for text translation. However, the early translation systems produce translations by translating each word individually, using a bilingual dictionary to look them up, locating the translations in the target language, and then generating the output in the same order as the source language. The approach was undoubtedly inadequate, and a rearrangement of word sequences was needed, which led to the requirement of syntactic analysis [28].

In 1964, The Automatic Language Processing Advisory Committee (ALPAC) was established by the government funders of MT to assess the potential of MT. Its renowned 1966 study concluded that MT was slower, less precise, and twice as expensive as human translation and advised against further investment [29]. However, the research for MT was continued on a much smaller scale.

Linguistic theories have significantly impacted systems, leading to adopting the 'indirect' approach, which includes syntactic analysis. Although Weaver had proposed the potential of translating through an intermediate language in his memorandum [8], it was not until the 1960s that linguistics could provide any models to use. The 'interlingual' MT method entails a two-step procedure where the source language is converted into an interlingua, and the target language is then converted from the interlingua [30]. Analysis and synthesis programs operate independently using different dictionaries and grammar for the source and target languages.

Following its interlingua efforts in the middle of the 1970s, a transfer approach has been effectively employed. The source and target languages have unique deep structural representations in this approach. As a result, translation is a three-step process that includes text analysis into representation in the source language, transfer to representations in the target language, and text synthesis in the target language.

Until the end of the 1980s, the MT framework was based on various linguistic rules, such as rules for morphology, lexical transfer, syntactic creation, and syntactic analysis. The examples of RBMT include the transfer-based machine translation [31], inter-lingual-based machine translation [12], and dictionary-based machine translation [32].

2.1.2 Corpus Based Machine Translation

However, since 1989, alternative approaches and techniques referred to as 'corpus-based' methodologies have emerged, shattering the supremacy of the rule-based approach [33]. The corpus-based approach can self-learn using bilingual corpora that require a considerable volume of bilingual content in both the source and target languages. It acquires translation information using these parallel data.

- **Statistical Machine Translation (SMT) :**

  In the 1990s, the focus of MT research switched to statistical methods. These systems employed statistical models trained on large bilingual corpora rather than depending on several linguistic rules. SMT is corpus-based; it requires a source language corpus and a parallel target language corpus that humans properly translate [34].

  Word-based SMT - A significant step that led to SMT was the development of the word-to-word statistic model (a noisy channel model) known as the IBM Model 1 in 1990 [35]. This approach introduced the idea of lexical translation by aligning words in parallel corpora and calculating translation probabilities. Due to word ambiguity across languages, the MT system chooses the words from parallel corpora, while multiple-choice translations often appear. Consequently, the MT system must calculate the translation probability for maximum likelihood estimation. Nevertheless, during translation, words could be omitted or added.

  Consequently, a word may not be in one-to-one alignment. Researchers studied more advanced methods of statistical translation. In the following years, the model was improved to the IBM Model 5, which includes reordering features and additional alignment [35, 14]. Although word-based improvements at the time were remarkable and notable, they lacked robustness, which prevented translation systems from producing adequate precise outputs. For instance, the word-based translation approach makes dealing with synonyms, word lattices, and POS challenging.

  Syntax Based SMT - The syntax-based MT was then introduced [36, 37]. This method incorporated syntactic information to guide the translation process, allowing for improved sentence structure management. Compared to word-based MT (IBM Models), syntax-based MT models produce more accurate word alignments because they eliminate the limitation of word-based MT that only manages structurally similar language pairs. The basic principle of syntax-based MT is analyzing syntactic order and its correspondence. The syntax-based models are often derived directly from Penn Treebank (PTB) style parse trees by composing treebank grammar rules [38]. Using treebanks, models may learn to evaluate sentences and rearrange them according to predefined grammar rules [14].

Phrase Based SMT - In the early 2000s, Phrase-Based Statistical Machine Translation (PB-SMT) gained popularity [39]. The introduction of phrase-based MT substantially enhanced the performance of MT [13]. These systems break sentences into phrases (strings of words) and translate them using statistical patterns obtained from training data. Various methods enhanced translation quality, including incorporating linguistic features, word reordering modeling, and handling rare words.

- **Neural Machine Translation (NMT):**

  The development of deep learning algorithms revolutionized MT. In the middle of the 2010s, the artificial neural network-based NMT model was adopted as the new standard for MT. Researchers began researching the use of artificial neural networks. NMT primarily employs Recurrent Neural Networks (RNN) and Transformer Neural Networks (TNN).

  RNN-based Models - In RNN-based models, an NMT system requires an RNN encoder-decoder to perform the sequence-to-sequence transfer. The main advantage of NMT is that it can be learned rapidly and accurately by directly training the source and target texts in an encoder-decoder system [22]. The encoder-decoder model mutually computes translation probabilities for a given source-target language pair. Both an encoder and a decoder are used to process the sentences of each language. Then, the sentences are encoded by the encoder into a vector of fixed length, and the translation is produced as a result by the decoder [40]. Although the encoder-decoder design is efficient, it still has issues when processing long sentences. The quality of the translation degrades dramatically with increasing sentence length and the number of unfamiliar words [21]. To address the shortcomings, the attention-based RNN architecture for encoder-decoder was developed [15, 41].

Developing the sequence-to-sequence (Seq2Seq) model with attention mechanisms in 2014 marked a significant advancement in NMT. Instead of encoding the input sentences into a fixed-length vector, the attention-based model encodes them into a sequence of vectors, enabling NMT, a state-of-the-art MT approach. After the launch of Seq2Seq with attention, research on NMT advanced significantly. Various architectural changes and training methodologies were investigated to improve translation quality.

The Transformer model - The introduction of the Transformer model in 2017 again marked a significant development for NMT [26]. The network's architecture is based primarily on the attention mechanism. The model has succeeded in RNN using self-attention mechanisms, enabling parallel processing and improving the capturing of long-range dependencies [42]. Compared to conventional RNN-based models, transformers provided much faster training and enhanced translation quality. The incorporation of transfer learning and fine-tuning in NMT is another significant advancement. Pre-trained models, like the Transformer-based models trained on vast amounts of multilingual data have been made accessible, enabling effective transfer to particular language pairings or domains. Fine-tune pre-trained models on smaller task-specific datasets have achieved state-of-the-art with fewer computing resources.

NMT research is an active field of study, with continuous initiatives to address issues, including processing rare words, domain adaptation, and enhancing stability to noisy or ambiguous input. To increase the capabilities of NMT, methods including unsupervised or semi-supervised learning, incorporating reinforcement learning, and integrating additional knowledge sources are being explored. By producing more accurate, fluent, and human-like translations, NMT has transformed MT. As it can manage long-range dependencies and incorporate contextual information, NMT has become the dominant approach in MT, significantly enhancing translation quality. Traditional neural translation models include the sequence-to-sequence model [22], encoder-decoder model [15], attention mechanism [43], transformer [26, 44], and others. Despite NMT's dominance in MT, it does not scale well with small corpora and low-resource datasets. The back-translation [45] and pivot (bridging) language [46, 47] have been shown to improve the quality of translations for low-resource bilingual datasets significantly. Additionally, knowledge graph [48, 49, 50, 51, 52] plays a bigger part in enhancing MT.

2.2 Related Works on Machine Translation

The development of MT provides insights into its evolution, impact on communication, and technological advancements. Several research works have been adopted in the field of MT. The MT has evolved significantly over the past decades, transitioning from rule-based systems to sophisticated neural networks. Technological advancement has yielded more promising results, fluency, and better translation quality. The study of language for MT entails investigating the complex relationships among linguistics, structure, computational models, and translation techniques to develop systems that automatically translate text or speech from one language to another. MT between different language pairs can present various challenges, primarily due to the structural, lexical, and cultural differences between languages. These challenges can affect the accuracy and fluency of translations, requiring specialized techniques and adaptations to achieve quality results.

Research also explored domain adaptation and low-resource languages. Methods for adapting MT systems to specific domains were reviewed [53], while strategies for translating languages with limited resources were discussed [54]. Multilingual and zero-shot translation capabilities were examined, demonstrating that a single NMT model could handle multiple language pairs [55], and large-scale multilingual models supporting over 100 languages were developed [56]. Additionally, advanced techniques like back-translation have improved MT performance through data augmentation [57]. The practical application of these advancements was showcased with the Google Neural Machine Translation system, which applied NMT to large-scale translation tasks [58].

Evaluation metrics also saw significant advancements. It integrates intelligent quality detection models that leverage deep learning and allow for real-time translation quality assessment, addressing issues like over-translation and under-translation. The BLEU score provided a quantitative measure for translation quality based on n-gram overlaps between machine-generated and reference translations [59]. METEOR offered an alternative metric incorporating synonyms, stemming, and word order into its evaluation, addressing some limitations of BLEU [60].

Another critical area is the development of low-resource NMT. This area addresses the challenge of translating languages with limited data using different models to improve translation quality and accessibility. Systematic literature reviews have identified significant works and methodologies that enhance NMT for low-resource languages. The adoption of transfer learning in MT is another significant development. Pre-trained language models such as BERT [44], cross-lingual models [61], and GPT [62], have been fine-tuned for specific translation tasks, significantly improving performance, especially in low-resource language pairs. These models leverage vast amounts of monolingual data to build robust language representations that can be adapted to translation tasks with limited bilingual data.

2.2.1 Low Resource Machine Translation

Natural language is divided into three categories based on the availability of resources. The categories include high, medium, and low-resource. The resources comprise works of native speakers, online data, and computational resources. The resource-poor languages are classified into the low-resource category that has restricted online resources [63, 64]. Moreover, a low-resource language pair is considered based on the minimal data required for training a model [65]. The proper definition of low-resource language pair poses a challenging research question. However, if the training data is under 1 million parallel sentences, it is considered a low-resource language pair [66]. The native speakers play a vital role in different aspects of the language, including the quality and quantity of the data.

In several language pairs, NMT achieves state-of-the-art performance. However, the performance degrades when working with low-resource languages. For low-resource pair translation like English to Vietnamese and English to Farsi, NMT improved performance through the recurrent units with multiple blocks and a trainable routing network [67]. Different strategies such as monolingual data, back translation, multilingual, multimodal, and dealing with tonal language have been devised by researchers to enhance NMT's limited resource language.

- **Monolingual Data** - Monolingual data is a collection of texts written in a single language. It plays a crucial role in MT. While parallel corpora (text pairs in source and target languages) are the primary resource for training MT systems, monolingual data contributes significantly to various aspects of MT. It can generate synthetic parallel data, expanding the training dataset, which is crucial for low-resource languages in MT. The integration of monolingual data for NMT was initially examined by [66] [68]. They independently trained language models for the target side using monolingual data, which were then incorporated into the NMT model during decoding [69].

  A monolingual data-based NMT has been introduced without modifying system architecture [70]. With monolingual data to address the low-resource language problem, a filtering approach for the pseudo-parallel corpus is proposed [71] to increase the parallel training corpus. Unsupervised pre-train-based NMT is introduced [72, 61], where monolingual data of both source and target sentences are pre-trained and then fine-tuned in the trained model with original parallel data. Thus, monolingual data is a vital resource in MT, offering numerous benefits in language modeling, data augmentation, and domain adaptation. Despite domain mismatch and data quality challenges, techniques like back-translation, denoising autoencoders, and transfer learning effectively leverage monolingual data to enhance MT systems. As research progresses, integrating monolingual data will be crucial in developing robust and high-quality MT systems, particularly for low-resource languages and specialized domains.

- **Back Translation** - Out of various techniques to improve the restricted resource language of NMT, the most prominent and influential approach is back translation (BT) [70]. BT is a data augmentation method in MT for a low-resource language. It trains a translation model backward to create a synthetic parallel corpus from the target monolingual data. Applying BT to low-resource language, monolingual data can generate low-resource target sentences using the NMT-trained model. Then, the obtained synthetic parallel data can be used as additional parallel training data. However, the NMT performance degrades by directly augmenting BT data in the original parallel data. Therefore, to improve NMT performance, BT data filtering is necessary before adding original parallel data [73, 71]. Additionally, the concept of BT is improved by iterative back translation (IBT), where both the forward and backward directions of translations are used for mutual training. A set of low-resource language models augmented via IBT were trained, improving the output [74].

- **Multilingual Language** - Multilingual NMT systems can handle translation between numerous language pairings. The main idea behind these strategies is to transfer knowledge from high-resource to low-resource translation. Recent research suggests that multilingual models outperform bilingual models, especially when only a few languages are present in the system, and that the degree of relatedness between the languages also affects performance [75]. Moreover, among similar language pair translations in WMT19, NMT systems performed remarkably on Hindi-Nepali [76].

  The diversity of multilingual datasets enables the creation of robust and versatile MT systems capable of handling multiple languages. Despite challenges such as data quality and availability of resources, ongoing advancements in data collection, cleaning, and alignment techniques continue to enhance the effectiveness of multilingual data in improving MT performance. As the field progresses, integrating more diverse and comprehensive multilingual datasets will be crucial in advancing the capabilities of MT systems.

- **Multimodal Language** - Multimodal machine translation (MMT) deals with information extraction from multiple modalities, assuming the additional modalities will provide valuable perspectives on the source data. It is, therefore, utilized to enhance the translation of low-resource language pairs using extra modalities. Hence, it enables NMT to be an efficient approach for both low and high-resource language pairs. There are several significant multimodal translation studies for language pairs such as English-German [77], English-Hindi [78] [79], English-Assamese [80]. Multimodal Neural Machine Translation (MNMT) is developed by employing a doubly attentive decoder, using two independent attention techniques for the source textual and visual components, respectively [77]. For the English-Hindi language pair, the Hindi Visual Genome 1.1 [78] is deployed for CNN with VGG19 to extract local and global characteristics from the pictures. It outperformed text-only NMT in terms of automatic evaluation metrics. An MNMT system in English–Hindi language pair [79] utilized synthetic data to build a translation system for image features. Further, MNMT surpasses text-only NMT for English-Assamese translation using the Assamese visual genome 1.1 [80]. Despite the varied viewpoints on the effectiveness of MT, it is proved that visual input produces a better translation quality than only text input [81].

- **Tonal Language** - Tonal languages, where variations in pitch and tone within syllables can alter word meanings, offer distinct challenges and opportunities in MT. Examples of tonal languages include Mandarin Chinese, Vietnamese, Thai, and various African languages. Translating tonal languages requires careful attention to the tonal words that affect meaning. MT systems can significantly improve their performance in tonal languages by incorporating tone information, using contextual embeddings, and leveraging techniques like data augmentation and transfer learning. The ongoing development and refinement of these approaches will be crucial in achieving high-quality translations for tonal languages, thus broadening the applicability and effectiveness of MT technology. In a low-resource tonal language like Burmese with English pair, NMT with BT strategy shows remarkable performance [82]. On the other hand, no previous work focuses on the tonal words of the Mizo language in MT.

In summary, recent works in MT highlight the integration of transformer architectures, transfer learning, multimodal approaches, improved evaluation metrics, and hardware optimization. These advancements collectively enhance the capabilities of automated language translation, ensuring better performance and broader applicability across diverse languages and contexts.

2.2.2 Machine Translation on Indian Languages

MT for Indian languages presents challenges and opportunities due to the region's linguistic diversity and cultural richness. India is a linguistically diverse country with 22 officially recognized languages and hundreds of dialects. Developing MT systems for Indian languages presents unique challenges due to the various scripts, grammatical structures, and vocabulary. Languages such as Hindi, Tamil, Bengali, and others exhibit complex morphological structures and syntactic variations that pose significant challenges for traditional MT systems [83]. Many of these languages often lack sufficient parallel corpora (bilingual data) required for training robust MT models, which is crucial for achieving high translation accuracy.

Recent advancements in NMT have shown promise in addressing some of these challenges by capturing intricate linguistic patterns and context dependencies inherent in Indian languages [17]. For morphologically rich Indian limited-resource languages [84], an NMT model utilizing self-attention multihead and byte- pair-encoded is presented to construct an effective translation strategy that overcomes the Out-Of-Vocabulary barrier. Moreover, applying multimodal approaches, integrating visual or other contextual information alongside textual data, can enhance translation quality further [85].

The NMT has been investigated with RNN for low-resource pairs like English to Punjabi, English to Tamil, and English to Hindi and observed that performance increases with an increase in parallel train data [17]. The NMT shows promising English to Hindi translation results on the benchmark dataset [86, 18]. MT for Indian languages is an evolving field with significant challenges and opportunities. Addressing data scarcity, script diversity, and complex morphology requires tremendous effort. By leveraging advanced neural models, multilingual training, and community-driven initiatives, the quality and accessibility of MT for Indian languages can be significantly enhanced, promoting better communication and understanding across India's diverse linguistic landscape.

2.2.3 Prior Works on English↔Mizo Machine Translation

The Mizo language can be categorized under the language group, with words having diacritics [87], as the tone markers represent the tonal words. Mizo language poses significant challenges for MT due to limited resources and unique linguistic characteristics. However, it is observed that Mizo words with tone markers are less frequent than those without tone markers[1], unlike Vietnamese [87], Yorùbá [88] and Arabic language [89].

As for the English↔Mizo language pair, limited work exists in the MT [16, 27, 90, 91]. It is mainly due to the need for more resources, as Mizo is a low-resource language. The prior works focus on developing English↔Mizo (En-Mz) parallel data to overcome the problem of the availability of resources for the En-Mz MT task. An English↔Mizo parallel corpus has been prepared and built [16]. A comparative study was performed on the prepared corpus between PBSMT and NMT models. Automatic evaluation metrics such as BLEU, METEOR, F-measure, and HE scores have been used to assess the performance of translations predicted by the trained models. Here, NMT outperforms SMT. Multiple attention-based NMT models, such as RNN, BRNN, and transformer, were also employed to analyze the English-

[1] `https://vanglaini.org/`

Mizo pair [27]. Subsequently, an enhanced NMT technique is implemented on English-Mizo language pair[90]. The predicted output is evaluated based on various aspects, including sentence length alteration, comparison to current baselines, and analysis of predicted errors. The model's prediction errors and prediction accuracy are analyzed depending on changes in sentence length. Consequently, all these attention-based models have been examined with parallel data only. Monolingual data of the Mizo language have yet to be incorporated to improve such low-resource pair translation. Furthermore, to the best of available knowledge, no research has addressed linguistic challenges such as the tonal words of Mizo for English$\leftrightarrow$Mizo translation.

Besides MT, the Mizo language has been studied in several fields in NLP, such as multiword expression of Mizo language[92], defining rules for identifying named entity classes in Mizo language[93], the creation of a Mizo-to-English dictionary and a developing part-of-speech tag set for the Mizo language, which will be utilized to create an automatic POS tagger [1, 2, 94, 95].

Chapter 3

The Study of the Structure and Challenges of Mizo language

3.1 The Mizo Language

The Mizo[1] language belongs to the Sino-Tibetan family of languages. It is spoken natively by the Mizo people (also known as Lushai) in the Mizoram state of India and Chin State in Burma. The Mizo language is the official language used in the state of Mizoram, which is situated in the northeastern part of India. It shares borders with three states in northeast India: Tripura, Assam, and Manipur. Additionally, the state shares a border with two neighboring countries: Myanmar and Bangladesh. The name Mizoram comes from the words *'Mi'* which means *'people'*, *'Zo'* which means *'hill'*, and *'Ram'* which means *'land'*. Thus, the word *'Mizoram'* implies a *'hilly people's land'* [1]. It holds the second least populated state with a population of 1,097,206 according to the 2011 Census of India[2]. The Mizo sub-tribes have their respective dialect. The Mizo language [92, 1] is mainly based on the *Lusei* dialect, and many words are also derived from its surrounding Mizo sub-tribes and sub-clan. The Lusei dialect is accepted overall as the lingua franca of the Mizo people, with considerable influence from other sub-tribes, and also due to its widespread and exclusive use by Christian missionaries and the later young generation.

The writing system of the Mizo language is based on the Roman script. The Mizo alphabet has 25 letters as in Table 3.1. It was devised by the first British Christian missionaries of Mizoram, Rev. J.H. Lorrain and Rev. F.W.Savidge [96].

[1] `https://en.wikipedia.org/wiki/Mizo_language`

[2] `https://www.census2011.co.in/census/state/mizoram.html`

A	**AW**	**B**	**CH**	**D**
E	**F**	**G**	**NG**	**H**
I	**J**	**K**	**L**	**M**
N	**O**	**P**	**R**	**S**
T	**Ṭ**	**U**	**V**	**Z**

Table 3.1: Alphabet of Mizo Language

The missionaries built up an approach to composing Mizo writings dependent on the Hunterian interpretation framework, created in India during the eighteenth century and derived from a framework by William Jones. In these alphabets are three letters with a combination of two letters represented as one letter: *AW, CH, NG*. Among the alphabets are six vowels, which are *A, AW, E, I, O* and *U*. A circumflex (ˆ) was added to the vowels to demonstrate long vowels, which were inadequate to express Mizo's tone completely. A *vowel* is a syllabic language unit pronounced without stricture within the vocal tract. Each of the vowels has its meaning by itself, representing the tone of a word. All the other alphabets are consonants that have no meaning but can be merged to form a syllable with a vowel. A *consonant* is a speech tone articulated with a complete or partial closure of the vocal tract in articulatory phonetics. Since the Christian missionaries of Mizoram devised the writing system, the numbering system of Mizo is also similar to the English numbering system; the only difference is in the pronunciation.

3.2 Low Resource Pair: English↔Mizo

Many of the world's languages are recognized as low-resource based on the availability of resources. The MT works are limited in India's north-eastern region, and the low-resource languages include Assamese, Boro, Manipuri, Khasi, Kokborok, and Mizo. Even though there are many MT systems accessible for significant Indian dialects, there are minimal resources for studying the Mizo language in MT.

As English is the most widely spoken language globally, English↔Mizo MT will help the Mizo community overcome its shortcomings in the modern age. However, the relationship between language pairs also significantly impacts the output of MT. Every language has unique linguistic characters that make up its phonological structure. With MT techniques, a closely related language pair gives better results than a language pair with substantial structural diversity. As for the English Mizo language pair, both languages are very different from each other as in the following:

- **Language Origin:** Each language has its unique linguistic framework. Similarly, the linguistic framework of a language pair can potentially affect the performance of MT. Regarding the English↔Mizo language pair, both languages are substantially distinct in their originating languages. English originated from the Indo-European family of languages, while the Mizo language belongs to the Sino-Tibetan family[3].

- **Tonality:** In terms of a tonal language, English is a non-tonal language, while Mizo refers to it as a tonal language where the tone of a syllable affects the lexical meaning of words. There are four tones in the Mizo language: high, low, rising, and falling [97]. A unique tone marker is used to signify unique tone variation.

- **Word-Order:** The syntactic order of both languages is different. The English language follows a word order of Subject-Verb-Object (SVO), while the declarative word order of Mizo is Object-Subject-Verb (OSV). Fig. 3.1 presents an example of a Mizo sentence with its translation in English. However, the word order in Mizo is not as rigid as in English. The word order of the Mizo language can alter depending on the formation of a phrase. Mizo language follows Object-Subject-Verb (OSV), and sometimes it follows Subject-Object-Verb (SOV) based on the sentence. Therefore, Mizo Language can be considered both OSV and SOV word order, as shown in Table 3.2.

- **Gender Distinction:** The proper names in English do not distinguish between genders. E.g., Dave, Robin. However, in Mizo, a gender indicator is appended to the suffix of all proper names to identify them. A letter 'i' suf-

[3] https://en.wikipedia.org/wiki/Mizo_language

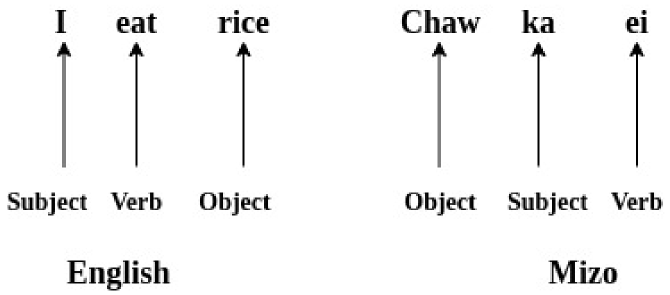


Figure 3.1: Example of English–Mizo word order

Sentence	**Structure**
Basketball ka khel (I play basketball)	OSV
Mawia'n savawm a kap (Mawia shot a bear)	SOV

Table 3.2: Word-order of Mizo

fixes all the female proper names, and a letter 'a' suffixes all the proper names of males.

– Rami *(indicates female since a letter **i** is append at the end)*

– Mawia *(indicates male since a letter **a** is append at the end).*

Additionally, in terms of personal pronouns, there is an appropriate gender distinction in English, like 'he' and 'she,' while it is impossible to determine gender in Mizo.

Based on the restricted availability of resources and the dissimilarities between the two languages, English↔Mizo can be classified as a low-resource pair and possess a challenging MT task. According to ISO, the language code of Mizo[4] is 'lus' and 'eng' is the language code for English[5]. However, En-Mz is used for English↔Mizo language pair, where *'Mz'* acronym is used for the Mizo language, which is the state abbreviation of Mizoram, and *'En'* acronym is used for the English language.

[4] https://iso639-3.sil.org/code/lus

[5] https://iso639-3.sil.org/code/eng

3.3 Challenges of English↔Mizo Machine Translation

Translation of a language is a complex task. Several challenges must be dealt with when translating one language to another. Like many other languages, the Mizo language deals with several challenges. Although both languages, Mizo and English, use the Roman script in the writing system and have the same number system, there are several linguistic differences between the two languages, as mentioned in Section 3.2.

Aside from the various strategies developed by researchers to enhance restricted languages in NMT, one of the most essential metrics is dealing with the linguistic challenges of a specific language. Therefore, tackling the research gaps between a language pair is one of the factors that will enhance the performance of the NMT. Several research gaps for English↔Mizo pair in the aspect of MT are as follows:

3.3.1 Tonal Words

A language is treated as tonal when its tone influences the word's meaning. Mizo language is undoubtedly a tonal language, which can lead to specific challenges for MT. Variations in tones and contour tones can alter the meaning of particular words. The type of pitch used can automatically determine the grammatical forms of that specific word. Many linguists have concluded the Mizo language to be of four tones, while some conclude it to be more than four tones by considering two vowel sound ways: long and short. However, the Mizo tone framework accepts four tones: High (H), Low (L), Rising (R), and Falling (F) [98]. The tones are also named in Mizo as *'Ri sang'*, *'Ri hniam'*, *'Ri lawn'* and *'Ri kuai'* respectively. The linguist created a tone marker for each tone to indicate the tone variation in the Mizo Language, which is listed in Table 3.3. The four tones used in Mizo words can indicate different meanings in the English word, as shown in Table 3.4. For

Type	**Tone (a)**
High tone	á
Low tone	à
Rising tone	ă
Falling tone	â

Table 3.3: Variation of tone (a) in Mizo

example, the Mizo word *'buk'* can indicate different meanings in English words like *'bushy'*, *'weight'*, *'hut/camp'*, *'unstable'*, which is to determine based on the tone used. Although the Mizo is undeniably a tonal language, the indication of tonal words in the writing system is neglected and not correctly considered. Even though there are four different tones with four tone markers, most of the writings in Mizo use only circumflex (ˆ) for indication of tone. Furthermore, it is also an understudied language with limited resources in terms of tones.

3.3.2 Tonal Polysemy

Mizo language is also rich in polysemy words where the intonation is the same, yet its meaning differs. Polysemy is a side of linguistic ambiguity that considers the multiplicity of word meanings. Table 3.5 presents examples of tonal-polysemy words in Mizo. It is a simple fact of common parlance, and people gleefully interpret correct results without conscious effort. However, polysemy is largely impervious to any generalized NLP task. As tonal languages go, the Mizo language is one of the most complicated languages in terms of MT. It is a tonal language where not only does a particular word have several tones, but also it is a language in which the word's pitch defines the meaning. However, polysemy is the association of a word with at least two distinct purposes. Since polysemy words have the same tone, the word's pitch alone cannot define the word. Therefore, understanding the nearby word or the whole sentence's context is necessary. A few polysemy words in the Mizo language can also act as nouns and verbs. For example:

- Engzat nge **mikhual** in thlen ? **(Noun)**
  I lo zin hunah ka **mikhual** ang che **(Verb)**

Mizo Word	**Tone**	**English Meaning**	**Mizo**	**English**
buk	High tone *(búk)*	Hut	Kan ramah **búk** sak ka duh.	I want to build a hut on our land.
	Low tone *(bùk)*	Bushy	He Ui hian mei a nei **bùk** hle mai.	This dog has a bushy tail.
	Rising tone *(bŭk)*	Unstable	He dawhkan hi a **bŭk** ania.	This table is unstable.
	Falling tone *(bûk)*	Weight	Khawngaihin heng hi min lo **bûk** sak teh.	Please weigh this for me.
lei	High tone *(léi)*	Tongue	Doctor in ka **léi** chhuah turin min ti.	The doctor asked me to stick out my tongue.
	Low tone *(lèi)*	Soil	Thlai chí tuh nan **lèi** an chǒ.	They dig up the ground to plant seeds.
	Rising tone *(lĕi)*	Buy	Thil ka **lĕi**.	I am buying something.
	Falling tone *(lêi)*	Bridge	**Lêi** ah ka ding.	I am standing on the bridge.
awm	High tone *(áwm)*	To be present	Vawiin seminar ah a **áwm** m?	Is he present today at the seminar?
	Low tone *(àwm)*	To look after/stay	Ka naute chu kan nauàwmtu in a **àwm**	My baby is look after by our nanny.
	Rising tone *(ăwm)*	Chest	A **ăwm** nat avangin doctor hnenah a inentir.	She went to the doctor complaining of chest pains.
	Falling tone *(âwm)*	Probably/likely	Inneihna ah a kal a **âwm** viau ani.	It is very likely that he will go to the wedding.

Table 3.4: Examples of the tonal words in Mizo

Tonal-Polysemy	Tone	English Meaning	Mizo	English
ăng	Rising tone (ă)	To open the mouth	I ka **ăng** rawh le.	Open your mouth.
		Talk angrily	Kha kha ti suh a tia, a **ăng** vak a.	'Don't do that!' she shouted angrily.
búl	High tone (ú)	Beginning	A **búl** atangin lehkha kha chhiar rawh.	Read the paper from the beginning.
		Stump	Kawtah sawn thing **búl** a awm.	There is a tree stump at the courtyard.
		Near	Helai **búl** velah hian thingpui dawr a awm hnai m?	Is there a restaurant nearby?

Table 3.5: Example of tonal-polysemy words in Mizo

- Ruah a sur dawn sia, **púk** ah hian awm mai ang u **(Noun)**

  I pawisa ka lo **púk** ang e, I phal em? **(Verb)**

Additionally, a few extraordinary words can change their tone depending on the phrase used but still have the same meaning. For example:

- **lĕi - Buy** (Raising tone)
  - Thil ka **lĕi** – > I am buying something. (Raising tone)
  - Khawiah nge I **lêi?** – > where did you buy? (sound as falling tone)
- **áng** – **will** (High tone)
  - Ka ti vek **áng** – > I will do everything. (High tone)
  - Chhang hi ka zai sak **àng** che – > I will cut this cake for you. (sound as low tone)

Types of clothes	**Different vocabulary for *'wear'* in Mizo**
Pants or shirt	Hà
Skirt	Féng
Shoes or Socks	Bún
Hat	Khúm
Belt	Hréng

Table 3.6: Vocabulary distinction

- **Chhûm – boil** (Falling tone)
  - Naute tui I pek dăwn chuan I **chhûm** so phawt dawn nia – > If you give water to a small baby, you have to boil it first. (falling tone)
  - I **chhúm** zawhah gas off rawh – > Off the gas after you boil. (sound as high tone)

3.3.3 Vocabulary Distinction

Although the English vocabulary is undeniably rich, Mizo has a richer vocabulary than English for a few specific words. For example, as in Table 3.6, the word *'Wear'* in English can refer to different vocabulary based on the clothes wear in Mizo.

3.3.4 Symbolic Words

Apart from the tone, symbols such as – and **'n** are used in the writings of the Mizo sentence.

- **Hyphen – :** In many places, – (hyphen) is found, which is used for continuing English (non-Mizo) words with Mizo words to appear as one word. Hyphen is also used when combining figures with words.
- **'n :** *'n* is used after the noun to show possession of the noun. It works as putting (an apostrophe) in an English sentence.

	Hyphen	**'n**
Symbolic Words	8,307-in	worker-te'n
	database-ah	Lalruatkima'n
	district-a	20-te'n
	police-te	hnathawktute'n

Table 3.7: Examples of symbolic words in Mizo

Table 3.7 demonstrates symbolic words in Mizo. Moreover, a few words are significant with having affix words. For instance, *'ah'* is an affix word that is a preposition (can be used as: at, on, upon, in, into) depending on the sentence. Combining the same word and the affixed word to produce one syllable of a linguistic unit may lead to a different meaning but an entirely correct Mizo word. For example:

- *Ru-ah* (steal) $->$ *Ruah* (Rain)
- *Chi-ah* (salt) $->$ *Chiah* (Dip)

3.4 Development of Parallel Corpus: English↔Mizo

The two prominent approaches of MT, SMT, and NMT, require a large amount of training data to provide promising results, which is a significant problem for low-resource pairs like English↔Mizo. It is a challenging task to prepare the parallel and monolingual corpora for English↔Mizo.

3.4.1 Corpus Details

Building a parallel corpus for MT involves several steps and considerations to ensure that the corpus is comprehensive, high-quality, and valuable for training MT models. An English↔Mizo parallel corpus is built to explore the field of MT. The parallel corpus is collected from various online sources namely, Bible[6], online dictionary

[6] https://www.bible.com/

Corpus	Source	Sentences
Parallel	Bible	26,086
	Online Dictionary (Glosbe)	70,496
	Government Websites	31,518
	Elementary Textbooks	45,058
	mCovid-19 Websites	21,480
	Movie Subtitles	10,529
	Tonal and Symbolic words	46,509
	Total	251,676
Monolingual	Web Pages/Blogs/Text Book	2,061,068

Table 3.8: Sources and statistics of English↔Mizo Corpus

(Glosbe)[7], government websites[8] [9], elementary textbooks[10], mCovid-19 websites[11], various movie subtitles and different web pages/blogs. Much of the data is also manually gathered to deal with tonal words. Monolingual data on the Mizo language has also been prepared. Table 3.9 demonstrates example sentences collected from various sources where the tonal words in the sentences are marked as bold.

3.4.2 Corpus Extraction Approaches

Corpus extraction refers to retrieving linguistic data from various sources to create a corpus, a structured collection of texts used for linguistic analysis or language model training. It is a fundamental step in linguistic research, NLP, and computational linguistics, providing valuable data for various applications such as MT, text mining, sentiment analysis, and language modeling.

Scrapy, an open-source framework, a web crawling technique, was used to extract the corpus. Acquiring parallel data through web scraping involves automatically extracting bilingual data from online resources. In Scrapy, xpath of each element is coded with a degree of generalization, which helps to crawl numerous web pages by replicating multiple web pages. To extract text from the PDF or image files, the

[7] `https://glosbe.com/en/lus`

[8] `https://finance.mizoram.gov.in/`

[9] `https://dipr.mizoram.gov.in/`

[10] `https://scert.mizoram.gov.in/`

[11] `https://mcovid19.mizoram.gov.in/`

Corpus	En	Mz	Source
Parallel	In the beginning God created the heavens and the earth.	A **tîrin** Pathianin lei leh **vân** a siam a.	Bible
	He will guide the humble in justice.	Retheite chu dik takin ro a **rêlsak** ang.	
	What questions do we need to answer?	Eng zawhnate nge kan **chhân** ang?	
	What is humility?	**Inngaihtlâwmna** chu eng nge ni?	Glosbe
	GSDP which is at an approximate level compared to previous year's figure.	GSDP atanga **chhût** erawh hi chu nikum dinhmun nen a intluk tlang a ni.	
	And the gate was shut as soon as the pursuers had gone out.	A **ûmtute** chu an chhuah veleh kulh kawngka chu an **khâr** ta a .	Government Website
	advance	**hmasâwn**	Tonal Word
	punch	**hnék**	(Manually Prepared)
	At Famous	'Famous'-ah	Symbolic word
	God for ever	kumkhua-in—Pathian	(Manually Prepared)
Monolingual		Schedule tribe-te chu income tax **àwl** an ni thin tih sawiin Zoramthanga chuan.	Web pages/Blogs/Text Book
		Mi **tlâwmte** chu a kawng a **zirtîr thîn**.	

Table 3.9: Examples of parallel and monolingual sentences

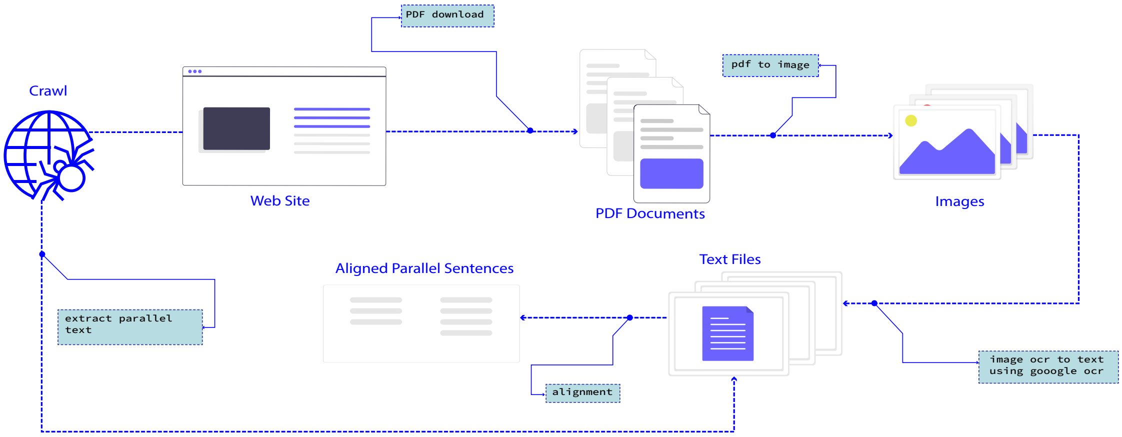


Figure 3.2: Data acquisition

Google OCR[12] tool is used. It is mainly used to extract Mizo data from textbook[13] (Government website). Fig. 3.2 depicts the overall data acquisition.

Furthermore, a manual effort has been used to prepare parallel data, mainly extracted from government websites. From the monolingual data of Mizo, tonal and symbolic words are extracted and translated manually to their corresponding English words. The alignment of the manual procedure took several months. Additionally, the Mizo sentences are cross-verified by a linguistic expert of Mizo, who is a native speaker and possesses linguistic knowledge of Mizo.

3.4.3 Data Cleaning and Split

The prepared corpus contains noise like too many special characters, web links (URLs), blank lines, and duplicates. As a result, the redundant phrases and noise must be eliminated. During data cleaning, conversion of lower-case and punctuation removal is not performed to maintain the semantic contextual meaning [99]. The prepared corpus is then split into three categories: training data, validation data, and test set. A sentence with tonal words is considered during the partition of validation and test data.

[12] https://cloud.google.com/vision/

[13] https://scert.mizoram.gov.in/

3.4.4 Domain Coverage

Domain coverage in MT refers to the capability of an MT system to accurately translate texts from various specific fields or subject areas. Achieving broad and deep domain coverage is essential for making MT systems useful across diverse real-world applications. By leveraging advanced techniques such as transfer learning, terminology integration, and active learning, MT systems can achieve higher accuracy and reliability in specialized domains. The ongoing advancements in AI and machine learning are expected to enhance further the capability of MT systems to cover a broader range of domains with greater precision.

Various domains have been covered to develop parallel corpus in English↔Mizo language pair. They are:

- **Bible :** The Bible is one of the most translated texts in the world and has been widely used as a resource for developing and testing MT systems. Its availability in many languages, consistent structure, and varied linguistic content make it valuable for MT research and development.
- **Government Websites :** Data from government websites are a rich source for creating a parallel corpus for MT. The government website is useful for the English-Mizo language pair since it contains data in both languages[14]. It is a good domain for MT due to its broad content range. Government documents often contain specialized terminology and ensure high-quality translation, regularly updating and maintaining data.
- **Movies Subtitle :** Movie subtitles are a valuable resource for developing and improving MT systems due to their informal language and the conversational nature of their content. Subtitles include colloquial language, idioms, slang, and cultural references, which are crucial for training models to handle everyday speech.

[14] `https://dipr.mizoram.gov.in/`

- **Online Newspaper :** Online newspapers offer diverse topics, regularly updated content, and formal language, ideal for training and testing MT systems. Therefore, a leading daily newspaper in Mizoram *Vanglaini*[15] is used for monolingual data of the Mizo language. It was published in the Mizo language. It has a column for local, northeast, country, world, youth, and sport, providing rich and diverse content for MT.

- **Elementary Textbook :** Using elementary textbooks for MT is beneficial as they provide clear contextual information with simple sentence structures, making them easier for MT models to process.

- **Online Dictionary(Glosbe) :** Glosbe is a multilingual online dictionary that provides translations, contextual examples, and user-contributed content for various language pairs. It can significantly enhance MT by providing rich linguistic resources, including vocabulary, phrases, idioms, and contextual examples. Especially for low-resource language pairs like English-Mizo, Globse is a valuable resource for the development of Corpus in MT.

- **General Domain :** Although developing a parallel corpus with specific domain data is helpful, it is necessary to have general domain data to handle a wide range of topics. It refers to text that is not specialized and encompasses a wide range of everyday topics, including news articles, novels, blogs, and general web content. The broad mix of text from various sources provides a baseline capability across many domains. By focusing on the general domain, MT developers can create more versatile, robust, and contextually aware translation systems that enhance global communication and understanding.

[15] `https://www.vanglaini.org/`

3.5 Conclusion

Developing effective MT systems for English↔Mizo language pairs presents unique challenges due to tonal language, data scarcity, linguistic complexity, and resource limitations. However, significant improvements can be made through collaborative efforts, dealing with linguistic challenges, and leveraging advanced techniques. As technology advances with focusing more on low-resource languages such as back-translation, transfer learning, and community involvement, it is possible to develop effective MT systems for English↔Mizo pairs. Continuous efforts in data collection, resource development, and collaborative research will be crucial in enhancing the quality and usability of MT systems for the English↔Mizo language pair, thus preserving and promoting linguistic diversity in the digital age.

Chapter 4

Machine Translation for English↔Mizo Pair Encountering Tonal Words

4.1 Introduction

MT attempts to minimize the linguistic barrier through automatic translation among natural or human-spoken languages. However, several challenges in the language pair need to be addressed for the MT task. Low-resource language in MT faces many challenges in terms of accuracy and comprehension due to limited exploration of the language. Limited resources called a low-resource in MT include limited online data [63] or computational tools [100]. However, to improve the performance in the output of an MT system, a large amount of bilingual corpus is needed, which is often a significant problem in low-resource pair translation.

There is a need for the bilingual corpus to encounter linguistic challenges for corpus-based MT. The MT system is considered for English↔Mizo pair by encountering challenges of Mizo tonal words. Although few research studies have been conducted, no prior work has addressed Mizo tonal words in low-resource English↔Mizo pair translation.

4.2 EnMzCorp1.0: English↔Mizo Corpus

As the low-resource English↔Mizo (En-Mz) pair has limited available options for parallel and monolingual data of Mizo, different viable resources have been explored to prepare the corpus. An En-Mz parallel corpus is prepared that contains a total of 130,441 sentences. Also, Mizo monolingual data is prepared. Table 4.1 presents

Corpus	Source	Sentences	Tokens	
			En	Mz
Parallel	Bible	26,086	684,093	866,317
	Online Dictionary (Glosbe)	70,496	1,438,445	1,674,435
	Government Websites	31,518	402,900	605, 365
	Tonal and Symbolic words (Manually Prepared)	2,341	2,341	2,341
	Total	130,441	2,524,779	3,148,458
Monolingual	Web Pages/Blogs/Text Book	1,943,023	-	25,813,315

Table 4.1: Corpus sources and statistics

Monolingual Data	Sentences	Tokens
Mz	14,957	243,657

Table 4.2: Extracted (Mz) data using Google OCR

the corpus of sources with statistics. The statistic of the extracted monolingual data is shown in Table 4.2

The prepared corpus has issues such as excessive special characters, web links (URLs), blank lines, and redundant entries. After eliminating noise and duplicate sentences, the total number of parallel sentences decreased to 118,449. Table 4.3 presents the split data for the training, validation, and test data. When splitting the data into validation and test sets, sentences containing tonal words are considered. Two test sets were also considered: Test Set-1 for in-domain data from the split data and Test Set-2 for out-domain data, which included different types of tonal words with a maximum length of 15 and was prepared manually. Following [101], small test data was compared to training data because it is used for the baseline system. In the train data, out of 115,249 Mizo sentences, 44,604 have tonal words. The corpus EnMzCorp1.0 covers various domains including Bible[1], online dictionary (Glosbe)[2], government websites[3] [4] and different web pages/blogs.

[1] https://www.bible.com/

[2] https://glosbe.com/en/lus

[3] https://finance.mizoram.gov.in/

[4] https://dipr.mizoram.gov.in/

Type	**Sentences**	**Tokens**	
		En	**Mz**
Train	115,249	1,308,563	1,462,070
Validation	3,000	78,083	82,470
Test Set-1	200	5,181	5,523
Test Set-2	200	1,312	1,608

Table 4.3: Statistics for train, valid and test set

4.3 Baseline System

PBSMT [13] and sequence-to-sequence model-based NMT (RNN, BRNN) were considered for baseline systems to provide benchmark translation accuracy for both the directions of translations in English↔Mizo pair. The EnMzCorp1.0 dataset and monolingual data of English (3 million sentences) from WMT16[5] were utilized.

4.3.1 Satistical Machine Translation

SMT consists of three modules: Translation Model (TM), Language Model (LM), and Decoder. Consider the translation task of English to Mizo, where the best Mizo translation (m_{best}) for the source English sentence (e) is formulated using Eq.4.1.

$$m_{best} = \arg\max_{m} P(m \mid e) \tag{4.1}$$

To estimate $P(m \mid e)$ for given source-target sentences, the probability distribution of all possible target sentences must be considered. This involves understanding what makes a good translation, which should possess two key aspects: adequacy and fluency. Adequacy means the target sentence retains the same meaning as the source sentence, while fluency means the target sentence is grammatically and stylistically correct. A good translation balances both adequacy and fluency. This concept can be formulated using Bayes Theorem, extending Eq. 4.1 as shown in Eq. 4.2.

[5] http://www.statmt.org/wmt16/translation-task.html

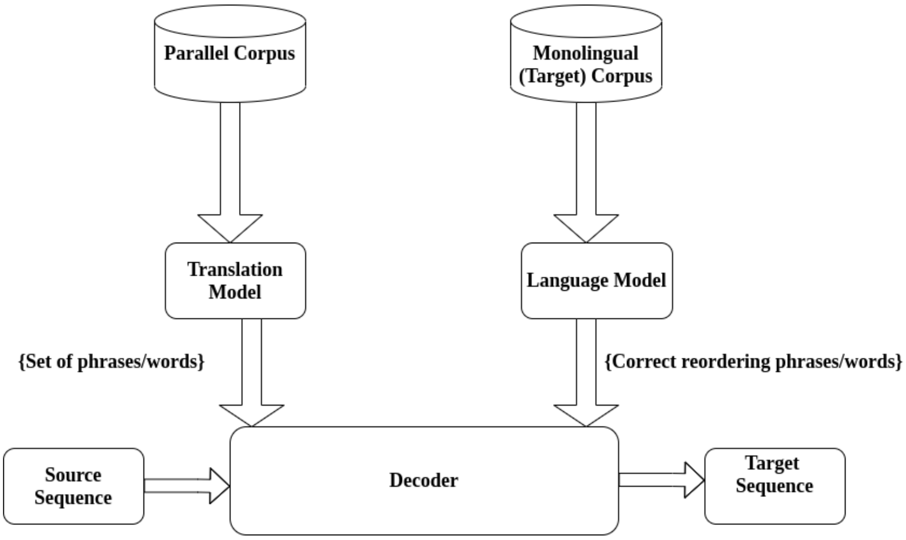


Figure 4.1: Abstract diagram of phrase-based SMT

$$\begin{aligned} m_{best} &= \arg\max_{m} adequacy(e \mid m) * fluency(m) \\ &= \arg\max_{m} P(e \mid m) * P(m) \end{aligned} \tag{4.2}$$

In the SMT, TM and LM compute $P(e \mid m)$ and $P(m)$. The decoder is responsible for $\arg\max_m$ to search for the best translation. The TM model collects phrase pairs from parallel data and then estimates the probable target words/phrases as shown in Eq.4.3.

$$P(e \mid m) = \frac{count(e, m)}{\sum_e count(e, m)} \tag{4.3}$$

The LM reorders the obtained target words/phrases from TM to predict syntactically correct target sentences to ensure fluency in translation. The LM is estimated from monolingual target data, where the conditional probability of each word models the target sentence given the previous words in the sentence. This modeling is also known as n-gram LM. The decoder uses a beam search strategy to find the best possible translation. The abstract pictorial representation of SMT is shown in Fig. 4.1.

4.3.2 Neural Machine Translation

RNN updates and maintains a memory known as a state during the processing of each word. The Eq. 4.4 represents probability of a sentence S and $S_1, S_2, ..S_n$ denotes a sequence of n words. The RNN-based LM [102] can be represented by considering the Equation 4.5, where next word S_{t+1} is predicted for the given current word S_t and previous words $S_1....S_{t-1}$.

$$P(S) = P(S_1, S_2....Sn) = \prod_{t=1}^{n} P(S_t \mid S_1....S_{t-1}) \tag{4.4}$$

$$\begin{aligned} S_{t+1} &= \arg\max_{t+1}(P(S_{t+1} \mid S_1....S_t)) \\ &= \arg\max_{t+1}(p_t) \\ p_t &= softmax(y_t) \\ y_t &= h_t^1 W_0 \\ h_t^1 &= tanh([h_t^0; h_{t-1}^1]W_h) \\ h_t^0 &= S_t E \end{aligned} \tag{4.5}$$

The RNN-based LM processes each word in a sentence at every time step t to predict the next word. Here, the vocabulary size and all hidden layers are considered V and M, respectively. The current word S_t is transformed into continuous space representation via indexing into the embedding matrix E provides h_t^0. The embedding S_t having vector size V is equivalent to the vocabulary size, where indexing is performed through one-hot vector representation. In one-hot vector representation, a '1' indicates the current word's index position; for all other positions, it is denoted by a '0'. It helps to create the embedding by multiplying the one-hot vector with size $1 \times V$ with the embedding matrix of size $V \times M$.

The RNN-based LM maintains memory using a hidden state called the previous state, h_{t-1}^1. The previous state is set to all zeros' vectors when the first word is encountered in the sequence. The previous state generates the concatenated vector and the embedding h_t^0 and then multiply with the matrix W_h having size $2 \times M \times M$ followed by the $tanh$ non-linear function. As a result, the hidden state h_t^1 is obtained at the current timestamp t, and the current hidden state represents the previous state for the next consecutive word in the sequence. The obtained h_t^1 is used for mapping to a vector y_t having size N via multiplication with the matrix W_0. Then, the softmax function is used to transform the vector y_t (also known as logits) into probability values, which provides the vector p_t. The predicted next word is the optimum probability value corresponding to the index position. This process of predicting S_{t+1} given S_t is required to update the neural network parameters by computing the cross-entropy loss between next word predictor p_t and the actual next word S_{t+1}. The cross-entropy loss is calculated for the entire sequence in the forward pass of the neural network. Then, the obtained total loss is used to calculate the prediction error through the backward pass.

Further, RNN considers long short-term memory (LSTM) [103] or gated recurrent unit (GRU) [104] for encoding and decoding to enhance learning long-term features. There are two main units of NMT: encoder and decoder where the encoder is used to compact the whole input/source sentence into a context vector, and the context vector is decoded to the output/target sentence by the decoder. Such basic encoder-decoder-based NMT cannot capture all important information if the sequence is too long. Therefore, the attention mechanism comes into existence [15, 41] that allows the decoder to focus on different segments of the sequence locally (part of the sequence) as well as globally (associating all the words of the sequence). Fig.4.2 depicts attention-based RNN, where the input Mizo sentence *'Thil ka lĕi'* is translated into the target English sentence *'I am buying something'*.

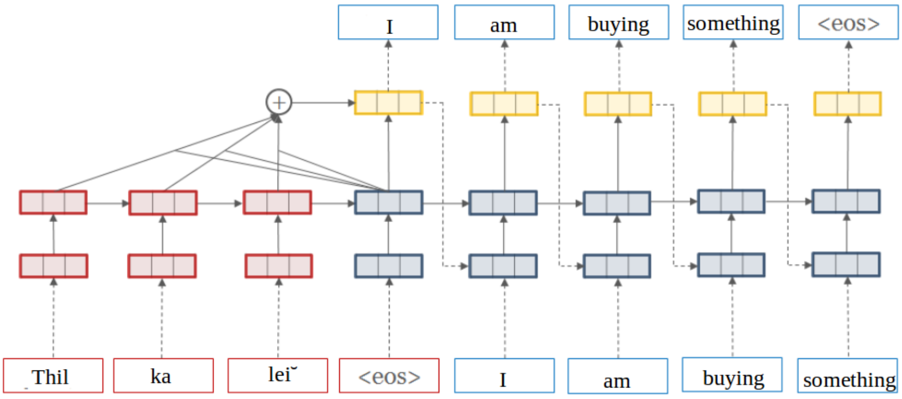


Figure 4.2: English–Mizo NMT system (attention-based RNN)

4.3.3 Experimental Setup

SMT and NMT setup have been followed by employing Moses[6] and OpenNMT-py[7]toolkit respectively. The SMT and NMT setup builds phrase-based (PBSMT), RNN, and BRNN-based NMT systems. For PBSMT, GIZA++ and IRSTLM [105] are utilized to produce phrase pairs and language models following the default settings of Moses. A 2-layer long short-term memory (LSTM) network of encoder-decoder architecture with attention is used for RNN and BRNN [15]. The LSTM contains 500 units at each layer. The Adam optimizer with a learning rate 0.001 and drop-outs of 0.3 is used in RNN and BRNN models. Unsupervised pre-trained word vectors of monolingual data using Glove[8] [106] and pre-trained up to 100 iterations with embedding vector size 200 were utilized.

4.3.4 Results

Automatic evaluation metrics and HE are considered when evaluating the baseline predicted sentences. The automatic evaluation metrics viz. BLEU [59], TER [107], METEOR [60] and F-measure.

[6] http://www.statmt.org/moses/
[7] https://github.com/OpenNMT/OpenNMT-py
[8] https://github.com/stanfordnlp/GloVe

- **BLEU**: The BLEU score is a metric designed to assess the quality of a machine-translated text. It measures how closely the generated translations match a set of reference translations, typically by comparing sequences of words or ‘n-grams’. It utilizes the modified precision of n-grams by comparing the n-grams of the candidate (predicted) translation’s n-grams with the reference translation. Eq. 4.6 represents the formula for the computation of the BLEU score. Here, P_l and R_l denote the predicted and reference translation length, respectively. Pd_i represents precision score of i^{th} gram. As for translation that is too short, a brevity penalty equal to 1.0 is considered when the candidate translation length is the same as the length of any reference translation. It is recommended to consider the lower value of n when the translation system is not adequate [16].

  $$BLEU = min(1, \frac{P_l}{R_l})(\sum_{i=1}^{n} Pd_i)^{\frac{1}{n}} \tag{4.6}$$

  In this work, the n-gram $n = 3$ is considered because the BLEU score tends to zero while crossing the tri-gram score. Table 4.4 presents the BLEU scores of baseline systems for both directions of translation.

Translation	**Test Data**	**PBSMT**	**RNN**	**BRNN**
En to Mz	Test Set-1	16.23	18.27	18.41
	Test Set-2	2.60	3.24	3.44
Mz to En	Test Set-1	17.18	19.20	20.12
	Test Set-2	2.75	3.47	3.49

Table 4.4: BLEU scores of baseline systems

- **TER**: The TER score is an automatic metric used to evaluate the quality of MT outputs. It calculates the actions required to update a candidate translation to align with the reference translation. It is a technique used in MT to measure the amount of post-editing effort needed for the output of MT. TER is computed in Eq. 4.7 by dividing the number of edits (N_{ed}) needed to adjust the candidate translation to match the reference translation by the reference translation's length (L_{rw}).

$$TER = \frac{N_{ed}}{L_{rw}} \tag{4.7}$$

Possible edits include insertion, deletion, substitution of single words, and shift of word sequences. The cost of all the edits is the same. Consider the following scenario of candidate translation and reference translation where italics highlight the mismatch:

Reference translation: Fruits are healthy *tasty and nutritious* loaded with *fiber* and vitamin

Candidate translation: Fruits are *tasty and* healthy loaded with *minerals antioxidant* and vitamin

From the above scenario, even if the candidate translation is fluent, TER, on the other hand, would not accept it as an exact match. The possible edits are as follows:

tasty and: shift (1 edit)

nutritious: insertion (1 edit)

minerals antioxidant: substitution for fiber (2 edits)

The total number of edits is 4 (one shift, one insertion, and two substitutions). The length of the reference word is 11. Therefore, the TER score becomes $\frac{4}{11} = 36\%$. Lowering the value of the TER score will improve accuracy. Table 4.5 presents TER scores for the baseline systems.

Translation	**Test Data**	**PBSMT**	**RNN**	**BRNN**
En to Mz	Test Set-1	80.1	78.90	75.0
	Test Set-2	102.6	102.4	101.8
Mz to En	Test Set-1	76.80	74.50	73.60
	Test Set-2	95.30	93.80	93.40

Table 4.5: TER (%) scores of baseline systems

- **METEOR and F-measure**: The METEOR score is a different metric designed to assess the quality of the machine-translated text. The metric relies on the harmonic mean of unigram precision and recall, emphasizing recall more than precision. It is calculated by computing a word alignment based on matching the three modules: an explicit word, stem word, and synonym word between the predicted and reference translation. These three modules work together to ensure the alignment between the two translations. The uni-gram precision P_{ug} and uni-gram recall R_{ug} are calculated using Eq. 4.8 and 4.9.

$$P_{ug} = \frac{T_m}{p_n} \tag{4.8}$$

$$R_{ug} = \frac{T_m}{r_n} \tag{4.9}$$

Where p_n and r_n are a number of uni-grams in the predicted reference translation, respectively. T_m denotes the number of matched uni-grams words between candidate and reference translation.

During the computation of the METEOR score, the F-measure score is calculated, which is the harmonic mean of precision P_{ug} and recall R_{ug} as shown in Eq. 4.10.

$$F - measure = \frac{2 \times P_{ug} \times R_{ug}}{P_{ug} + R_{ug}} \tag{4.10}$$

Also, the F-mean is calculated by the parameterized harmonic mean of the precision P_{ug} and recall R_{ug}. Then, METEOR is computed using Eq. 4.11.

$$METEOR = (1 - Pen) \times F_{mean} \tag{4.11}$$

Where fragmentation penalty (Pen) is calculated by fragmentation fraction (frag) and γ in Eq. 4.12 to account for the degree to which the uni-grams in both translations are in the same order. γ is the maximum penalty determined by the value ranging from 0-1. To compute fragmentation fraction (frag), the number of chunks (ch), which is a group of matched uni-grams that are adjacent to each other with having the same word order in both translations, is divided by the number of matches (m) as given in Eq. 4.13. METEOR and F-measure are assigned, ranging from 0 to 1 in each segment.

$$Pen = \gamma \times Frag \tag{4.12}$$

$$Frag = \frac{ch}{m} \tag{4.13}$$

Table 4.6 and 4.7 present METEOR and F-measure scores for the baseline systems.

Translation	Test Data	PBSMT	RNN	BRNN
En to Mz	Test Set-1	0.1626	0.1795	0.1812
	Test Set-2	0.0792	0.0794	0.0811
Mz to En	Test Set-1	0.1783	0.1856	0.1904
	Test Set-2	0.0893	0.0920	0.0925

Table 4.6: METEOR scores of baseline systems

Translation	Test Data	PBSMT	RNN	BRNN
En to Mz	Test Set-1	0.3832	0.4139	0.4179
	Test Set-2	0.1872	0.1961	0.1970
Mz to En	Test Set-1	0.4049	0.4175	0.4316
	Test Set-2	0.2103	0.2114	0.2140

Table 4.7: F-measure scores of baseline systems

- **Human Evaluation**: HE is a manual evaluation metric used to evaluate the predicted sentence of the MT systems [16]. Although automated metrics such as BLEU, METEOR, and TER offer rapid and consistent assessments, HE offers a more profound comprehension of translation quality by capturing nuances that automated approaches may overlook. The linguistic expert engaged in HE is acquainted with Mizo and English. The expert is well-versed in the complexities and challenges of the Mizo language. A human evaluator evaluates the predicted translations based on adequacy, fluency, and overall rating.

  Adequacy is measured using the contextual meaning of the predicted translation that corresponds to the reference translation. Fluency is measured by considering the good formation of the predicted sentence in the target language, regardless of whether it corresponds to the reference translation. The overall rating is measured by computing an average score of both adequacy and fluency.

Consider an example of a reference translation as *'Small businesses have been exempted from the tax increase'* and the predicted translation as *'I am putting my hand on my table'*. The predicted translation is considered inadequate since it contains a different contextual meaning than the corresponding reference translation. The predicted sentence is also fluent; although the meaning differs entirely from the reference translation, it is well-formed in the target language. The overall rating[9] considers the average of the adequacy as well as fluency. The assessment criteria are measured on a scale of 1-5, with higher values indicating better performance [16]. The rating score is assigned for 50 predicted test sentences (randomly chosen). The HE scores were calculated using Eq. 4.14.

$$HE(Overall\ Rating) = \frac{n_{TAR}}{n_{TBR}} \times 100\% \tag{4.14}$$

Where n_{TAR} is the total average adequacy and fluency rating scores, n_{TBR} is calculated by multiplying the best rating score with a total number of questions, i.e., $5 \times 50 = 250$. Table 4.8 reports HE scores for the baseline systems.

Translation	**Test Data**	**PBSMT**	**RNN**	**BRNN**
En to Mz	Test Set-1	28.56	29.40	31.92
	Test Set-2	17.40	18.80	19.60
Mz to En	Test Set-1	29.24	30.08	32.92
	Test Set-2	18.60	19.20	20.80

Table 4.8: HE (Overall Rating (%)) scores of baseline systems

4.4 System Description for Encountering tonalily

The proposed approach is based on BT [70] strategy without modifying the model architecture. It consists of three operations. First, Mizo sentences having tonal words are extracted from Mizo's monolingual data. Secondly, extracted Mizo tonal

[9] `https://nlp.amrita.edu/mtil_cen/#results`

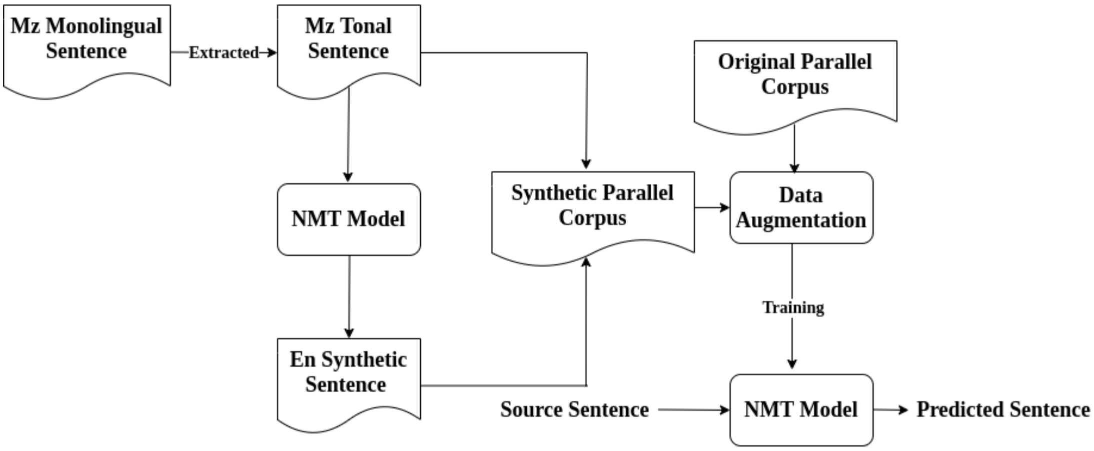


Figure 4.3: Proposed Approach

Parallel Corpus	Sentences	Tokens	
		En	Mz
Synthetic	33,229	550,238	610,376
Synthetic + Original	148,478	1,858,801	2,072,446

Table 4.9: Augmented train data statistics

sentences are used to generate the English synthetic sentences via Mizo's best translation model (BRNN) to English obtained from the baseline system. Then, the synthetic parallel corpus is augmented with the original parallel corpus. The main goal of the first two operations is to expand the parallel train data by increasing the Mizo tonal sentences. Lastly, the augmented data is used to train the NMT model (BRNN) independently for each translation direction. Fig. 4.3 depicts the pictorial diagram of the proposed approach. The original train data contains only 44,604 Mizo tonal sentences. Therefore, 44,000 Mizo tonal sentences were extracted using a maximum word length of 10. However, blank lines and single-word sentences were removed from the synthetic English sentences and their corresponding Mizo sentences. Thus, the synthetic parallel corpus contains 33,229 sentences, as shown in Table 4.9. For both Test Set-1 and Test Set-2, the automatic evaluation results of the proposed approach are reported in Table 4.10, and the human evaluation results of the proposed approach are shown in Table 4.11.

Translation	Test Data	BLEU	TER (%)	METEOR	F-measure
En to Mz	Test Set-1	20.21	73.4	0.1851	0.4272
	Test Set-2	4.04	100.5	0.0868	0.1992
Mz to En	Test Set-1	20.31	71.9	0.2022	0.4501
	Test Set-2	4.10	92.4	0.0931	0.2230

Table 4.10: Automatic evaluation results of proposed approach

Translation	Test Data	HE
		Overall Rating (%)
En to Mz	Test Set-1	32.24
	Test Set-2	20.40
Mz to En	Test Set-1	33.48
	Test Set-2	21.80

Table 4.11: Human evaluation results of proposed approach

4.5 Analysis

Among the automatic evaluation scores in Table 4.4, 4.5, 4.6, 4.7, and 4.10 on both test data, the proposed approach attains higher accuracy than baseline systems. As for human evaluation, the proposed system Table 4.11 also outperforms the baseline system Table 4.8. Furthermore, Test Set-1 (in-domain) accuracy is better than Test Set-2 (out-domain) for both automatic and human evaluation. It is noticed that Mizo to English translational evaluation scores outperform English to Mizo. Since train data contains more Mizo tokens than English tokens, as mentioned in Table 4.3. Therefore, the model encoded more Mizo word frequency, and the decoder can better translate Mizo to English. It is observed that all the system's output encountering tonal words has poor translational quality. Moreover, predicted output suffers under translation, is impotent in named-entity prediction, and has out-of-vocabulary issues. By the following notations, the predicted sentence samples are considered below to inspect the errors.

- ST: Source Test sentence.
- RT: Reference/Target sentence.

- PD1: Predicted sentence by the proposed approach.
- PD2: Predicted sentence by the BRNN.
- PD3: Predicted sentence by the RNN.
- PD4: Predicted sentence by the PBSMT.

1. Sample predicted sentence for En-to-Mz (partial adequacy but good in fluency)

ST: *They dig up the ground to plant seeds.*
RT: *Thlai chí tuh nan lèi an chŏ.*
PD1: *Mau hmanga lei laih an rél a.*
PD2: *Lung chi hrang hrang an han thlen chuan.*
PD3: *Lung angin lei an khuar a.*
PD4: *Pialtlêp chu an kân a.*

Discussion: The PD1 has encountered the tonal words *'lei'* meaning *'ground'* and generated the relevant meaning of the tonal words. However, it cannot detect the tone marker *è*. The word *'dig'* in the source sentence is predicted as *'laih'*, which is correct and also has a similar meaning as the tonal word *'chŏ'* in the reference sentence. The English meaning of the proposed approach is *'They decide to dig the ground with bamboo'*. In the predicted sentence, the word *'Mau'* means *'bamboo'* and a tonal word *'rél'* means *'decide'* are encountered which are not relevant to the source sentence. Both PD2 and PD3 predictions are inadequate and not fluent. However, PD4 translation is also inadequate but fluent. Thus, in terms of total words, the proposed approach can identify the tonal words, but the other baseline systems do not consider it for translation. Compared to baseline systems translation, the proposed approach has the best-predicted sentence since most words are correctly predicted. Therefore, it attains partial adequacy but is good in terms of fluency.

2. Sample predicted sentence for Mz-to-En (partial adequacy but good in fluency)

ST: *Naupang ruàlin pawnah an nghak.*

RT: *A group of children waited outside the door.*

PD1: *They are waiting for the child.*

PD2: *shun*

PD3: *There*

PD4:*books*

Discussion: The PD1 has identified the tonal word *'ruàlin'* in the source sentence and predicted it as 'they', which can be accepted as having a similar meaning with *'group'* in the reference translation. However, PD2, PD3, and PD3 do not recognize the tonal word. They incorrectly predicted the sentence by only one word, which is inadequate and not fluent. However, in PD1, the predicted sentence's contextual meaning is partially adequate compared to the reference translation. In terms of fluency, it is a well-formed Mizo sentence.

3. Sample predicted sentence for En-to-Mz (inadequacy but good in fluency)

ST: *There are many ants that crowd around sugar.*

RT: *Fanghmir tam deuhin chini an bâwm luai luai mai.*

PD1: *Gas agency tam tak an awm a.*

PD2: *Hnathawk vêlah chuan huaisen takin a awm a.*

PD3: *Hnathawk vêlah chuan mipa tam tak an awm a.*

PD4: *Chutah chuan mipa tam tak an awm a.*

Discussion: Both the PD1 and PD4 have not generated tonal words in their predicted sentence, while a tonal word *'bâwm'* means *'crowd'* appears in the reference text. However, PD2 and PD3 have both generated a tonal word *'vêlah'*, which means *'about'* in their predicted sentence, which is not relevant to the reference translation. The predicted translations are contextually inadequate for all the systems and have completely different meanings than the reference translations. However, in terms of fluency, the predicted sentences of all the systems are good in fluency.

4. Sample predicted sentence for Mz-to-En (inadequacy but good in fluency)

ST: *A hma a ka lo tilo kha ka ă hle mai.*

RT: *I was foolish not to have done it before.*

PD1: *I was very sorry that he had not come before him.*

PD2: *I did not know how I was good.*

PD3: *I didn't know how he didn't know it.*

PD4: *I did not know him until he was saying.*

Discussion: A tonal word *'ă'*, which means *'foolish'*, appears in the source sentence, but none of the systems can detect the source's tonal word. Here, the contextual meaning of all the predicted sentences is completely different from the reference translation. Therefore, they are termed as inadequate. As the predicted sentence is well-formed and proper in the target language, it is considered to be fluent.

5. Sample predicted sentence of named-entity error (En-to-Mz)

ST: *They moved the goal posts wider apart.*

RT: *Goal bàn an sawn zau.*

PD1: *Ruahpui vânâwn chu nasa takin an chelh a.*

PD2: *Thalai chu an tum ber tur tlat a ni.*

PD3: *latitudinal*

PD4: *Mitin chuan an ramri chu an pan ta a.*

Discussion: A tonal word *'vânâwn'* means *'down pour'* is generated in the PD1. However, there is no relevant word in the reference translation. On the other hand, a tonal word *'bàn'* appears in the reference translation, but all the systems cannot generate the tonal word in their predicted sentence correctly. Multiple errors exist in the named entity as the word *'goal'* appears in both source text and reference text. However, none of the systems have correctly generated their predicted sentence. Therefore, due to huge errors in named entities and contextually different predictions, the predicted sentences of all the systems are inadequate. Regarding fluency, parts of the prediction in PD1 and PD2 are correct, so they are partially fluent. However, PD3 predicts non-Mizo words, and PD4 predicts a proper Mizo sentence. Therefore, it is good in fluency but inadequate.

6. Sample predicted sentence of named-entity error (Mz-to-En)

ST: *I ka ăng rawh le.*

RT: *Open your mouth.*

PD1: *hushaby*

PD2: *I make it for you.*

PD3: *Let me get your grave.*

PD4: *I have to make it for y.*

Discussion: A tonal word *'ăng'* which means *'open mouth'* appears in the source sentence, but none of the systems can detect the source's tonal word. All the systems have encountered named-entity errors in their predicted sentences. While the reference translation is *'Open your mouth'*, none of the systems predicted the word *'open'* and *'mouth'*. PD1 predicts as *'hushaby'*, which is completely inadequate but fluent. Likewise, PD2 and PD4 have predicted contextually different sentences but are perfectly fluent. However, PD3 predicts an improper English sentence, which is also inadequate.

7. Sample predicted sentence of over-prediction (En-to-Mz)

ST: *Two children answered the teacher's question simultaneously.*

RT:*Naupang pahnih chuan zirtirtu zawhna a ruálin an chhăng.*

PD1: *Naupangte chuan zawhna an chhâng a, zawhna pahnih an chhâng a.*

PD2: *Fa pahnih chuan <unk> zawhna pakhat chu an chhâng a.*

PD3: *Fapa pahnih chuan zawhna pakhat chu an chhâng a.*

PD4: *16 Naupang pahnih chuan zawhna pakhat a chhâng a.*

Discussion: Two tonal words *'chhăng'* and *'ruálin'* appear in the reference translation. A word *'answered'* in the source text is correctly predicted by all the systems as *'chhâng'*. However, in all the predicted sentences, the tone marker is changed in *'chhâng'*, which is a falling tone, while in the reference translation, it is a rising tone. However, a tonal word *'ruálin'* from the reference translation, which means *'simultaneously'*, is unable to be generated by all the systems in their predicted sentence. From all the predicted sentences, it can be noticed that all of the systems encountered over-prediction. As the number of questions is not mentioned in the source test sentence, all the systems have predicted a sentence that includes the number of questions. PD1 predicts two questions, while PD2, PD3, and PD4 predict one question. Even though the predicted sentences by all the systems are incomplete and inadequate, all are well-formed, and therefore, it is fluent.

8. Sample predicted sentence of over-prediction (Mz-to-En)

ST: *kha kha ti suh a tia, a ăng vak a.*

RT: *Don't do that! she shouted angrily.*

PD1: *And do not do it in judgment and in crook.*

PD2: *Do not do that which is great in the eyes of him who is <unk>*

PD3: *not*

PD4:*And don't do not do that which is right in the eyes of Yahweh .*

Discussion: A tonal word *'ăng'* which means *'shouted'* appears in the source sentence, but none of the systems can detect the source's tonal word. The PD1 is over-predicted by adding *'judgment and in crook'*, which does not appear in the reference translation. Similarly, PD2 and PD4 have also been over-predicted by adding several words apart from the reference sentence. Although it is inadequate, it is good in fluency. Besides, *'<unk>'* is detected as part of the predicted sentence in PD2. However, PD3 has predicted only a single word *'not'*, which is inadequate.

9.Sample predicted sentence of under-prediction (En-to-Mz)

ST: *There was a bomb blast yesterday.* RT: *Niminah bàwm a puak.*

PD1: *<unk> puak a awm a.*

PD2 : *Nimin puak puak thei a awm.*

PD3: *Zanin chu a puak puak.*

PD4:*Niminah tu a lo awm.*

Discussion: A tonal word *'bàwm'* means *'bomb'* appears in the reference translation. However, none of the systems can correctly generate the tonal word in the predicted sentences. In the PD1, *'<unk>'* is generated as part of the predicted sentence. However, the prediction of all the systems encountered under-prediction as *'bomb'* and *'yesterday'* are not generated in the PD1. Although it is inadequate, it is good in fluency. Similarly, PD2 and PD3 have not mentioned *'bomb'*, and PD4 does not mention *'bomb blast'*. The predicted sentences of PD2, PD3, and PD4 are inadequate and are not well-formed Mizo sentences.

10. Sample predicted sentence of under-prediction (Mz-to-En)

ST: *kan bill kan pek hnuah èngzah nge la bâng áng?*

RT: *How much will we have left over once we've paid our bill?*

PD1: *And when we give the bill.*

PD2: *When our bill of our bill.*

PD3: *After the bill of our bill.*

PD4: *And when we get the Memorial, what does it <unk>*

Discussion: Three tonal words *'èngzah'* means *'How much'*, *'bâng'* means *'left'* and *'áng'* means *'will'* is encountered in the source sentence, but none of the systems can detect the source's tonal word. All the systems encountered under prediction where the predicted sentence predicts only part of the reference translation. It is inadequate as the reference translation has a contextual meaning that differs from the systems' predicted sentence. In terms of fluency, it is not a well-formed Mizo sentence.

4.6 Conclusion

In this chapter, EnMzCorp1.0 has been developed for the English-Mizo corpus, and the same has been used to build baseline systems for English to Mizo and vice-versa translations encountering tonal words. The proposed approach attains higher translation accuracy than baseline systems. From the analysis of predicted translations, it is realized that the system needs to be improved to encounter Mizo tonal words. Increasing the size of the dataset and exploring the knowledge-transfer-based NMT approach will enhance the performance.

Chapter 5

Building Low Resource English-to-Mizo NMT Model with Post Processing

5.1 Introduction

A multilingual country like India has an enormous linguistic diversity. The demand for developing Indian language resources is growing, with implications for various MT applications. However, MT efforts in India's north-eastern regions are limited, with many languages, including the Mizo language, considered low-resource. Low-resource language translation poses challenges in the field of MT. The challenges include the availability of corpus and differences in linguistic information. Therefore, building a parallel corpus, i.e., English-Mizo Corpus, is crucial for exploring MT tasks and contributing to advancing Indian language resources.

For a low-resource language pair: English-to-Mizo, NMT is employed. It has attained a promising approach in MT because of its context analysis ability and deal with long-range dependency problems [15, 26]. However, it needs sufficient training data, which is challenging for the low-resource language pair translation [54]. With Mizo being the tonal language, a distinct tone marker is used to represent the tonal words contextually. Based on the previous investigation in Sec 4.3, the baseline translation of English-Mizo MT suffers in handling the tonal words and their corresponding context. Table 5.1 shows an example where the baseline predicted sentence could not capture accurate tone markers (marked as 'bold'). Without tone markers, the meaning of the predicted sentence is ambiguous, corresponding to the source sentence. It can mean either 'What is the price?' or 'What did he catch?' but with a specific tonal marker, it is defined as the exact meaning of the sentence, i.e., 'What did he catch?'. As a result, the contextual meaning is not clear. To tackle the problem, a technique is proposed for encountering context-specific tonal words to improve the predicted sentence during the post-processing step.

Source / Target	**Predicted**
What did he catch? (Source)	Eng nge a man? (baseline)
Èng nge a m**á**n? (Target)	**È**ng nge a m**á**n? (Current Objective)

Table 5.1: Example of a predicted sentence (tone markers are marked as bold)

5.2 Low Resource NMT

Although NMT has struggled with low-resource language, various works have been done by several researchers to deal with it. The NMT methodologies can be categorized into supervised, semi-supervised, and unsupervised. A large-scale bilingual corpus is crucial for supervised NMT, lacking in low-resource languages. Semi-supervised techniques depend on the presence of parallel corpora in addition to monolingual corpora. Unsupervised learning approaches aim to build an MT system on a language pair with no parallel corpus. Both semi-supervised and unsupervised techniques deal with low-resource language pairs. Various approaches are considered to improve the quality of limited-resource pair translation. These approaches can be categorized into two groups: the traditional approach and the knowledge-transfer approach.

5.2.1 Traditional Approach

It mainly deals with data scarcity, the issue of rare words, and out-of-vocabulary. To obtain better translation quality, NMT requires sufficient parallel data. However, with low resource language, monolingual data is more accessible than parallel data from the viewpoint of resource availability. Therefore, the data augmentation technique is introduced by utilizing monolingual data to increase the size of the dataset. It is known as the self-training approach [108] under semi-supervised settings [109, 110]. Researchers have been investigating improving NMT by utilizing monolingual data[70, 109, 69] and BT. However, large-scale noise in BT data decreases the training performance and, as a result, lowers the translation quality [69]. Furthermore, a COPY model is set up, which is an alternative approach to BT. Although the out-of-vocabulary issue is improved in the COPY model, it increases the vocabulary only in the target data [69]. For this reason, filtering-based approaches are introduced to refine the synthetic data obtained from BT [19, 71].

With an unsupervised setting, pre-train word embeddings on monolingual data enhance the performance of low-resource NMT. The issues of rare words and out-of-vocabulary (OOV) are the crucial challenges of low-resource NMT [54] and byte-pair-encoding (BPE) [111] introduce to handle such problems.

5.2.2 Knowledge-Transfer Approach

It is primarily concerned with insufficient data for low-resource NMT through various approaches: transfer learning, multitasking, zero-shot, multilingual, and multimodal-based NMT. The multitasking approach executes multiple tasks in a limited time by employing various downstream NLP tasks, such as part-of-speech tagging, role-labeling, named-entity-recognition, and image caption generation [112]. The objective behind multitasking is to obtain train weights from extensive data and enhance the model's generalization capability. Furthermore, the investigation has been conducted by integrating parallel and monolingual data simultaneously to improve translation quality [61] using translation language modeling (TLM) and masked language modeling (MLM) objectives.

5.3 Approaches of Low Resources NMT

Researchers have developed various ways to improve the limited resource language in NMT. The following strategies analyze techniques for handling such problems. By applying these strategies, the limitation of low-resource language may be addressed to some extent, resulting in a significant improvement for NMT.

5.3.1 Back Translation (BT)

BT has emerged as a critical technique for enhancing MT systems, particularly for low-resource languages. It generates a synthetic parallel corpus from the target monolingual data by training a translation model in the backward direction. The model is then re-trained using this synthetic parallel data and the original one. The purpose was to create synthetic source sentences to maximize the number of parallel training data. BT has proven to be a substantial gain in NMT [70]. IBT is also proposed [113] to improve the idea of BT, where both forward and backward translations directions are utilized for training.

BT is applied to several African and Indic languages, showing significant improvements in translation accuracy [114]. It is highlighted that carefully curated monolingual data and diverse BT strategies could yield substantial gains. The benefits of BT for low-resource languages are manifold. It addresses the fundamental challenge of data scarcity by augmenting training datasets with high-quality synthetic data. BT also enables the exploitation of vast monolingual corpora in the target language, which are often more readily available than parallel data. It also facilitates the development of robust MT models capable of handling the complexities and nuances of low-resource languages. Additionally, BT has been shown to improve domain adaptation by focusing on specific domains.

Overall, BT has proven to be a powerful tool for advancing MT in low-resource settings. Recent studies highlight the importance of innovations in synthetic data generation, multilingual models, and domain-specific adaptations. As research continues to evolve, BT is poised to play an increasingly pivotal role in bridging the gap between high-resource and low-resource languages in MT systems.

5.3.2 Word Embedding

Word embedding techniques have revolutionized NLP, enabling the representation of words in dense vector spaces where semantic and syntactic similarities are preserved. For low-resource languages, where annotated datasets and linguistic resources are scarce, word embeddings provide a crucial foundation for developing robust language models by leveraging monolingual and cross-lingual data. Several concepts, such as Word2Vec [115] and GloVe [84], used word embedding techniques. The Word2Vec model capture linguistic patterns through distributed representations. The BPE is one of the techniques used for word segmentation [111]. It aims to break down words into subword units, making dealing with rare and unknown words easier. It is a data compression method in which the most often occurring pair of bytes in a sequence are replaced.

Studies have made significant advancements in word embeddings tailored to low-resource settings. FastText model uses subword information to generate embeddings, making it particularly beneficial for morphologically rich and low-resource languages [116]. Word embeddings enable efficient use of limited data, enhance cross-lingual transfer, and support downstream tasks like machine translation, information retrieval, and sentiment analysis. Additionally, subword-based methods mitigate the challenges of rare and out-of-vocabulary words, while multilingual embeddings foster language understanding even without extensive training data. The word embedding technologies continue to bridge the gap between low-resource and high-resource languages, driving advancements in NLP for diverse linguistic communities.

5.3.3 Transfer Learning

Transfer learning has emerged as a transformative approach in NLP, enabling the development of effective models for low-resource languages by leveraging knowledge from high-resource counterparts. This paradigm involves pre-training a model on a large dataset in one or multiple languages and fine-tuning it for specific tasks or languages with limited data. Transfer learning addresses the fundamental challenge of data scarcity, a persistent issue in low-resource language processing.

BERT, a pre-trained transformer model fine-tuned for a wide range of NLP tasks, including low-resource languages, was introduced [44]. Similarly, adapter modules were used for task-specific additions to pre-trained models [117], allowing efficient fine-tuning without modifying the entire model. The benefits of transfer learning for low-resource languages are profound. Reusing knowledge from high-resource settings reduces the dependency on costly and time-consuming data annotation processes. Additionally, transfer learning promotes inclusivity by extending NLP capabilities to underrepresented languages, fostering linguistic diversity in computational systems. It also supports the development of downstream applications like translation, sentiment analysis, and information retrieval, even for languages with limited resources. In conclusion, transfer learning has become a cornerstone for advancing NLP in low-resource settings.

5.3.4 Zero shot NMT

Zero-shot NMT has become a pivotal method for enabling translation between low-resource language pairs without requiring parallel data. This approach leverages a shared encoder-decoder architecture trained on high-resource language pairs to learn general translation patterns. By extending these learned patterns, the model can translate between unseen language pairs, effectively bridging the gap in low-resource scenarios.

Recent advancements have significantly refined zero-shot NMT methodologies. An unsupervised NMT framework has been introduced [118] that relies on shared subword embeddings and back-translation, demonstrating success in language pairs without direct supervision. A multilingual NMT model was proposed [119] with language-specific tags, enabling the system to perform zero-shot translation by aligning shared latent spaces across multiple languages. Furthermore, innovations in pre-trained multilingual language models such as mBART [120] and MASS [121] have enhanced zero-shot capabilities by incorporating robust sequence-to-sequence pretraining on diverse multilingual datasets.

By eliminating the need for direct parallel corpora, the approach significantly reduces resource constraints, making it feasible to include underrepresented languages in translation systems. It also promotes cross-linguistic inclusivity, enabling the development of tools and applications like multilingual chatbots, cross-cultural communication platforms, and low-cost translation services. Zero-shot NMT additionally fosters linguistic preservation by supporting endangered and regional languages, which might otherwise lack computational resources.

5.3.5 Multilingual Approach

The multilingual approach has established itself as a pioneering approach in NLP, particularly for addressing the challenges of low-resource languages. By training a single model on data from multiple languages, this approach exploits shared linguistic features to improve performance in languages with limited resources. The underlying concept is that similarities between languages, such as grammar, syntax, and vocabulary, can be leveraged through a shared representation space.

A multilingual NMT system uses language-specific tags to enable translation across multiple language pairs, including those without direct parallel corpora [119]. This system demonstrated the power of cross-lingual transfer learning. Further, a robust cross-lingual language model XLM-R was developed that trained on massive multilingual datasets [122]. XLM-R set new benchmarks for low-resource languages in tasks like translation, sentiment analysis, and named entity recognition. The introduction of mBART [120] provided a sequence-to-sequence pre-training framework specifically designed for multilingual applications, enabling fine-tuning for low-resource languages.

The multilingual approach has become a foundation for advancing NLP for low-resource languages. Harnessing the power of cross-lingual representations and transfer learning addresses critical challenges while enabling equitable access to language technology. As research continues to refine these methods, the potential for multilingual models to democratize NLP grows, ensuring that no language is left behind.

5.3.6 Multimodal Approach

The multimodal approach NLP represents an innovative framework integrating multiple data types, such as text, images, audio, and video, to enhance language understanding and generation. This approach is particularly advantageous for low-resource languages, where traditional text-based data may be scarce, but rich contextual information from other modalities can supplement the learning process.

MURAL [123], a multimodal universal representation model, demonstrated how combining text and image data could benefit multilingual translation and cross-lingual tasks. Speech-based multimodal approaches, such as Speech2Vec [124], have utilized audio-text alignment to improve performance in languages with limited textual resources. The models leverage massive multimodal datasets to learn robust cross-modal representations. Furthermore, multimodal NMT systems incorporate visual data during training to enhance the quality of translations in resource-constrained scenarios. It reduces reliance on extensive text-only datasets by incorporating complementary data from other modalities.

In conclusion, the multimodal approach is a promising avenue for advancing NLP for low-resource languages. Integrating diverse data modalities provides a more prosperous and inclusive representation of language. The multimodal methods support the preservation of linguistic and cultural heritage through multimedia archives, benefiting endangered languages.

Mizo Tonal Words in Corpus	**Sentences**
Sentence without Mizo tonal words	77,067
Sentence with Mizo tonal words	44,168
Total sentence	**121,235**

Table 5.2: Mizo Tonal Words in Corpus

5.4 Corpus Preparation

Parallel data and Mizo monolingual data were prepared manually and from various online resources to build a language resource for the English-Mizo corpus. Online resources include Elementary Textbook[1], mCovid-19 websites[2] and movie subtitles, while manually prepared parallel sentences cover the general domain sentences. A dataset of 121,235 parallel sentences has been prepared, including 44,168 Mizo sentences with tonal words to encounter context-specific tonal words of the Mizo language. as shown in Table 5.2.

The parallel corpus contains 118,895 sentences from online sources (98.06%) and manually prepared 2,340 sentences (1.93%) as in Table 5.3. The difference between online parallel sentences and manually prepared sentences is that online parallel sentences include both with and without tonal sentences, whereas manually prepared sentences only include tonal words to enhance the number of parallel sentences with tonal words. The monolingual Mizo data of 2,061,068 sentences is highlighted in Table 5.4. It is prepared from various newspapers, web pages, blogs, and books.

Online/Manual	**Sentences**	**Percentage**
Online	118,895	98.06%
Manually	2,340	1.93%
Total sentence	**121,235**	

Table 5.3: Corpus Statistics

[1] https://scert.mizoram.gov.in/

[2] https://mcovid19.mizoram.gov.in/

Monolingual Data	Sentences
Mz	2,061,068

Table 5.4: Monolingual Data on Mizo Language

Type	Sentences	Tokens	
		En	Mz
Train	118,035	1,468,044	1,314,131
Validation	2,000	55,316	52,320
Test	1,200	11,943	10,168

Table 5.5: Statistics for train, valid and test set

Web crawling[3] techniques are used to collect data from online sources. To allow for replication over several web pages, each element's `xpath` is formatted/encoded with a degree of generalization. It aided in crawling and retrieving information from many web pages. Before splitting a parallel corpus, duplicates and noise (such as web links, excessive special characters, and blank lines) are removed. Further, the dataset is verified by hiring a linguistic expert with linguistic knowledge of both languages. The data statistics of the train, valid, and test set are shown in Table 5.5. During the split, parallel sentences with tonal words are considered for validation and test data. The test and validation set include 98% and 2% sentences from online and manually prepared sentences, and also, the train set includes 1.92% of and 98.07% sentences from manually prepared and online sources. The percentage of tonal words present in the train, validation, and test set are 11.20%, 10.50%, and 10.30%.

5.5 System Description

To build an English-to-Mizo NMT system, as shown in Figure 5.1, the approach consists of three phases:

- Initially, for the first phase, Mz tonal sentences are extracted from the monolingual data of Mz. Then, the extracted Mz tonal sentences are used to

[3] `https://scrapy.org/`

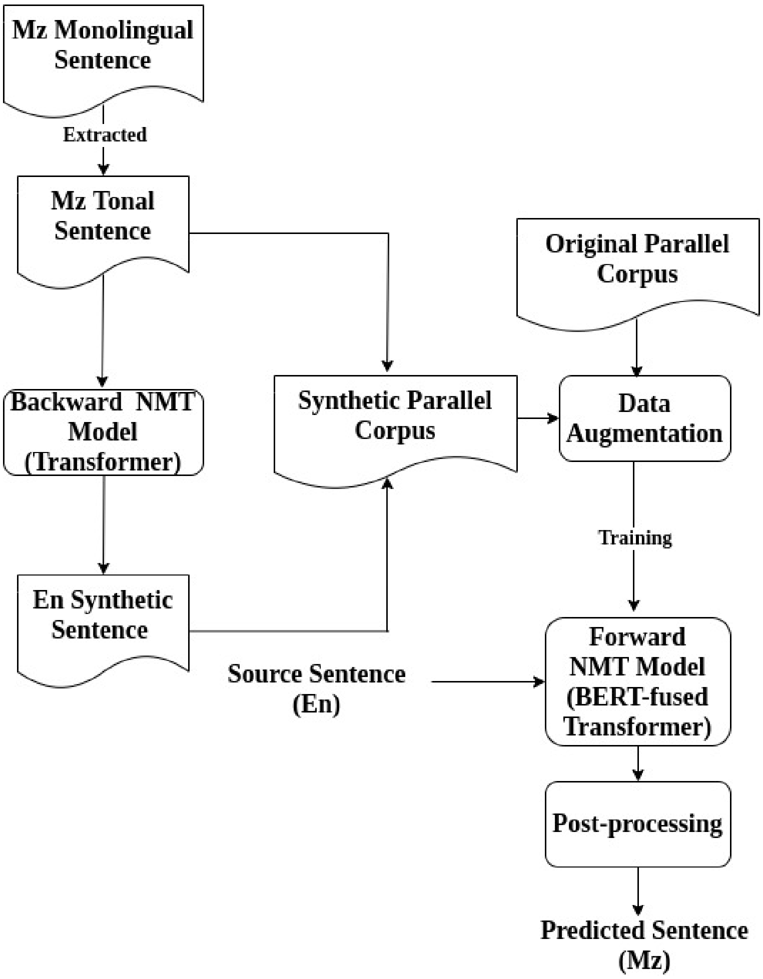


Figure 5.1: English-to-Mizo NMT System

generate En synthetic sentences using the backward NMT model (Mz-to-En). The conventional transformer model [26] was used in this case. Blank lines and under-translated sentences (single or double words) were removed from En synthetic sentences, and the corresponding Mz tonal sentences. Thus, a total of 33,021 synthetic parallel sentences were prepared, as given in Table 5.6.

Sentences	**Tokens**	
	En	**Mz**
33,021	6,08,586	5,49,822

Table 5.6: Synthetic parallel data statistics

- The synthetic parallel corpus is augmented with the original parallel corpus in the second phase. Then, the technique of [125] was followed by augmenting the swapped sentences (Mz-to-En). Artificial tokens were added at the beginning of the source sentences to recognize the target sentences (such as <2mz> for Mizo and <2en> for English target sentences) and trained with BERT-fused NMT [126] for the forward (En-to-Mz) translation. BERT-fused NMT is utilized to leverage the pre-trained English model. Different configurations were investigated, namely, unidirectional and bidirectional parallel corpus (trained on En-to-Mz and Mz-to-En simultaneously). BERT processes an input sequence by first transforming it into representations. Through the BERT-encoder attention module, each NMT encoder layer processes each of the representations from the BERT module. Besides, each NMT encoder layer's self-attention continues to process the previous NMT encoder layer's representations. Finally, it generates fused representations through the encoder layers feed-forward network by merging both the output of BERT-encoder attention and self-attention. The decoder works similarly; the BERT-decoder attention is introduced to each NMT decoder layer. The obtained trained model is used to predict the target sentences.

- Lastly, an example-based post-processing step is proposed to improve the translation accuracy of encountering tonal words.

  Example-based post-processing: An example-based dictionary was created using the following steps for the post-processing step.

  - Keywords containing tonal words were extracted from monolingual data of Mizo using a language-independent keyword extraction tool known as YAKE [127], considering maximum n-gram size= 3.
  - The uni-gram words were discarded from the extracted keywords because they cannot represent the context-specific tonal words.
  - An example-based dictionary ($K_z||K_y$) was created. Here, K_y denotes extracted keywords, and K_z is prepared by removing the tonal markers from K_y.

  The example-based dictionary is utilized for the post-processing of the predicted sentences. Each keyword of K_z was searched in the predicted sentences, and if it is found, it is replaced with the keyword of K_y. The reason behind using the post-processing step is that if the trained model cannot capture the appropriate tone marker in the translation process, then the post-processing step attempts to correct the concerned tone marker using an example-based dictionary. An example-based dictionary was used because the tonal word is contextually dependent on the pre or post-word of the concerned tonal word. The proposed approach is based on the BERT-fused NMT (transformer model), bidirectional data augmentation with synthetic parallel corpus, and an example-based post-processing step.

5.6 Experiment and Result Analysis

Preliminary experiments were performed for both En-to-Mz and Mz-to-En translations using RNN [15] and transformer model [26] with sub-word segmentation technique i.e., BPE. The preliminary experiment results are reported in Table 5.7.

Translation	**Model**	**BLEU**
En-to-Mz	RNN	16.98
	Transformer	17.86
Mz-to-En	RNN	15.46
	Transformer	16.52

Table 5.7: BLEU scores of preliminary experiments

The result shows that for both translations, i.e., En-to-Mz and Mz-to-En, the Transformer model performs better than the RNN model.

Eight different models for En-to-Mz translations have been investigated. They are:

- **M1 :** Transformer Model
- **M2 :** Bert-fused Transformer Model
- **M3 :** Bert-fused Transformer Model with post-processing step
- **M4 :** Bert-fused Transformer Model with synthetic parallel corpus
- **M5 :** Bert-fused Transformer Model with synthetic parallel corpus and post-processing step
- **M6 :** Bert-fused Transformer Model with bidirectional parallel corpus
- **M7 :** Bert-fused Transformer Model with bidirectional parallel corpus and synthetic parallel corpus
- **M8 :** Bert-fused Transformer Model with bidirectional parallel corpus and synthetic parallel corpus and post-processing step

The quantitative results of all the models are evaluated in terms of automatic evaluation metric, BLEU[4] [59] and HE [16] on randomly selected 100 sample sentences by hiring a linguistic expert. The default configurations of the OpenNMT-py[5] toolkit were followed to implement the RNN and transformer models. The Adam optimizer

[4] Utilized multi-bleu.perl script

[5] `https://github.com/OpenNMT/OpenNMT-py`

Model	BLEU
M1 (UPC)	17.86
M2 (UPC)	18.39
M2 + PP (M3)	21.90
M2 + SPC (M4)	20.55
M4 + PP (M5)	23.82
M2 (BPC) (M6)	22.80
M6 + SPC (M7)	24.33
M7 + PP (M8)	**28.59**

Table 5.8: Comparative results (BLEU scores) of different models for En-to-Mz translation, M1: Transformer, M2: BERT-fused Transformer, SPC: Synthetic Parallel Corpus, PP: Post-processing, UPC: Unidirectional Parallel Corpus, BPC: Bidirectional Parallel Corpus

with a learning rate of 0.001 and drop-outs of 0.3 (for RNN) and 0.1 (for transformer) were used in the training process. Additionally, the default configurations of the Fairseq[6] toolkit were followed to implement BERT-fused NMT [126]. Table 5.8 presents the results for all eight models of BLEU for En-to-Mz translation. Likewise, Table 5.9 reports the comparative results of the HE for all eight models for En-to-Mz translation. From both Table 5.8 and Table 5.9, it shows that the proposed approach **(M8)** attains the best score.

Model	Adequacy	Fluency	Overall Rating
M1	2.58	2.76	2.67
M2	3.40	3.92	3.66
M3	3.76	4.54	4.15
M4	3.26	4.47	3.86
M5	3.92	4.68	4.30
M6	3.65	4.52	4.08
M7	3.32	4.64	3.98
M8	4.14	5.24	4.69

Table 5.9: Human evaluation scores of different models for En-to-Mz translation

To examine the effectiveness of the proposed approach **M8** in terms of encountering tonal words, a comparative analysis is presented in Figure 5.2. Although, the proposed approach **M8** encounters a higher frequency of tonal words than conventional transformer [26] and BERT-fused transformer [126] models, but much less than the frequency of tonal words in reference test sentences.

[6] `https://github.com/bert-nmt/bert-nmt`

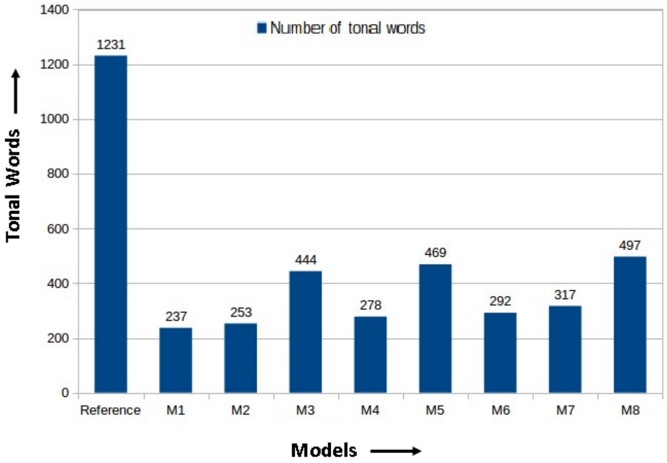


Figure 5.2: Comparative analysis on tonal frequency of words. Reference: Mizo target sentences (test data)

To provide a detailed analysis of the performance of the eight different models, an example output sentence is presented below:

Output Sentence:

English : It is nice

Mizo : **A thà lutùk.**

M1 : A tha lutuk.

Discussion: Model M1 produces the exact output as the reference, but the tone markers *à* and *ù* are missing.

M2 : A tha lutuk.

Discussion: Model M2 produces the exact output as the reference, but the tone markers *à* and *ù* are missing.

M3 : A tha lutùk.

Discussion: Model M3 produces an output as the reference sentence, but the tonal word in *'thà'* is not generated

M4 : A thà khawp mai.

Discussion: Model M4 produces an output that is not exact with the reference sentence but has the same meaning. The tone maker in the word *'mai'* is missing.

M5 : A thà khawp mài.

Discussion: Model M5 produces an output that is not exact with the reference sentence but has the same meaning. Here, the tonal words are correctly generated using the tone marker.

M6 : A tha lutuk.

Discussion: Model M6 produces the exact output as the reference, but the tone markers *à* and *ù* are missing.

M7 : A tha lutuk.

Discussion: Model M7 produces the exact output as the reference, but the tone markers *à* and *ù* are missing.

M8 : A thà lutùk.

Discussion: The proposed approach, Model M8, produces the exact output as the reference with the tonal words correctly generated.

Further, a few examples are presented in Table 5.10 to inspect qualitative analysis of encountering tonal words. It is observed that the conventional transformer (M1) and BERT-fused transformer (M2) models cannot encounter tone markers in the tonal words of the predicted sentences. However, with the post-processing approach, M3, M5, and M8 generate tonal words with appropriate tonal markers marked as 'bold.' By capturing tone markers in tonal words, the proposed approach **M8** significantly represents the contextual meaning of the sentences compared to other models.

5.7 Conclusion

This work aims to prepare an Indian language resource, i.e., English-Mizo corpus, and investigate En-to-Mz translation by encountering tonal words and exploring different NMT models on the developed dataset. The proposed approach is based on BERT-fused NMT with bidirectional data augmentation with synthetic parallel corpus. Since one of the main challenges in the Mizo language is tackling tonal words, post-processing steps are employed to manage the tonal words as they add to the

Source / Target	**Predicted**
	A tha lutuk. (M1)
	A tha lutuk. (M2)
	A tha lutùk. (M3)
It is nice. (En)	A thà khawp mai. (M4)
A thà lutùk. (Mz)	**A thà khawp mài.**(M5)
	A tha lutuk. (M6)
	A tha lutuk. (M7)
	A thà lutùk.(M8)
	Dawt sawi suh. (M1)
	Dawt sawi duh suh. (M2)
	Dáwt sáwi duh suh. (M3)
Don't tell lie. (En)	Dawt sawi suh. (M4)
Dáwt sáwi suh. (Mz)	**Dáwt sáwi suh.** (M5)
	Dáwt sawi suh. (M6)
	Dáwt sawi suh.(M7)
	Dáwt sáwi suh. (M8)
	Ka lokal ang. (M1)
	Ka lokal dawn nia. (M2)
	Ka lòkal dawn. (M3)
I'll be there. (En)	Ka kal dawn nia. (M4)
Ka lòkál áng. (Mz)	**Ka lokál áng e.** (M5)
	Ka lokál ang. (M6)
	Ka lokál ang.(M7)
	Ka lòkál áng. (M8)

Table 5.10: Analysis of different models for En-to-Mz translation

complexity on top of low-resource challenges for any NLP task. Better translation accuracy was achieved compared to a conventional transformer and BERT-fused NMT. This method enhances translation accuracy by effectively addressing tonal words in Mizo using a post-processing step, resulting in state-of-the-art outcomes in English-to-Mizo translation.

Chapter 6

Mizo Visual Genome 1.0: A Dataset for English↔Mizo Multimodal NMT

6.1 Introduction

Deep Learning offers diverse applications and has obtained state-of-the-art for different work during the last few years. NLP tasks, speech recognition, image classification, and other modalities are examples of deep learning applications. Modality refers to the representation format in which a specific data type is represented, such as audio, visual, textual, etc. Therefore, multimodal representation refers to the combination of several modalities. Though merging several modalities to improve performance is intuitively desirable, incorporating conflicts across modalities is complicated. However, for language translation, the multimodal approach surpasses textual translation as it integrates the features of an image with parallel textual data. Visual elements in MT may aid in the resolution of unclear words.

The multimodal method also improves low-resource pair translation, making NMT an effective analysis for both high and low-resource language pairs. However, developing multimodal MT requires an appropriate dataset, leading to a significant challenge for a low-resource language pair. Since no prior study has addressed the English↔Mizo MMT system, a multimodal translation is proposed for a low-resource English↔Mizo language pair to improve the translation with additional modalities. Although text-only parallel corpora for En-Mz translation are available [128, 90, 129], there was no multimodal dataset for Mizo. Therefore, the main objective is to implement a Mizo MMT system.

6.2 Multimodal NMT

MNMT enhances translation quality by integrating multiple data modalities, such as text, images, and audio. This approach leverages the complementary information from various sources, allowing models to capture richer semantic meanings and contextual nuances. By utilizing parallel datasets that pair text with relevant images or audio, MNMT can improve translation understanding and accuracy. Despite its potential, MNMT faces challenges, including data scarcity, especially for low-resource languages, handling code-switching, and developing effective alignment strategies between different modalities. By leveraging multiple modalities, the multimodal approach bridges language gaps and enhances user experience, making it a promising direction for future translation technologies.

The MNMT collects information from various modalities, including textual and visual sources. The MNMT model is an extension of the attention-based NMT model that includes spatial visual elements through the addition of a visual component[77, 130]. It comprises three components, including images in the attention-based NMT framework. These are:

6.2.1 Feature extraction from images

The extraction of local and global features from the given image dataset is performed using pre-trained CNN with VGG19 for multimodal translation [131]. The pre-trained 19-layer VGG network (VGG19- CNN) extracts visual features from the dataset and includes them in initializing the encoder and decoder hidden state.

6.2.2 Initialising Encoder

A BRNN with GRU [20] is employed to encode data. Instead of utilizing the zero vector $\overrightarrow{0}$ to initialize the hidden state of the encoder, two new feed-forward single-layer neural networks are used to construct the forward RNN and the backward

RNN, respectively. Equation (6.1) creates a vector d from the global image feature vector.

$$d = PM_I^2.\left(PM_I^1.q + v_I^1\right) + v_I^2 \tag{6.1}$$

Here, PM represents the projection matrix, and v represents the bias vector. PM_I^2 and v_I^2 project the picture characteristics into the exact dimensions as the source language encoder's hidden states. The following feed-forward networks are computed to initialize the encoder's hidden state:

$$\overrightarrow{s}_{init} = tanh(PM_f d + v_f) \tag{6.2}$$

$$\overleftarrow{s}_{init} = tanh(PM_b d + v_b) \tag{6.3}$$

Here, the forward and backward directions are shown by the subscripts f and b of PM and v, respectively.

6.2.3 Initialising Decoder

A doubly attentive RNN was utilized to integrate an image into the decoder. The decoder computes three essential components at each stage [77]:

- First, the previously hidden state and the previously emitted word are used to calculate the hidden states.
- Second, time-dependent source vectors are computed by the source-language RNN using an attention mechanism over its hidden states.
- Third, the time-dependent source context vector and hidden state proposal are used to determine the final hidden states.

Equations (6.4) and (6.5) demonstrate how a feed-forward single-layer network is used to calculate an expected alignment (e) between each input vector (hi) and the expected word to be produced at the current time step (t):

$$e_{t,i}^{src} = (v_a^{src})^T \, tanh \, (U_a^{src} \acute{s}_t + m_a^{src} s_i) \tag{6.4}$$

$$a_{t,i}^{src} = \frac{exp \left(e_{t,i}^{src} \right)}{\sum_{j=1}^{N} exp \left(e_{t,j}^{src} \right)} \tag{6.5}$$

here, $a_{t,i}^{src}$ indicate the alignment matrix between each input vector s_i and the expected word to be produced at time step t, and v_a^{src}, U_a^{src} and m_a^{src} are model parameters.

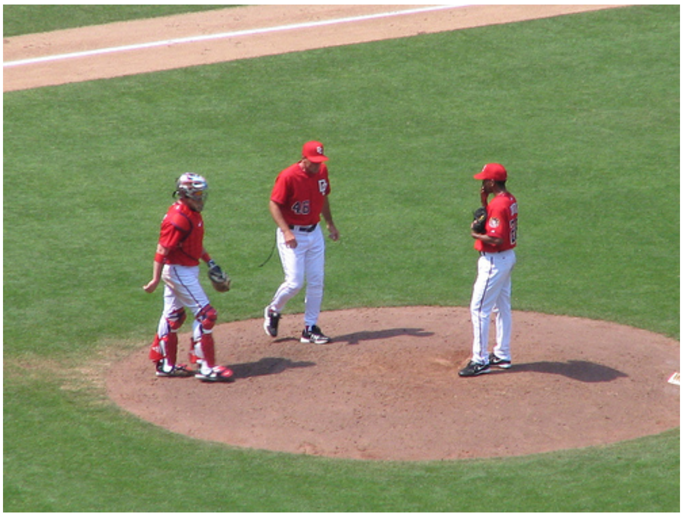
Finally, a source context vector with time-dependent cv_t is calculated as a sum of the weighted input vectors with the attention weight applied to each vector.$a_{t,i}^{src}$ as in Equation (6.6)

$$cv_t = \sum_{i=1}^{N} a_{t,i}^{src} s_i \tag{6.6}$$

6.3 Mizo Visual Genome 1.0 (MVG 1.0)

A multimodal Mizo Visual Genome 1.0 (MVG 1.0) corpus is created for the En-Mz language pair. The MMT system employs the text and the corresponding image to create the model that will be translated into the target language. Fig. 6.1 shows the MVG 1.0 comprises bilingual picture captions in the image. Generating a multimodal MVG corpus involves creating a comprehensive dataset that pairs Mizo language annotations with relevant images. The corpus is a foundational resource for training multimodal MT models and enhancing the understanding of contextual nuances in Mizo.

The experiment uses Hindi Visual Genome 1.1 [132] to create MVG 1.0 multimodal corpus. Hindi Visual Genome 1.1 is an improved version of Hindi Visual Genome 1.0 [133]. The Hindi Visual Genome comprises an image with corresponding English captions translated into Hindi captions. To construct the MVG 1.0



English Text: *Three baseball players are in the field*
Mizo Text : *Baseball player pathum tualzawlah an awm.*

Figure 6.1: An example of MVG 1.0 dataset: a picture with its caption in English and Mizo.

corpus, a variety of practical resources were examined, which are discussed in the following points:

- Train data in Hindi Visual Genome 1.1 includes 28,928 images and 28,930 captions. Three redundant captions (id:2385507, 2391240, and 2328549) were eliminated, along with one particular image having an ID of 232683, as its corresponding English caption is inaccessible. As a result, there are 28,927 images and English captions in the train set. The source and target sentences' training, validation, and test data were verified and corrected, as shown in Table 6.1. This table offers an overview of the multimodal corpus statistics for MVG 1.0.

Set	Data type	Instances	En-Mz (Token)
Train	Text/Image	28,927	148,146 / 172,460
Validation	Text (En-Mz)	998	5,098 / 5,519
	Image (En-Mz)	998	
Test	Text (En-Mz)	1,400	8,192 / 10,337
	Image (En-Mz)	1,400	

Table 6.1: Multimodal Corpus Statistics of MVG 1.0

- In May 2022, the Mizo language was added to Google Translate, which is then utilized as an automatic translator of English to the Mizo language. By utilizing Google Translate[1], the English captions of Hindi Visual Genome are translated into Mizo. However, Google Translate provides extremely inaccurate output, resulting in numerous errors. 97 % of the sentences required human correction. Therefore, linguistic experts manually verify and post-edit the data. The errors observed in the output of Google Translate include vocabulary issues, wrong word order, incorrect names of colors, etc. A few examples of Google Translate output with correction are shown in Table 6.2. Table 6.3 provides a few instances of parallel datasets.

Set	**English**	**Mizo (Google Translate Output)**	**Mizo (Mannually corrected)**
Train	leg of the chair	ke chu chair-ah a ding a	Chair ke ani
	man has brown hair	mipa chuan sam dum a nei a	mipa chuan sam uk a nei a
	woman wearing red scarf	hmeichhia scarf sen ha	hmeichhia scarf sen awrh
Validation	a car loading food on a plane	thlawhna chhunga ei tur phur car pakhat	thlawhna chhungah carin eitur a dah
	the man standing near the motorcycle	motor bula ding pa chu	motorcycle bulah pa pakhat a ding
	White short sleeved tshirt	Tshirt kawrfual dum a ni	Tshirt var ban bul
Test	white TV stand on floor	TV dum ding chu leiah a awm a	TV dahna var chu leiah a awm a
	charger stand on a desk	charger chu desk chungah a ding a	dawhkan chungah charger stand a awm
	second floor windows	second floor window a awm bawk	Chhawng hnihna tukverh

Table 6.2: Google Translate output with correction

[1] https://translate.google.co.in/

Set	English	Mizo
Train	wall is painted white	Bang chu rawng var a in hnawih
	Person crossing street with umbrella	Mi pakhatin nihliap kengin kawng a kan
	A plate full of food	Thleng eitur a khat
Validation	piece of sandwich stuff with lot of meat	sandwich pakhat a sa tam tak awmna
	a bunch of apples	apple bawr khat
	The dog has long fur	Ui hian hmul sei tak a nei a
Test	tail of cat forms a C	zawhte mei chuan C a siam a
	A display of different phone models	Phone chi hrang hrang phochhuah a ni
	a group of people watching the baseball players	mipui tam takin basketball players te an en

Table 6.3: Example of Parallel Train, Valid and Test Data

6.4 System Description

The multimodal translation system developed for the En-Mz language pair is described in detail. The En-Mz MNMT system is presented in Fig.6.2. OpenNMT-py [134] is used to set up both MNMT and text-only NMT systems. The major procedures used to independently develop the systems for En-to-Mz and Mz-to-En translation are preprocessing the data, system training, and system testing.

The image and text data are preprocessed separately by pre-trained CNN using the VGG19 model [131], publicly available in OpenNMT-py. Pre-trained CNN with VGG19 extracts local and global features from the image dataset for multimodal translation. The source-target sentences with a vocabulary size of 6991 for English and 8443 for Mizo are generated using the OpenNMT-py toolkit. During system training, the parallel source-target text data are used to fine-tune the retrieved features acquired from the preprocessing stage. An RNN encoder was employed along with a doubly-attentive RNN decoder[21]. Additionally, the BRNN encoder was considered for comparison purposes. The model was trained for 40 epochs on a single GPU using a 2-layer, 16 batch size, Adam optimizer, learning rate of 0.001,

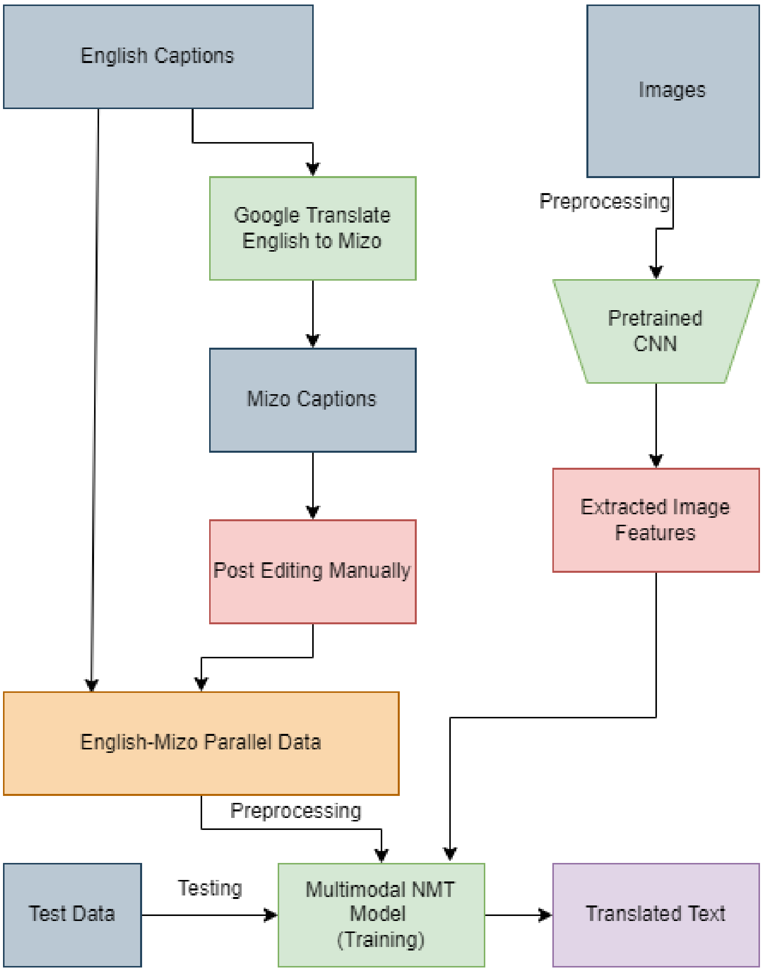


Figure 6.2: English↔Mizo MNMT System

and drop-outs of 0.3. The system is then tested using the best model developed during training. The obtained trained models translate the provided test data for both multimodal and text-only NMT systems separately.

6.5 Experiment and Result Analysis

For the experiment, two types of models—BRNN and RNN—are employed [21]. To evaluate the translation quality of the predicted sentences, automatic evaluation metrics viz. BLEU [59], RIBES (rank-based intuitive bilingual evaluation scores) [135], METEOR [60], TER [107], and F-measure scores are considered. Higher scores for BLEU, METEOR, RIBES, and F-measures indicate better accuracy, whereas the lower scores indicate better results for the TER score. Table 6.4 provides automatic evaluation scores on MNMT and text-only NMT systems for both forward (En-to-Mz) and backward (Mz-to-En) translation. The BRNN MNMT system performs better than the RNN MNMT system, as indicated in Table 6.4. Additionally, the MNMT system produces better output than the NMT method, which solely uses text since MNMT utilizes both textual and visual components.

Set	Translation	Model	BLEU	TER	RIBES	METEOR	F-measure
Multimodal	En-to-Mz	RNN	7.10	81.5	0.428798	0.141688	0.337810
		BRNN	7.39	81.3	0.428816	0.141711	0.337855
	Mz-to-En	RNN	9.85	78.7	0.429210	0.179498	0.404810
		BRNN	10.03	78.5	0.429255	0.179529	0.404847
Text-only	En-to-Mz	RNN	6.02	82.9	0.405535	0.133310	0.323508
		BRNN	6.18	82.8	0.405563	0.133328	0.323998
	Mz-to-En	RNN	8.20	78.9	0.421498	0.176710	0.399064
		BRNN	8.84	78.7	0.421517	0.176739	0.399092

Table 6.4: Automatic Evaluation Scores of En-Mz Multimodal and Text-only NMT Systems

Using the BRNN model, the best and worst examples of MNMT predicted sentences are shown in Fig. 6.3 and 6.4, respectively. Google translation output is also considered for comparative analysis. The predicted sentences are evaluated below to investigate the efficiency of the MNMT system.

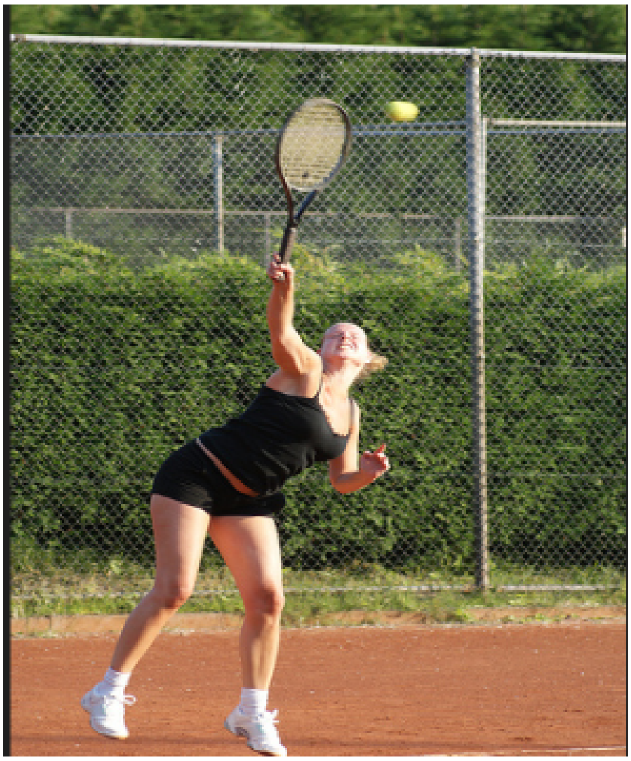
<table>
<tr><th colspan="2">Image ID: 2327005</th></tr>
<tr><td colspan="2"></td></tr>
<tr><td><u>En-to-Mz</u>
<b>Source Sentence</b>: small patch of grass beside a tennis court
<b>Reference Sentence</b>: tennis court bulah hnim tlem te a awm</td><td><u>Multimodal Translation</u>
<b>Predicted Sentence</b>: tennis court bula hnim tlemte awmna
<u>Text-only translation</u>
<b>Predicted Sentence</b>: tennis court bulah chuan grass patch te tak te a awm
<b>Google Translation</b>: tennis court bulah thinghnah te tak te a awm</td></tr>
<tr><td><u>Mz-to-En</u>
<b>Source Sentence</b>: Tennis court bulah hnim tlem te a awm
<b>Reference Sentence</b>: small patch of grass beside a tennis court</td><td><u>Multimodal Translation</u>
<b>Predicted Sentence</b>: small patch of grass on the tennis court
<u>Text-only translation</u>
<b>Predicted Sentence</b>: small round animals in a tennis court by a tennis court
<b>Google Translation:</b> a little sand near the tennis court</td></tr>
</table>

Figure 6.3: Examples of MNMT's Best Predicted Output with Text-only NMT

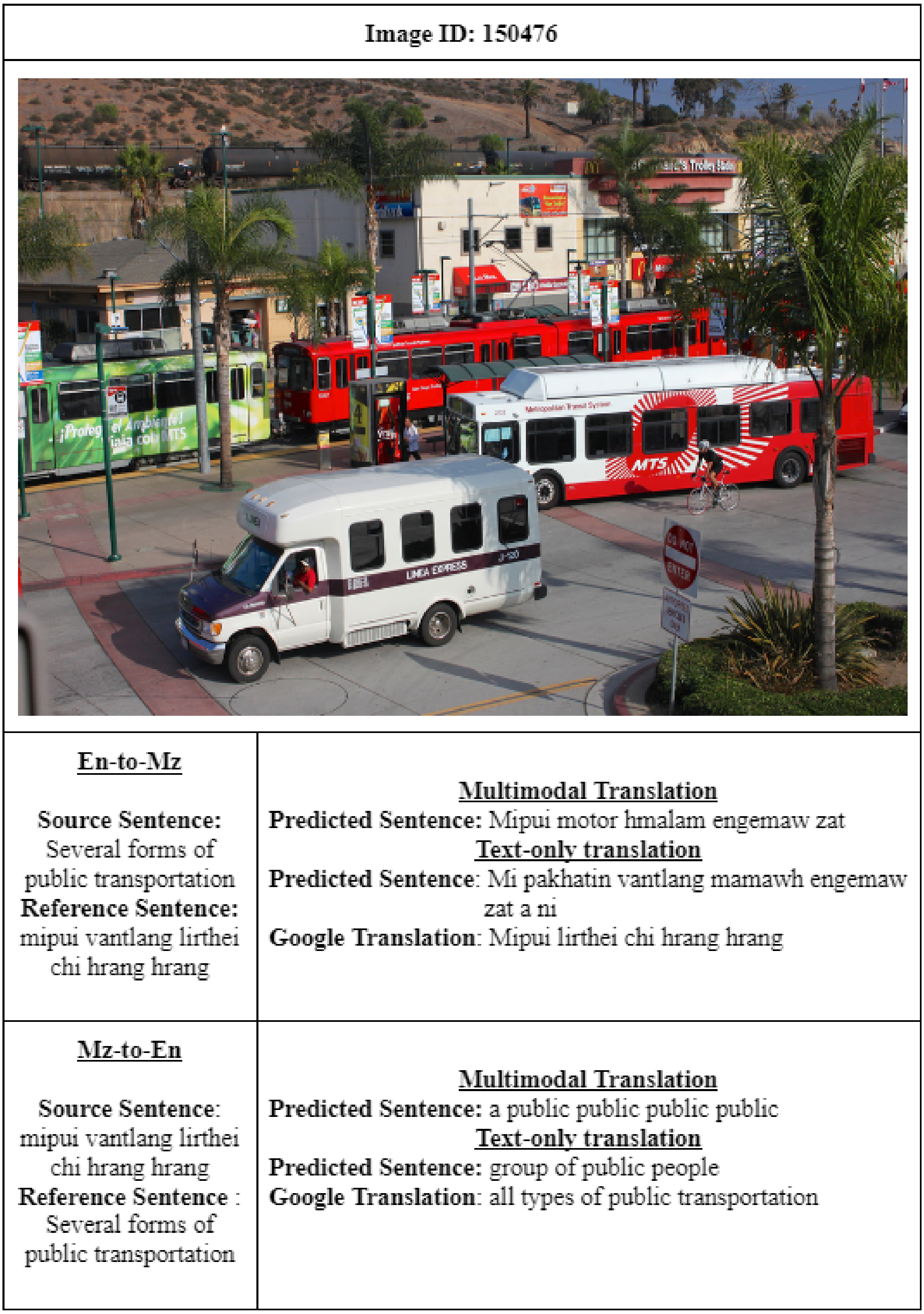


Image ID: 150476	
<u>**En-to-Mz**</u> **Source Sentence:** Several forms of public transportation **Reference Sentence:** mipui vantlang lirthei chi hrang hrang	<u>**Multimodal Translation**</u> **Predicted Sentence:** Mipui motor hmalam engemaw zat <u>**Text-only translation**</u> **Predicted Sentence**: Mi pakhatin vantlang mamawh engemaw zat a ni **Google Translation**: Mipui lirthei chi hrang hrang
<u>**Mz-to-En**</u> **Source Sentence**: mipui vantlang lirthei chi hrang hrang **Reference Sentence** : Several forms of public transportation	<u>**Multimodal Translation**</u> **Predicted Sentence:** a public public public public <u>**Text-only translation**</u> **Predicted Sentence:** group of public people **Google Translation**: all types of public transportation

Figure 6.4: Examples of MNMT's Worst Predicted Output with Text-only NMT

Evaluation of Fig. 6.3 of MNMT's Best Predicted Output

for En-to-Mz translation

- MNMT predicted sentence and reference sentence are contextually similar.
- As for text-only NMT translation, even though the translation is correct, words like 'grass patch' are not translated into the Mizo language.
- Google Translate translates the word 'grass' as 'thinghnah', meaning 'leaves'.

for Mz-to-En translation

- The predicted sentence of MNMT is acceptable, apart from the word 'beside' being predicted as 'on'.
- In the case of text-only NMT, the predicted sentence is poorly structured compared to the reference sentence,
- The translation in Google Translate is acceptable. Here, 'grass' is translated as 'sand'.

Evaluation of Fig. 6.4 of MNMT's Worst Predicted Output

for En-to-Mz translation

- The predicted sentence of both MNMT and text-only NMT are semantically distinct from the reference sentence and poorly structured.
- The Google Translate is acceptable

for Mz-to-En translation

- The predicted sentence of MNMT is meaningless with multiple similar word predictions.
- The text-only NMT predicts a meaningful sentence structure but contextually different concerning the reference sentence. Here, 'several forms' and 'transportation' are predicted as 'group' and 'people' respectively.
- The Google Translate is acceptable

Furthermore, Figs. 6.5 and 6.6 display the two average case examples of MNMT-predicted sentences. Regarding Fig. 6.5, in the En-to-Mz translation, the MNMT predicted sentence and reference sentence are contextually similar, but the word ‘blue’ is missing in the MNMT predicted sentence. For text-only NMT translation, the translation is partially correct, but the word ‘bang pawl’ in the source sentence is not encountered correctly. Google Translate translates it correctly, except the word ‘blue’ is translated as ‘dum’, meaning ‘black’ in the translated text. In Mz-to-En translation. the MNMT predicted sentence and reference sentence are contextually similar. For text-only NMT, although the translation is not contextually correct, it is acceptable. Google Translate wrongly translates some words ‘blue wall’ as ‘group of walls’.

As for Fig. 6.6, the MNMT predicted sentence and reference sentence are contextually similar in En-to-Mz translation. The translation for text-only NMT is partially correct. Google Translate translates it similarly to the reference sentence. In Mz-to-En translation. the MNMT predicted sentence and reference sentence are contextually similar. For text-only NMT, the translation is partially correct. As for Google Translate, the translated sentence and reference sentence are contextually similar.

6.6 Conclusion

The exploration of MMT for English↔Mizo has shown that integrating textual and visual data significantly improves translation accuracy and fluency. MVG 1.0, a multimodal corpus, has been developed to create a baseline MNMT (En-Mz) system. Automatic evaluation measures are employed to assess the system’s outputs. The multimodal approach demonstrates better translation quality compared to text-only NMT. The dataset size will be increased for subsequent work, and more experiments will be performed to enhance translation quality.

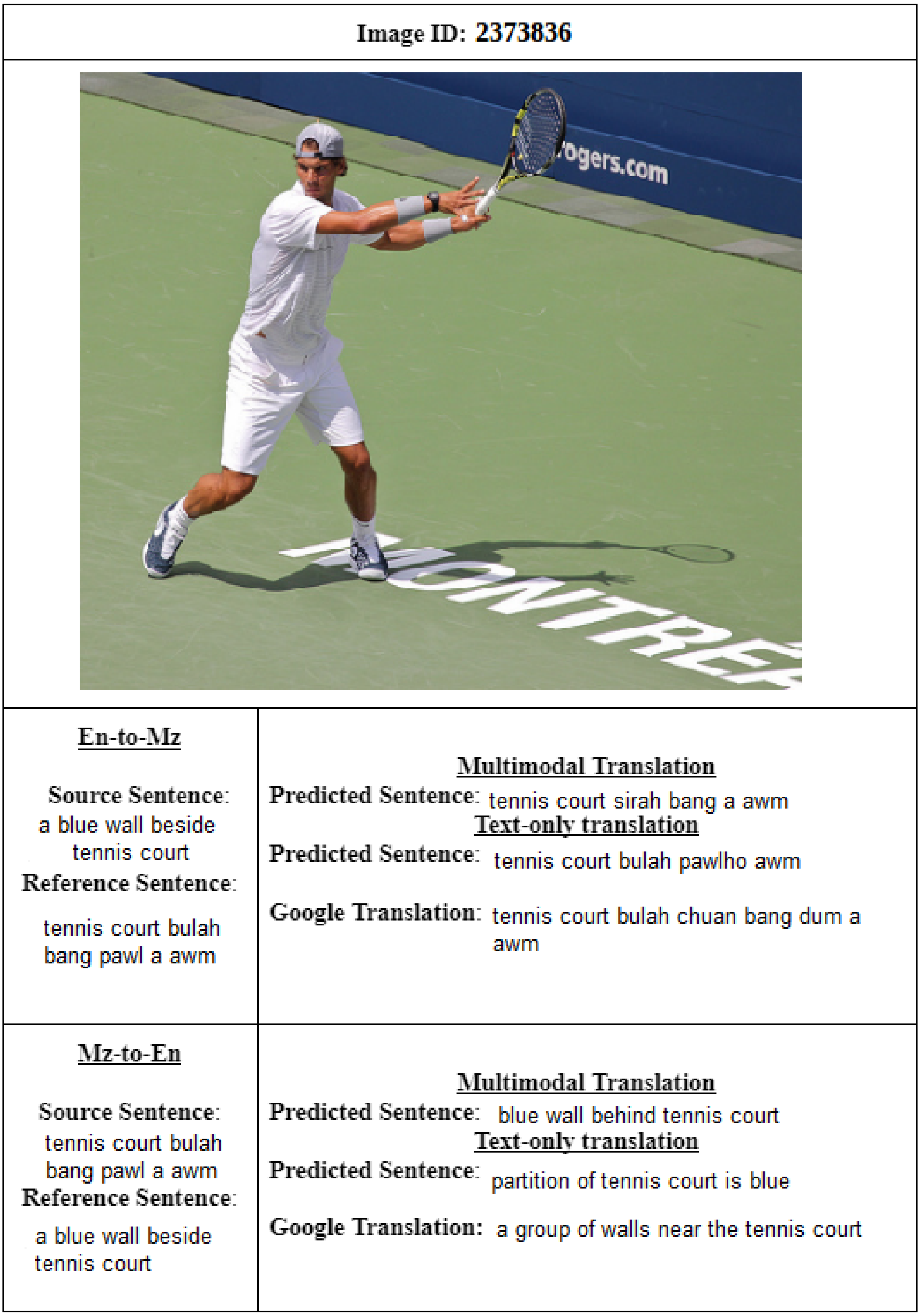
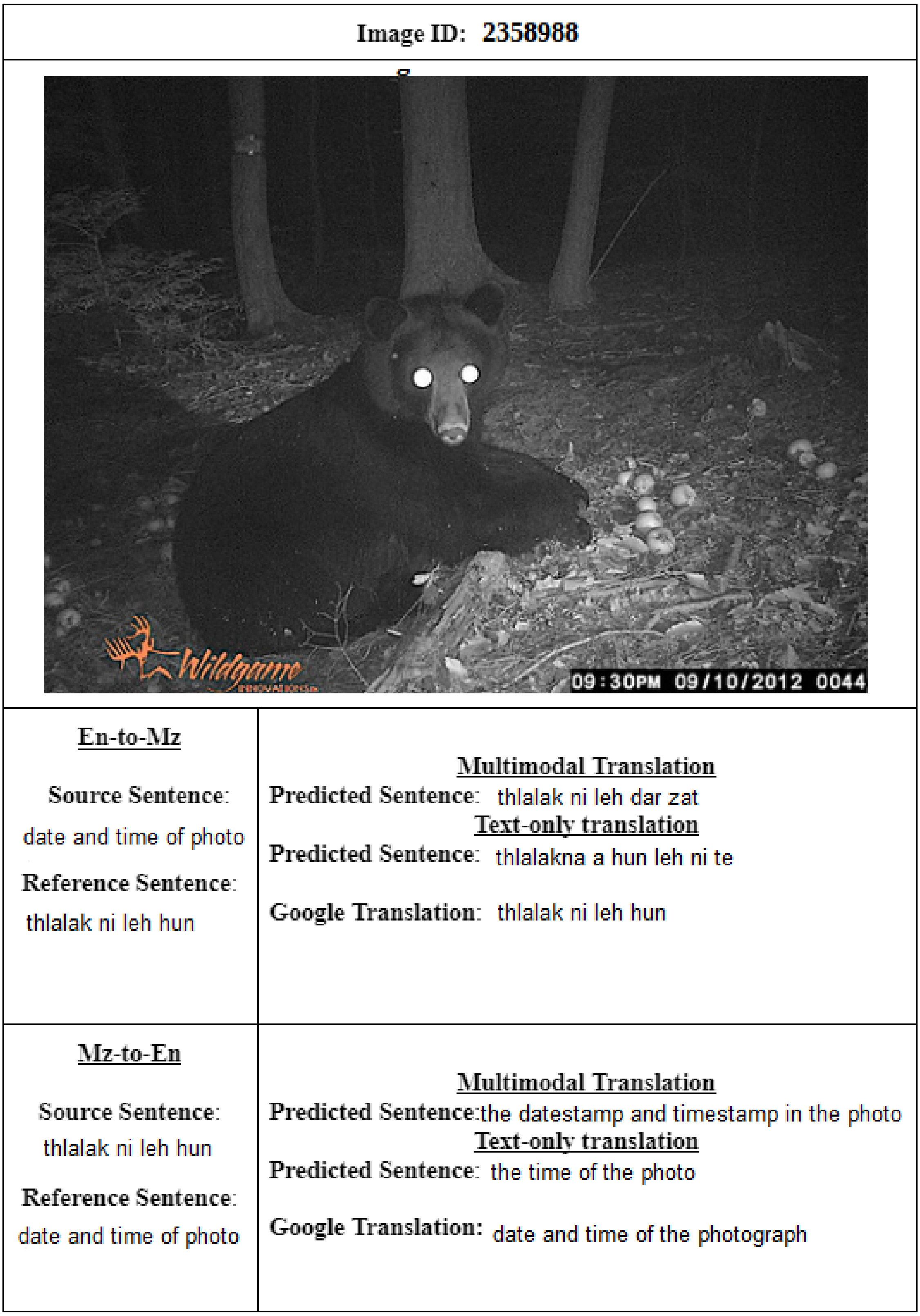


Image ID: 2373836	
En-to-Mz **Source Sentence**: a blue wall beside tennis court **Reference Sentence**: tennis court bulah bang pawl a awm	**Multimodal Translation** **Predicted Sentence**: tennis court sirah bang a awm **Text-only translation** **Predicted Sentence**: tennis court bulah pawlho awm **Google Translation**: tennis court bulah chuan bang dum a awm
Mz-to-En **Source Sentence**: tennis court bulah bang pawl a awm **Reference Sentence**: a blue wall beside tennis court	**Multimodal Translation** **Predicted Sentence**: blue wall behind tennis court **Text-only translation** **Predicted Sentence**: partition of tennis court is blue **Google Translation:** a group of walls near the tennis court

Figure 6.5: Example 1 of MNMT's Average Predicted Output with Text-only NMT



Wildgame
INNOVATIONS
09:30PM 09/10/2012 0044

Image ID: 2358988	
En-to-Mz **Source Sentence:** date and time of photo **Reference Sentence:** thlalak ni leh hun	**Multimodal Translation** **Predicted Sentence:** thlalak ni leh dar zat **Text-only translation** **Predicted Sentence:** thlalakna a hun leh ni te **Google Translation:** thlalak ni leh hun
Mz-to-En **Source Sentence:** thlalak ni leh hun **Reference Sentence:** date and time of photo	**Multimodal Translation** **Predicted Sentence:**the datestamp and timestamp in the photo **Text-only translation** **Predicted Sentence:** the time of the photo **Google Translation:** date and time of the photograph

Figure 6.6: Example 2 of MNMT's Average Predicted Output with Text-only NMT

Chapter 7

English↔Mizo NMT Using Language Model and Addressing Data Scarcity Problem

7.1 Introduction

NMT has emerged as a promising technique due to its context analysis ability and addresses issues with long-range dependencies [15] [26]. It attains state-of-the-art performance and has made significant progress [17]. However, it requires a substantial quantity of training data, which is a massive challenge for low-resource language pair [54]. With Mizo being the low-resource tonal language, several challenges must be addressed for MT. To encounter tonal words in English↔Mizo language pair, the NMT approach has yielded remarkable success[136]. Apart from the tonal words, the Mizo language has challenges in various fields, data scarcity issues, and linguistic divergence. It is essential to address the data scarcity and word-order linguistic divergence issues. To explore the language in the domain of the MT system, analyzing the linguistic characteristics of the language, followed by addressing the linguistic challenges when building the corpus, is highly suggested. The data scarcity challenge has been addressed by utilizing the IBT strategy [113], phrase-pairs extraction [137, 138] and pre-trained language model (LM) [139] to improve the translational performance of low-resource English↔Mizo NMT.

7.2 Transformer Based NMT

The Transformer model utilizes an encoder-decoder architecture that resembles the RNN models. It is the first model to construct representations of its input and output only via self-attention, without convolution or sequence-aligned RNN models.

It intends to address the long-term dependencies and limitations of parallelizing in RNN models. The transformer model's concept is to encode each location and to build entirely on attention mechanisms to link two separate words, which would then be parallelized to speed up learning. Unlike the classic attention mechanism, the self-attention mechanism calculates attention several times, known as multi-head attention. The encoder and decoder are formed by six (6) identical attention layers stacked on top of one another, as shown in Figure 7.1. The encoder comprises two sub-layers: the multi-head self-attention layer and a fully connected position-wise feed-forward network layer. There are three sub-layers in the decoder. Two of the three sub-layers are the same as those in the encoder. Another multi-head attention layer is utilized in the third sub-layer to focus on the encoder stack's output as in Figure 7.2.

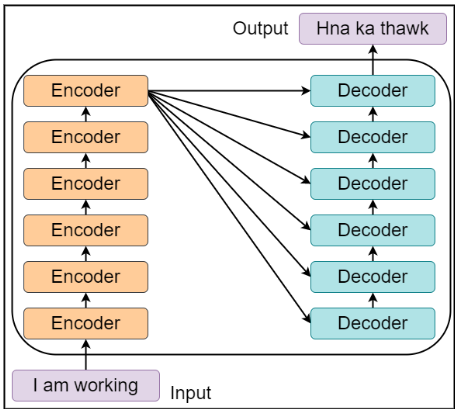


Figure 7.1: Transformer Architecture Model

The mathematical framework of the attention in the Transformer model is determined in Equation 7.1 using a Query (QY), Value (V), and Key (K) with d_k as

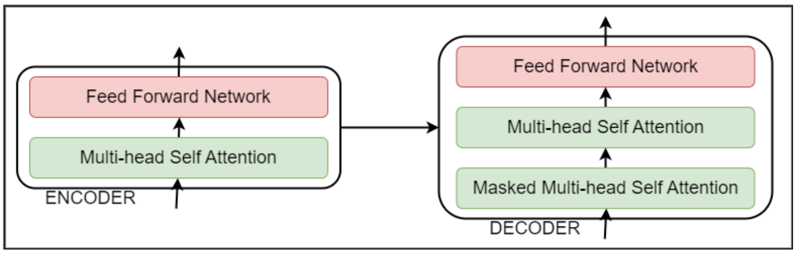

Feed Forward Network
Multi-head Self Attention
ENCODER
Feed Forward Network
Multi-head Self Attention
Masked Multi-head Self Attention
DECODER

Figure 7.2: Encoder and Decoder of Transformer Model

dimension. A dot product of the query with each key is computed, divided by d_k, and then the softmax function measures each word's weight at a specific position.

$$Attention(QY, K, V) = softmax(\frac{QYK^T}{\sqrt{d_k}})V \tag{7.1}$$

In contrast to single-head attention, the Transformer model suggests the idea of multi-head attention, which enables the model to handle various word representations through numerous locations.

$$\begin{aligned} MultiHead(QY, K, V) &= Concat(head_1, head_2, \\ &\qquad head_3...head_h)W^o \\ head_i &= Attention(QRW_i^{QR}, KW_i^K, VW_i^V) \end{aligned} \tag{7.2}$$

Where h=8 parallel attention heads are employed.
The parameter matrices $W_i^{QR} \in \mathbb{R}^{d_{model}Xd_k}$, $W_i^K \in \mathbb{R}^{d_{model}Xd_k}$, $W_i^V \in \mathbb{R}^{d_{model}Xd_v}$ and $W^o \in \mathbb{R}^{hd_vXd_{model}}$.

7.3 Development of English↔Mizo NMT

The data scarcity issue was addressed using the IBT strategy [113] to prepare synthetic parallel data and phrase pairs augmentation [137]. The approach consists of different phases: synthetic parallel data preparation, phrase-pairs extraction, and preparation of a language model for the target language. This work utilized a developed dataset in Sec 5.4 comprising parallel English↔Mizo data and Mizo monolingual data [91]. The English monolingual data from WMT16[1] is used. Further, the monolingual data of the target language (En/Mz) is used to train and generate an LM using the transformer model. The weight matrices are loaded from the pre-trained LM by initializing the decoder of an encoder-decoder architecture of transformer-based NMT. Figure 7.3 demonstrates the English↔Mizo NMT system.

[1] `http://www.statmt.org/wmt16/translation-task.html`

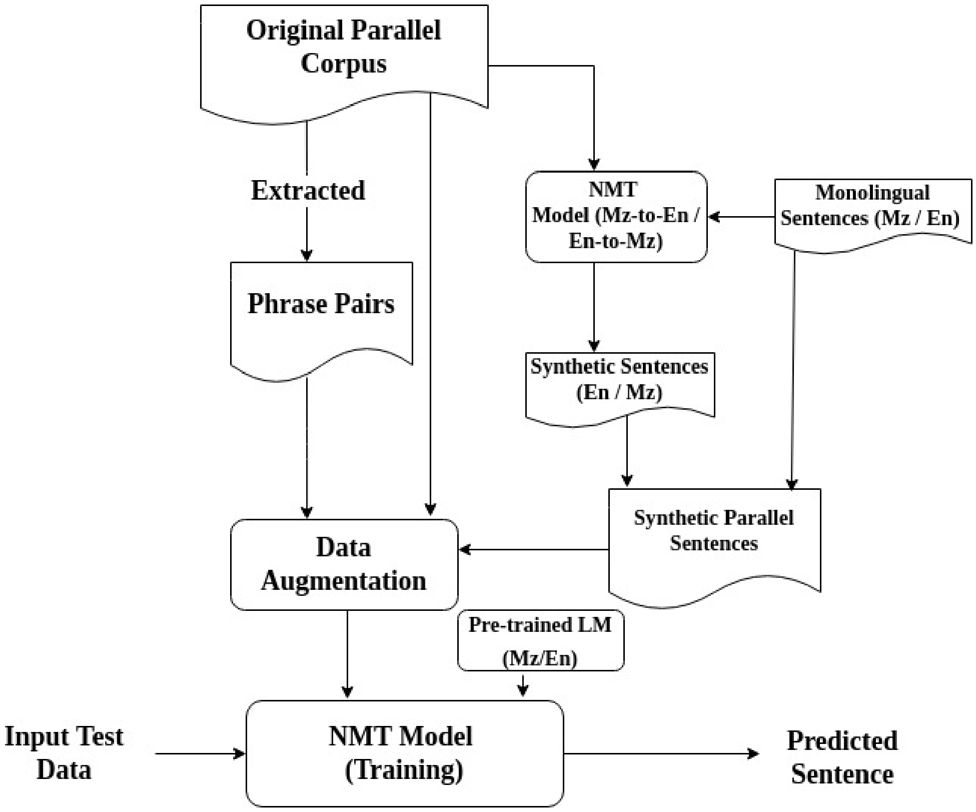


Figure 7.3: English↔Mizo NMT System

- Synthetic parallel data preparation: For synthetic parallel data preparation, both monolingual data (En/Mz) are utilized by following IBT [113] strategy. Experiments were performed in each direction by increasing synthetic data in the ratio of 'original parallel corpus: synthetic parallel'. For instance, the Mz-to-En transformer-based NMT model is used on Mz monolingual data to generate En sentences. The blank lines and under translations (single-word translations) are removed, and synthetic En-Mz parallel data is prepared. The obtained synthetic En-Mz parallel data is augmented with the original parallel data (train set) by performing different ratios of 'original parallel corpus: synthetic parallel'. This process is repeated several times until the convergence condition is reached. The intuition is that not all synthetic parallel sentences are of good quality. Better quality synthetic parallel data was identified by adopting the IBT technique [113]. In this work, the ratios of 1 : 4 and 1 : 3 showed improvement by utilizing Mizo and English monolingual data.

Sentences	Tokens	
	En	Mz
42,110	339,702	266,589

Table 7.1: Statistics for phrase-pairs

Consequently, both were merged, and synthetic parallel data was used by maintaining a ratio of 1 : 7.

- Phrase-pairs extraction: The phrase-pairs extraction strategy from [137, 138] was adopted. Here, phrase-based SMT is trained using Moses[2] toolkit on En-Mz original parallel data and extracted phrase-pairs by considering translation probability $p \geq 0.5$. Also, removed duplicates and the statistics of obtained pairs are presented in Table 7.1. By augmentation of the phrase-pairs train set, more word alignment information is passed to the training model, and the word-order divergence problem is addressed, in addition to tackling the data scarcity issue.

7.4 Experimental Result and Analysis

In the experiment, the publicly available Marian [140] toolkit is employed in three basic operations: data preprocessing, training, and testing. The word-segmentation technique, namely, BPE [111], is used with $32k$ merge operations. The vocabulary sizes of Mz and En are 28,006 and 26,834, respectively. Source-target vocabulary is shared during preprocessing, and the obtained merged vocabulary size is 49,229. The default configuration was followed [26], utilizing 6 layers, 8 attention heads, Adam optimizer with a learning rate 0.001 and drop-out of 0.1 for training the LM (target language) and the NMT system for En-to-Mz and Mz-to-En translation. The Marian toolkit[3] allows the use of custom LM during the training process of the NMT model. The models are trained on a single NVIDIA Quadro P2000 GPU.

[2] `http://www.statmt.org/moses/`

[3] `https://marian-nmt.github.io/docs/`

The predicted sentences from the experiment are evaluated using automatic evaluation metrics. For translation evaluation, the automatic evaluation metrics such as BLEU [59], TER [107], METEOR [60], and F-measure has been implemented with a result as shown in Table 7.2, 7.3, 7.4, and 7.5 respectively. HE was also conducted on 200 sample predicted sentences using a scale of 1-5 [16]. Human evaluators with linguistic knowledge of both languages were hired, and the average scores are reported in Table 7.6. The preliminary experiments in Sec 5.6 [91] show that the transformer-based NMT achieves higher accuracy than RNN-based NMT. Therefore, the transformer-based NMT models are explored in different flavors, which are as follows:

- Baseline: 'Original Parallel Sentences': (Train Set: 118,035)
- With SPA (Synthetic Parallel-Sentence Augmentation): 'Original Parallel Sentences' (Train Set: 118,035) + 'Synthetic Parallel Sentences (826,000)'
- With PPA (Phrase-Pairs Augmentation): 'Original Parallel Sentences' (Train Set: 118,035) + 'Phrase-Pairs (42,110)'
- With LM: 'Original Parallel Sentences' (Train Set: 118,035) + Pre-trained LM (Target Language: En/Mz)
- With SPA + PPA + LM: 'Original Parallel Sentences' (Train Set: 118,035) + 'Synthetic Parallel Sentences (826,000)' + 'Phrase-Pairs (42,110)' + Pre-trained LM (Target Language: En/Mz)

Model	En-to-Mz	Mz-to-En
Baseline	17.86	16.52
With SPA	23.43	22.16
With PPA	22.46	21.67
With SPA+PPA	30.78	28.42
With LM	18.74	17.78
With SPA+LM	26.64	25.18
With PPA + LM	23.72	22.67
With SPA+PPA+LM	32.54	30.26

Table 7.2: BLEU scores of En-to-Mz and Mz-to-En translation. SPA: Synthetic Parallel-Sentence Augmentation, PPA: Phrase Pairs Augmentation, LM: Language Model

Model	En-to-Mz	Mz-to-En
Baseline	63.40	67.52
With SPA	55.27	59.34
With PPA	55.10	60.40
With SPA+PPA	49.67	53.35
With LM	61.40	68.70
With SPA+LM	54.66	58.32
With PPA + LM	54.28	59.69
With SPA+PPA+LM	48.36	51.54

Table 7.3: TER scores of En-to-Mz and Mz-to-En translation, SPA: Synthetic Parallel-Sentence Augmentation, PPA: Phrase Pairs Augmentation, LM: Language Model

Model	En-to-Mz	Mz-to-En
Baseline	0.450914	0.427929
With SPA	0.508452	0.485692
With PPA	0.498104	0.472070
With SPA+PPA	0.546745	0.537843
With LM	0.449661	0.445672
With SPA+LM	0.526754	0.506754
With PPA + LM	0.507864	0.486546
With SPA+PPA+LM	0.558976	0.548794

Table 7.5: F-measure scores of En-to-Mz and Mz-to-En translation, SPA: Synthetic Parallel-Sentence Augmentation, PPA: Phrase Pairs Augmentation, LM: Language Model

Model	**En-to-Mz**	**Mz-to-En**
Baseline	0.182516	0.167540
With SPA	0.259766	0.237848
With PPA	0.236274	0.216831
With SPA+PPA	0.286678	0.268976
With LM	0.199985	0.181323
With SPA+LM	0.268843	2484562
With PPA + LM	0.245322	0.227643
With SPA+PPA+LM	0.295632	0.270842

Table 7.4: METEOR scores of En-to-Mz and Mz-to-En translation, SPA: Synthetic Parallel-Sentence Augmentation, PPA: Phrase Pairs Augmentation, LM: Language Model

Translation	**Model**	**Adequacy**	**Fluency**	**Overall Rating**
En-to-Mz	Baseline	2.58	2.76	2.67
	Present Work (best model)	4.95	5.98	5.46
Mz-to-En	Baseline	2.10	2.36	2.23
	Present Work (best model)	4.56	5.57	5.06

Table 7.6: Human evaluation scores of En-to-Mz and Mz-to-En translation, SPA: Synthetic Parallel-Sentence Augmentation, PPA: Phrase Pairs Augmentation, LM: Language Model

From the quantitative results, it is observed that transformer-based NMT with SPA + PPA + LM attains higher scores for both directions of translation and outperforms previous work in Sec 5.6 [91]. The comparison with previous work and Google Translate[4] are presented in Figure 7.4 and 7.5. It is noticed that En-to-Mz translation evaluation scores outperform Mz-to-En translational evaluation scores due to the number of En tokens in the train data as compared to Mz tokens. As a result, the model encoded more En word frequency; thus, the decoder can generate a better En-to-Mz translation. Moreover, the best model was evaluated on the benchmark dataset [141], and the BLEU score results are reported in 7.7.

Test Set	**En-to-Mz**	**Mz-to-En**
FLORES-200	7.32	5.14

Table 7.7: BLEU scores of En-to-Mz and Mz-to-En translation on FLORES-200 test data

[4] `https://translate.google.co.in/`

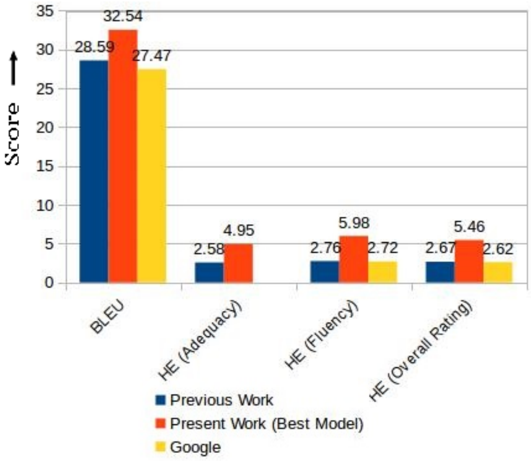


Figure 7.4: Comparison among the previous work, present work (best model: with SPA+PPA+LM), and Google Translate in terms of BLEU and HE scores for En-to-Mz

7.5 Error Analysis and Discussion

To further evaluate the efficiency of English↔Mizo NMT system, the quality of several predicted sentences produced by the transformer models was assessed from various viewpoints alongside Google Translate. The predicted sentences are compared against the reference sentence regarding adequacy and fluency. Adequacy measures how well a reference sentence's meaning is retained in the predicted translation. Fluency is indicated by the appropriate formation of the predicted sentence, regardless of the reference translation. Using the following notations, the predicted sentence samples are presented below to investigate the errors.

- MTS - Mz Test Sentence
- ETS - En Test Sentence
- Base_mz - Predicted sentence of baseline model for En-to-Mz
- Base_en - Predicted sentence of baseline model for Mz-to-En
- Best_mz - Predicted sentence of the best model (With SPA+PPA+LM) for En-to-Mz

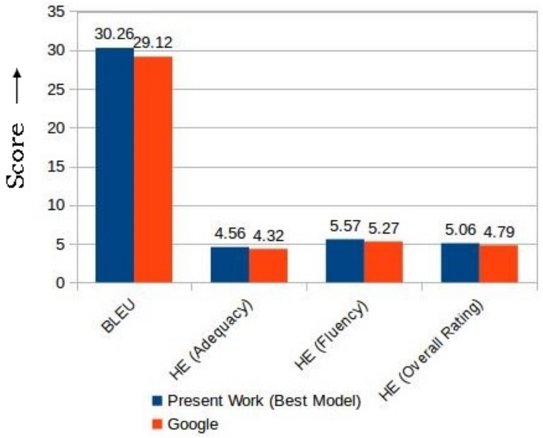


Figure 7.5: Comparison among the present work (best model: with SPA+PPA+LM) and Google Translate in terms of BLEU and HE scores for Mz-to-En

- Best_en - Predicted sentence of the best model (With SPA+PPA+LM) for Mz-to-En
- G_mz - Google Tranlate for En-to-Mz.
- G_en - Google Translate for Mz-to-En.

1. Sample predicted sentence for **adequacy and fluency** (English to Mizo)

MTS - *Jakoba thlàhte zàwng záwng chu mi sàwmsarih an ni.*

ETS - *All the descendants of Jacob were seventy persons.*

Base_mz - *Jakoba thlahte chu mi sawmsarih an ni.*

Best_mz - *Jakoba thlahte zàwng záwng chu sawmsarih an ni.*

G_mz - *Jakoba thlah zawng zawng chu mi sawmsarih an ni.*

2. Sample predicted sentence for **adequacy and fluency** (Mizo to English)

MTS - *Arona tiang chuan an tiangte chu a lem zo ta vék a.*

ETS - *Aaron's staff swallowed up their staffs.*

Base_en - *Aaron struck the staff of Aaron.*

Best_en - *Aaron's staff swallowed up their staffs.*

G_en - *And Aaron 's rod swallowed up their rods.*

Discussion for ***adequacy and fluency***: From the sample predicted sentences of 1 and 2, all the predicted sentences are well structured and contextually correct. Furthermore, they are well-formed and have proper sentence structure. Therefore, they are adequate and fluent in both directions. A tonal word *'thlàhte'*, *'zàwng'*, *'záwng'*, *'sàwmsarih'* and *'vék'* appears in MTS of sample 1 and 2, none of these tonal words are generated in the predicted sentences for all the models except for Best_mz in sample 1. Best_mz has correctly encountered the tonal word *'zàwng'* and *'záwng'* with a correct tone marker.

3. Sample predicted sentence for **inadequacy but fluency** (English to Mizo)
MTS - *Pûk kawngka kha han hawng rawh u.*
ETS - *Open the mouth of the cave.*
Base_mz - *biru rawh u.*
Best_mz - *Tukverh hawng rawh.*
G_mz - *Puk kawngkhar chu hawng rawh.*

4. Sample predicted sentence for **inadequacy but fluency** (Mizo to English)
MTS - *Josefa chu Egypt rámah chuan alo áwm tawh a.*
ETS - *Joseph was already in Egypt.*
Base_en - *Joseph took a party to the Egyptians.*
Best_en - *Joseph had enough of the Egyptians.*
G_en - *Joseph was already in Egypt.*

Discussion for ***inadequacy but fluency***: From the sample predicted sentences of 3 and 4, all of the predicted sentences have context-specific meanings that are entirely different from the reference sentence. As a result, they are considered to be inadequate. However, the predicted sentences are well-formed sentences, and hence they are fluent in both English as well as Mizo. A tonal word *'Pûk'*, *'rámah'* and *'áwm'* appears in MTS of samples 3 and 4, none of the models can correctly generate the tonal word in their predicted sentence.

5. Sample predicted sentence for **partially adequate but perfectly fluent** (English to Mizo)

MTS - *Chu chu thlang lampang rám chu a ni.*

ETS - *This forms the western side.*

Base_mz - *Chu chu chhim lampang chu a ni.*

Best_mz - *Hei hi thlang lampang thlengin a ni.*

G_mz - *Hei hian chhim lam a siam a ni.*

6. Sample predicted sentence for **partially adequate but perfectly fluent** (Mizo to English)

MTS - *Nile luia sanghate chu an thi áng.*

ETS - *The fish in the Nile shall die.*

Base_en - *The fish of the fish is dead.*

Best_en - *The fish in the Nile died.*

G_en - *The fish of the Nile River are dead.*

Discussion for ***partially adequate but perfectly fluent***: From the sample predicted sentences of 5 and 6, all the models have generated most of the words correctly in their predicted sentences. Hence they are partially adequate. However, all the predicted sentences are well formed with correct meaning. Therefore, they are perfectly fluent in both English and Mizo translations. A tonal word *'rám'* and *'áng'* meaning *'shall'* appears in MTS of samples 5 and 6 respectively, all the models are unable to predict the tonal word.

7. Sample predicted sentence for **inadequate and not fluent** (English to Mizo)

MTS - *Theitui ka bùnna lamah ka tibua.*

ETS - *I spilled the juice while I was pouring it.*

Base_mz - *A split a.*

Best_mz - *Ka tân chawhtawlh ka ha a.*

G_mz - *Ka leih lai chuan a tui chu ka theh chhuak a.*

8. Sample predicted sentence for **inadequate and not fluent** (Mizo to English)
MTS - *Tho tam tàk chuan rám chu a tichhe chiam a.*
ETS - *All the land was ruined by the swarms of flies.*
Base_en - *In the hand of all the country, all the people were under the land*
Best_en - *The land was ruined by the land*
G_en - *Then the land was destroyed.*

Discussion for ***inadequate and not fluent***: From the sample predicted sentences of 7 and 8, the predicted translations of all the models are completely different from the reference translation in terms of contextual meaning, and they are not well-formed. Therefore, they are inadequate and not fluent except for G_en in sample 8. As for the tonal word, in sample 7, *'bùnna'* appears in MTS, but all the systems are unable to generate the tonal word in their predicted sentence. However, a tonal word *'tân'* is generated in Best_mz, which is not relevant. In sample 8, a tonal word *'tam tàk'* meaning *'swarms'* and *'rám'* meaning *'land'* appears in MTS. None of the systems are able to detect a tonal word *'tàk'*. However, all the systems detect the tonal word *'rám'* and generate the correct meaning *'land'* in their predicted sentence.

9. Sample predicted sentence for **named-entity problem** (English to Mizo)
MTS - *Naomi mi ti suh u, Marai mi ti zâwk rawh u.*
ETS - *Do not call me Naomi, call me Mara.*
Base_mz - *Naomi min ko va , mi ko va.*
Best_mz - *Naomi mi ti suh , Naomi mi ti zâwk rawh.*
G_mz - *Naomi min ti suh la, Mara ti rawh.*

10. Sample predicted sentence for **named-entity problem** (Mizo to English)

MTS - *A fanu Zipporah chu Moses a pè a.*

ETS - *He gave Moses his daughter Zipporah.*

Base_en - *The daughter of his daughter , the servant of his daughter.*

Best_en - *He opened his daughter s wife.*

G_en - *And he gave Moses his daughter Zipporah.*

Discussion for ***named-entity problem***: From the sample predicted sentence of 9, the named-entity word *'Marai'* is unable to be predicted by all the models. Although it is partially correct in G_mz. A tonal word *'zâwk'* appears in MTS, which is correctly encountered in Best_mz but not in Base_mz and G_mz. As for the sample predicted sentence of 10, the named-entity word *'Moses'* and *'Zipporah'* is unable to be predicted in both Base_en and Best_en model. However, G_en has correctly predicted the named-entity word. A tonal word *'pè'* meaning *'gave'* appears in MTS, but both Base_en and Best_en are unable to detect the MTS's tonal word. However, it is correctly predicted in G_en.

11. Sample predicted sentence for **over-prediction** (English to Mizo)

MTS - *I nu leh pa te ti hèk suh.*

ETS - *Don't used up your parents's money.*

Base_mz - *In pawisa hman dah suh.*

Best_mz - *I nu leh pate chuan sum leh pai te chu hmang suh.*

G_mz - *I nu leh pa pawisa chu hmang zo suh.*

12. Sample predicted sentence for **over-prediction** (Mizo to English)

MTS - *Ka nunna hian a nghak réng a ni.*

ETS - *My soul waits.*

Base_en - *My life is for my life.*

Best_en - *My soul runs in wait for him.*

G_en - *My life is waiting for you.*

Discussion for ***over-prediction***: From the sample predicted sentence of 11, the predicted sentences are partially correct as well as partially fluent. However, the predicted sentence is overpredicted by adding words like *'chuan'*, *'hman'*, and *'chu'*, which is not in the reference sentence. As for sample sentence 12, words like *'life'*, *'for'*, *'runs' 'him'* and *'you'* are included in the predicted sentence which is also not included in the reference sentence. The tonal words, *'hèk'* and *'réng'*, appear in MTS of samples 11 and 12, respectively, but all the models are unable to generate the tonal word in their predicted sentence.

13. Sample predicted sentence for **under-prediction** (English to Mizo)
MTS - *Ní a, Gibeon khaw chúngah díng rèng rawh.*
ETS - *Sun, stand still at Gibeon.*
Base_mz - *Ngawi rawh u.*
Best_mz - *Gibeon-ah.*
G_mz - *Sun, Gibeon-ah chuan ding reng rawh.*

14. Sample predicted sentence for **under-prediction** (Mizo to English)
MTS - *Va chhuak la, va bèi mawlh rawh.*
ETS - *Go out now and fight with them.*
Base_en - *Go out, go out.*
Best_en - *Go out, and go.*
G_en - *Go out, go out, go out.*

Discussion for ***under-prediction***: From the sample predicted sentences of 13 and 14, all the models have encountered under-prediction since the predicted sentence predicts only part of the reference translation. Words like *'Ní a'*, *'khaw'*, and *'chúngah'* are not generated in the predicted sentences of sample 13. Likewise, *'now'*, *'and'*, *'fight' 'with'* and *'them'* are also not generated in the predicted sentences of sample 14. As for tonal words, *'chúngah'*, *'díng'*, *'rèng'* and *'bèi'* meaning *'fight'* appear in MTS of sample 13 and 14, but all the models are unable to generate or detect the tonal word in their predicted sentence.

7.6 Conclusion

The data scarcity problem for English-to-Mizo and vice-versa translation has been addressed using transformer-based NMT. It is performed by augmenting synthetic parallel sentences and phrase pairs to expand the training amount of data and LM at the target side. Several experiments have been conducted in this thesis. Encountering tonal words with data augmentation techniques, implementing Bert-fused NMT with post-processing steps, developing MMT for the Mizo language, and addressing data scarcity issues were implemented, resulting in promising output. The experimental results show that the current work attains the best translational accuracy compared to the previous work. Based on the discussions in error analysis, most predicted translations are acceptable in terms of adequacy and fluency for both the best model and Google Translate. Some of the tone markers in the Mizo language are also predicted and encountered correctly in the best model. In contrast, Google Translate has never predicted tonal words or tone markers.

Chapter 8

CONCLUSION

The thesis explored developing and evaluating an English↔Mizo MT system to bridge the language gap between English and Mizo. The research focused on designing a system capable of translating English text into Mizo with high accuracy, considering the unique linguistic characteristics of both languages.

Developing effective MT systems for English↔Mizo language pairs presents unique challenges. The research began by studying the structure and challenges of the Mizo language. Despite the advances made in MT technologies, the study acknowledges several limitations. Mizo is considered a low-resource language due to the scarcity of linguistic resources, which poses challenges for machine translation. Although NMT has struggled with low-resource language, various works have been done by several researchers to deal with it. Key challenges include the tonality of the Mizo language, limited data availability, linguistic complexity, and resource constraints. However, significant improvements can be made through collaborative efforts, dealing with linguistic challenges, and leveraging advanced techniques. As technology evolves, focusing on low-resource languages through back-translation, transfer learning, and community engagement offers promising pathways to develop effective MT systems for English↔Mizo translation.

Several studies have explored MT for English↔Mizo pairs. However, no existing research specifically addresses the challenge of Mizo tonal words in low-resource English↔Mizo translations. Consequently, overcoming the challenges of low-resource settings and tonal complexities necessitates the development of unique strategies to enhance translation quality. The English↔Mizo corpus EnMzCorp1.0 was then developed from different sources. The corpus consists of both parallel and monolingual data of Mizo. Various NMT models were investigated using the developed dataset. A data augmentation approach is proposed to encounter tonal words, resulting in better translation accuracy than the baseline system. Furthermore, an English↔Mizo NMT system was proposed, utilizing BERT-fused NMT with bidirectional data augmentation with synthetic parallel corpus. One of the primary challenges in the Mizo language is handling tonal words, which add to the complexity of any NLP task, especially given the low-resource nature of the language. Various post-processing steps are employed to address the issue. Better translation accuracy was achieved compared to a conventional transformer and BERT-fused NMT. The proposed system demonstrated promising results, with improved translation quality over previous models, by effectively managing tonal words in Mizo through a post-processing step, leading to state-of-the-art results in English-to-Mizo translation.

The concept of multimodality, which integrates textual and visual data, has been applied to English↔Mizo machine translation. The approach addresses the challenges of translating low-resource languages like Mizo by adding additional features. The Mizo Visual Genome 1.0 (MVG 1.0) dataset was created without a standard multimodal corpus for this language pair. MVG 1.0 features images paired with bilingual textual descriptions, designed explicitly for English↔Mizo multimodal machine translation. Evaluations using automated metrics reveal that multimodal neural machine translation (MNMT) significantly outperforms traditional text-only neural machine translation (NMT) regarding accuracy and fluency. By leveraging the complementary nature of visual and textual data, MNMT systems achieve a deeper understanding and improved translation quality. The English↔Mizo MNMT system represents the first effort in this field, establishing a baseline for future research on this low-resource language pair.

The exploration of MMT for English↔Mizo demonstrates the potential of combining textual and visual data to enhance translation quality. The development of MVG 1.0 has laid the foundation for further advancements, with plans to expand the dataset size and conduct additional experiments to improve system performance. This pioneering work addresses the challenges associated with low-resource languages and highlights the effectiveness of multimodal approaches in improving the overall quality of machine translation systems.

Various transformer-based NMT models such as data augmentation techniques, implementing BERT-fused NMT with post-processing steps, and developing a multimodal machine translation (MMT) system for the English↔Mizo pair were investigated to tackle data scarcity and the challenges posed by tonal words in English↔Mizo translation. To further mitigate these issues, transformer-based NMT models were employed, utilizing synthetic parallel sentence generation, phrase pair augmentation, and pre-trained language models (LM). These techniques expanded training data and improved target-side language modeling, enhancing overall translation performance. The system focuses on increasing the volume of training data and improving token alignment through phrase pair augmentation. By augmenting synthetic parallel data and phrase pairs, the problem of data scarcity is effectively mitigated. Additionally, integrating a pre-trained language model (LM) further enhances the quality of English↔Mizo translations. State-of-the-art results have been achieved on the English-to-Mizo and Mizo-to-English translation test data. The model correctly predicts and handles some of the tone markers in the Mizo language. In contrast, Google Translate fails to predict tonal words or tone markers. The proposed system demonstrates the effectiveness of the developed system in producing accurate translations for the English↔Mizo language pair, particularly in handling the unique tonal characteristics of Mizo, which remain a challenge for other translation systems.

The study has demonstrated that while MT between English and Mizo presents particular challenges, significant progress has been made in improving translation quality. The research emphasizes the significance of linguistic resources, such as bilingual corpora and linguistic rules, in enhancing the performance of MT systems. After conducting multiple experiments, the research has contributed to developing the English↔Mizo NMT system, which establishes a robust foundation for future advancements in low-resource MT and sets a benchmark for translations in the language pair. Additionally, the development of MMT further enhanced translation accuracy by leveraging visual context alongside textual data. By leveraging these models, the MT system produced reasonably accurate translations, although some issues with fluency and idiomatic expressions remain.

Key findings from this research include the importance of developing a bilingual corpus and incorporating linguistic rules to enhance the translation process. The quality of the translations was heavily dependent on the size and variety of the training data, as well as the linguistic richness of the Mizo language resources. Despite many advancements throughout the experiments, several limitations of the current system were identified. One of the primary challenges was the limited scope of the bilingual corpus, which restricted the system's ability to handle diverse domains and specialized vocabulary. Moreover, translating sentences with complex syntax or cultural references often lacked fluency, resulting in low quality. As compared to more widely spoken languages, the absence of a rich set of linguistic resources for Mizo also presents a barrier to achieving high-quality translations across various contexts.

In future research, several areas could be explored to enhance the English-Mizo MT system. Improving the dataset with abundant parallel corpus from different domains, incorporating deep learning techniques, and developing more sophisticated models to handle Mizo's unique syntactic and morphological structures will help enhance the performance of NMT in both directions by achieving better translation accuracy. Additionally, exploring the integration of Mizo dialects and incorporating domain-specific translation models would further enhance the system's utility and accuracy.

In conclusion, the English↔Mizo MT system developed in this thesis represents an important step toward bridging the language gap between English and Mizo. While the current system has limitations, it provides a foundation for future research and development in the field, potentially improving communication, education, and digital accessibility for Mizo speakers. Continuous efforts in data collection, resource development, and collaborative research will be crucial in enhancing the quality and usability of MT systems for the English↔Mizo language pair, thus preserving and promoting linguistic diversity in the digital age.

Bibliography

[1] P. Pakray, A. Pal, G. Majumder, and A. Gelbukh, "Resource building and parts-of-speech (pos) tagging for the mizo language," in *2015 Fourteenth Mexican International Conference on Artificial Intelligence (MICAI)*, pp. 3–7, 2015.

[2] M. Nunsanga, D. P. Pakray, C. Lallawmsanga, and L. Singh, "Part-of-speech tagging for mizo language using conditional random field," *Computación y Sistemas*, vol. 25, 12 2021.

[3] M. T. H. K. Tusar and M. T. Islam, "A comparative study of sentiment analysis using NLP and different machine learning techniques on US airline twitter data," *CoRR*, vol. abs/2110.00859, 2021.

[4] S. Ilias and M. F. Shamsudin, "Customer satisfaction and business growth," *Journal of Undergraduate Social Science and Technology*, vol. 2, Mar. 2020.

[5] F. Al-Attar and K. Shaalan, "Using artificial intelligence to understand what causes sentiment changes on social media," *IEEE Access*, vol. PP, pp. 1–1, 04 2021.

[6] N. Jain and S. Rastogi, "Speech recognition systems – a comprehensive study of concepts and mechanism," *Acta Informatica Malaysia*, vol. 3, pp. 01–03, 01 2019.

[7] H. Li and Z. Li, "Text classification based on machine learning and natural language processing algorithms," *Wireless Communications and Mobile Computing*, vol. 2022, pp. 1–12, 07 2022.

[8] W. Weaver, "Translation," in *Machine Translation of Languages* (W. N. Locke and A. D. Boothe, eds.), pp. 15–23, Cambridge, MA: MIT Press, 1949/1955. Reprinted from a memorandum written by Weaver in 1949.

[9] V. H. Yngve, "Sentence-for-sentence translation," *Mech. Transl. Comput. Linguistics*, vol. 2, pp. 29–37, 1955.

[10] V. H. Yngve, "A framework for syntactic translation," *Mech. Transl. Comput. Linguistics*, vol. 4, pp. 59–65, 1957.

[11] B. Vauquois, G. Veillon, and J. Veyrunes, "Syntax and interpretation," *Mech. Transl. Comput. Linguistics*, vol. 9, pp. 44–54, 1966.

[12] R. H. Richens, "Interlingual Machine Translation," *The Computer Journal*, vol. 1, pp. 144–147, 01 1958.

[13] P. Koehn, F. J. Och, and D. Marcu, “Statistical phrase-based translation,” in *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 127–133, 2003.

[14] P. Koehn, *Statistical machine translation.* Cambridge University Press, 2009.

[15] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, pp. 1–15, 2015.

[16] A. Pathak, P. Pakray, and J. Bentham, “English–mizo machine translation using neural and statistical approaches,” *Neural Computing and Applications*, vol. 30, pp. 1–17, Jun 2018.

[17] A. Pathak and P. Pakray, “Neural machine translation for indian languages,” *Journal of Intelligent Systems*, pp. 1–13, 06 2018.

[18] S. R. Laskar, A. Dutta, P. Pakray, and S. Bandyopadhyay, “Neural machine translation: English to hindi,” in *2019 IEEE Conference on Information and Communication Technology*, pp. 1–6, 2019.

[19] S. R. Laskar, P. Pakray, and S. Bandyopadhyay, “Neural machine translation: Hindi-Nepali,” in *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, (Florence, Italy), pp. 202–207, Association for Computational Linguistics, Aug. 2019.

[20] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, “Learning phrase representations using RNN encoder–decoder for statistical machine translation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (Doha, Qatar), pp. 1724–1734, Association for Computational Linguistics, Oct. 2014.

[21] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio, “On the properties of neural machine translation: Encoder-decoder approaches,” in *Proceedings of SSST@EMNLP 2014, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, Doha, Qatar, 25 October 2014* (D. Wu, M. Carpuat, X. Carreras, and E. M. Vecchi, eds.), pp. 103–111, Association for Computational Linguistics, 2014.

[22] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks,” in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS’14, (Cambridge, MA, USA), p. 3104–3112, MIT Press, 2014.

[23] S. H. Ramesh and K. P. Sankaranarayanan, “Neural machine translation for low resource languages using bilingual lexicon induced from comparable corpora,” in *Proceedings of the 2018 Conference of the North American Chapter*

of the Association for Computational Linguistics: Student Research Workshop, (New Orleans, Louisiana, USA), pp. 112–119, Association for Computational Linguistics, June 2018.

[24] N. Kalchbrenner and P. Blunsom, "Recurrent continuous translation models," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, (Seattle, Washington, USA), pp. 1700–1709, Association for Computational Linguistics, Oct. 2013.

[25] J. Gehring, M. Auli, D. Grangier, and Y. Dauphin, "A convolutional encoder model for neural machine translation," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (Vancouver, Canada), pp. 123–135, Association for Computational Linguistics, July 2017.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems, December 4-9, 2017, Long Beach, CA, USA*, pp. 5998–6008, 2017.

[27] C. Lalrempuii and B. Soni, "Attention-based english to mizo neural machine translation," in *Machine Learning, Image Processing, Network Security and Data Sciences*, (Singapore), pp. 193–203, Springer Singapore, 2020.

[28] W. J. Hutchins, "The evolution of machine translation systems," in *Translating and the Computer: Practical experience of machine translation*, (London, UK), Aslib, Nov. 5-6 1981.

[29] S. Nirenburg, H. L. Somers, and Y. Wilks, "Alpac: The (in)famous report," 2003.

[30] N. Chomsky, *Aspects of the Theory of Syntax.* The MIT Press, 50 ed., 1965.

[31] O. Furuse and H. Iida, "An example-based method for transfer-driven machine translation," in *Proceedings of the Fourth Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, 1992.

[32] D. A. Hull and G. Grefenstette, "Querying across languages: A dictionary-based approach to multilingual information retrieval," in *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '96, (New York, NY, USA), p. 49–57, Association for Computing Machinery, 1996.

[33] S. K. Mondal, H. Zhang, H. M. D. Kabir, K. Ni, and H.-N. Dai, "Machine translation and its evaluation: a study," *Artificial Intelligence Review*, Feb 2023.

[34] C. Hardmeier, "Discourse in statistical machine translation: A survey and a case study," *Discours*, 12 2012.

[35] P. F. Brown, V. J. D. Pietra, S. A. D. Pietra, and R. L. Mercer, "The mathematics of statistical machine translation: parameter estimation," *Comput. Linguist.*, vol. 19, p. 263–311, jun 1993.

[36] S. M. Shieber and Y. Schabes, "Synchronous tree-adjoining grammars," 1991.

[37] S. M. Shieber and Y. Schabes, "Generation and synchronous tree-adjoining grammars," *Computational Intelligence*, vol. 7, no. 4, pp. 220–228, 1991.

[38] M. Marcus, B. Santorini, and M. A. Marcinkiewicz, "Building a large annotated corpus of english: The penn treebank," 1993.

[39] F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Computational linguistics*, vol. 29, no. 1, pp. 19–51, 2003.

[40] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[41] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, (Lisbon, Portugal), pp. 1412–1421, Association for Computational Linguistics, Sept. 2015.

[42] D. Rothman, *Transformers for Natural Language Processing: Build innovative deep neural network architectures for NLP with Python, PyTorch, TensorFlow, BERT, RoBERTa, and more.* Packt Publishing Ltd, 2021.

[43] Y. Cui, Z. Chen, S. Wei, S. Wang, T. Liu, and G. Hu, "Attention-over-attention neural networks for reading comprehension," *arXiv preprint arXiv:1607.04423*, 2016.

[44] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (J. Burstein, C. Doran, and T. Solorio, eds.), (Minneapolis, Minnesota), pp. 4171–4186, Association for Computational Linguistics, June 2019.

[45] S. Ramnath, M. Johnson, A. Gupta, and A. Raghuveer, "Hintedbt: Augmenting back-translation with quality and transliteration hints," *arXiv preprint arXiv:2109.04443*, 2021.

[46] H. Wu and H. Wang, "Pivot language approach for phrase-based statistical machine translation," *Machine Translation*, vol. 21, pp. 165–181, 2007.

[47] C.-H. Liu, C. C. Silva, L. Wang, and A. Way, “Pivot machine translation using chinese as pivot language,” in *Machine Translation: 14th China Workshop, CWMT 2018, Wuyishan, China, October 25-26, 2018, Proceedings 14*, pp. 74–85, Springer, 2019.

[48] B. Ahmadnia, B. J. Dorr, and P. Kordjamshidi, “Knowledge graphs effectiveness in neural machine translation improvement,” *Computer Science*, vol. 21, 2020.

[49] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, “Freebase: a collaboratively created graph database for structuring human knowledge,” in *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pp. 1247–1250, 2008.

[50] Y. Lu, J. Zhang, and C. Zong, “Exploiting knowledge graph in neural machine translation,” in *Machine Translation: 14th China Workshop, CWMT 2018, Wuyishan, China, October 25-26, 2018, Proceedings 14*, pp. 27–38, Springer, 2019.

[51] D. Moussallem, A.-C. Ngonga Ngomo, P. Buitelaar, and M. Arcan, “Utilizing knowledge graphs for neural machine translation augmentation,” in *Proceedings of the 10th international conference on knowledge capture*, pp. 139–146, 2019.

[52] Y. Zhao, J. Zhang, Y. Zhou, and C. Zong, “Knowledge graphs enhanced neural machine translation,” in *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pp. 4039–4045, 2021.

[53] C. Chu and R. Wang, “A survey of domain adaptation for neural machine translation,” 2018.

[54] P. Koehn and R. Knowles, “Six challenges for neural machine translation,” in *Proceedings of the First Workshop on Neural Machine Translation*, (Vancouver), pp. 28–39, Association for Computational Linguistics, Aug. 2017.

[55] M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Viégas, M. Wattenberg, G. Corrado, M. Hughes, and J. Dean, “Google’s multilingual neural machine translation system: Enabling zero-shot translation,” 2017.

[56] R. Aharoni, M. Johnson, and O. Firat, “Massively multilingual neural machine translation,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (J. Burstein, C. Doran, and T. Solorio, eds.), (Minneapolis, Minnesota), pp. 3874–3884, Association for Computational Linguistics, June 2019.

[57] S. Edunov, M. Ott, M. Auli, and D. Grangier, “Understanding back-translation at scale,” 2018.

[58] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, Łukasz Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean, "Google's neural machine translation system: Bridging the gap between human and machine translation," 2016.

[59] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, (Stroudsburg, PA, USA), pp. 311–318, Association for Computational Linguistics, 2002.

[60] A. Lavie and M. J. Denkowski, "The meteor metric for automatic evaluation of machine translation," *Machine Translation*, vol. 23, p. 105–115, Sept. 2009.

[61] A. Conneau and G. Lample, "Cross-lingual language model pretraining," in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada* (H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, eds.), (Canada), pp. 7057–7067, Advances in Neural Information Processing Systems (NeurIPS), 2019.

[62] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," 2018.

[63] K. Megerdoomian and D. Parvaz, "Low-density language bootstrapping: the case of tajiki persian," in *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco*, pp. 3293–3298, European Language Resources Association, 2008.

[64] K. Probst, R. D. Brown, J. G. Carbonell, A. Lavie, L. Levin, and E. Peterson, "Design and Implementation of Controlled Elicitation for Machine Translation of Low-density Languages," 6 2003.

[65] J. Gu, H. Hassan, J. Devlin, and V. O. Li, "Universal neural machine translation for extremely low resource languages," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, (New Orleans, Louisiana), pp. 344–354, Association for Computational Linguistics, June 2018.

[66] T. Kocmi, "Exploring benefits of transfer learning in neural machine translation," *CoRR*, vol. abs/2001.01622, 2020.

[67] P. Zaremoodi, W. Buntine, and G. Haffari, "Adaptive knowledge sharing in multi-task learning: Improving low-resource neural machine translation," in

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), (Melbourne, Australia), pp. 656–661, Association for Computational Linguistics, July 2018.

[68] Ç. Gülçehre, O. Firat, K. Xu, K. Cho, L. Barrault, H. Lin, F. Bougares, H. Schwenk, and Y. Bengio, "On using monolingual corpora in neural machine translation," *CoRR*, vol. abs/1503.03535, 2015.

[69] F. Burlot and F. Yvon, "Using monolingual data in neural machine translation: a systematic study," in *Proceedings of the Third Conference on Machine Translation: Research Papers*, (Brussels, Belgium), pp. 144–155, Association for Computational Linguistics, Oct. 2018.

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

[71] A. Imankulova, T. Sato, and M. Komachi, "Filtered pseudo-parallel corpus improves low-resource neural machine translation," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 19, Oct. 2019.

[72] D. Variš and O. Bojar, "Unsupervised pretraining for neural machine translation using elastic weight consolidation," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, (Florence, Italy), pp. 130–135, Association for Computational Linguistics, July 2019.

[73] L. Wu, Y. Wang, Y. Xia, T. Qin, J. Lai, and T.-Y. Liu, "Exploiting monolingual data at scale for neural machine translation," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, (Hong Kong, China), pp. 4207–4216, Association for Computational Linguistics, Nov. 2019.

[74] I. Feldman and R. Coto-Solano, "Neural machine translation models with back-translation for the extremely low-resource indigenous language Bribri," in *Proceedings of the 28th International Conference on Computational Linguistics*, (Barcelona, Spain (Online)), pp. 3965–3976, International Committee on Computational Linguistics, Dec. 2020.

[75] A. Mueller, G. Nicolai, A. D. McCarthy, D. Lewis, W. Wu, and D. Yarowsky, "An analysis of massively multilingual neural machine translation for low-resource languages," in *Proceedings of the 12th Language Resources and Evaluation Conference*, (Marseille, France), pp. 3710–3718, European Language Resources Association, May 2020.

[76] S. R. Laskar, A. F. U. R. Khilji, P. Pakray, and S. Bandyopadhyay, "Hindi-Marathi cross lingual model," in *Proceedings of the Fifth Conference on Machine Translation*, (Online), pp. 396–401, Association for Computational Linguistics, Nov. 2020.

[77] I. Calixto, Q. Liu, and N. Campbell, "Doubly-Attentive Decoder for Multi-modal Neural Machine Translation," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (Vancouver, Canada), pp. 1913–1924, Association for Computational Linguistics, July 2017.

[78] S. R. Laskar, A. F. U. R. Khilji, P. Pakray, and S. Bandyopadhyay, "Multi-modal neural machine translation for English to Hindi," in *Proceedings of the 7th Workshop on Asian Translation*, (Suzhou, China), pp. 109–113, Association for Computational Linguistics, Dec. 2020.

[79] K. Dutta Chowdhury, M. Hasanuzzaman, and Q. Liu, "Multimodal neural machine translation for low-resource language pairs using synthetic data," in *Proceedings of the Workshop on Deep Learning Approaches for Low-Resource NLP*, (Melbourne), pp. 33–42, Association for Computational Linguistics, July 2018.

[80] S. R. Laskar, B. Paul, S. Paudwal, P. Gautam, N. Biswas, and P. Pakray, "Multimodal neural machine translation for english-assamese pair," in *2021 International Conference on Computational Performance Evaluation (ComPE)*, pp. 387–392, 2021.

[81] O. Caglayan, L. Barrault, and F. Bougares, "Multimodal attention for neural machine translation," *CoRR*, vol. abs/1609.03976, 2016.

[82] R. Wang, H. Sun, K. Chen, C. Ding, M. Utiyama, and E. Sumita, "English-Myanmar supervised and unsupervised NMT: NICT's machine translation systems at WAT-2019," in *Proceedings of the 6th Workshop on Asian Translation*, (Hong Kong, China), pp. 90–93, Association for Computational Linguistics, Nov. 2019.

[83] R. N. Patel, P. B. Pimpale, and M. Sasikumar, "Machine translation in indian languages: Challenges and resolution," 2018.

[84] H. Choudhary, S. Rao, and R. Rohilla, "Neural machine translation for low-resourced indian languages," *CoRR*, vol. abs/2004.13819, 2020.

[85] T. Tayir, L. Li, B. Li, J. Liu, and K. A. Lee, "Encoder-decoder calibration for multimodal machine translation," *IEEE Transactions on Artificial Intelligence*, vol. PP, pp. 1–9, 01 2024.

[86] S. Saini and V. Sahula, "Neural machine translation for english to hindi," in *2018 Fourth International Conference on Information Retrieval and Knowledge Management (CAMP)*, pp. 1–6, 03 2018.

[87] J. Náplava, M. Straka, P. Straňák, and J. Hajič, “Diacritics restoration using neural networks,” in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, (Miyazaki, Japan), European Language Resources Association (ELRA), May 2018.

[88] D. I. Adelani, D. Ruiter, J. O. Alabi, D. Adebonojo, A. Ayeni, M. Adeyemi, A. Awokoya, and C. España-Bonet, “The effect of domain and diacritics in yoruba-english neural machine translation,” in *Proceedings of the 18th Biennial Machine Translation Summit - Volume 1: Research Track, MTSummit 2021 Virtual, August 16-20, 2021*, pp. 61–75, Association for Machine Translation in the Americas, 2021.

[89] A. Fadel, I. Tuffaha, B. Al-Jawarneh, and M. Al-Ayyoub, “Neural arabic text diacritization: State of the art results and a novel approach for machine translation,” in *Proceedings of the 6th Workshop on Asian Translation, WAT@EMNLP-IJCNLP 2019, Hong Kong, China, November 4, 2019*, pp. 215–225, Association for Computational Linguistics, 2019.

[90] C. Lalrempuii, B. Soni, and D. P. Pakray, “An improved english-to-mizo neural machine translation,” *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 20, pp. 1–21, 05 2021.

[91] V. Khenglawt, S. R. Laskar, S. Pal, P. Pakray, and A. K. Khan, “Language resource building and english-to-mizo neural machine translation encountering tonal words,” in *In Proceedings of the WILDRE-6 Workshop @LREC2020, Marseille, European Language Resources Association (ELRA)*, pp. 48–54, June 2022.

[92] G. Majumder, P. Pakray, Z. Khiangte, and A. Gelbukh, “Multiword expressions (mwe) for mizo language: Literature survey,” in *Computational Linguistics and Intelligent Text Processing* (A. Gelbukh, ed.), (Cham), pp. 623–635, Springer International Publishing, 2018.

[93] J. Bentham, P. Pakray, G. Majumder, S. Lalbiaknia, and A. Gelbukh, “Identification of rules for recognition of named entity classes in mizo language,” in *2016 Fifteenth Mexican International Conference on Artificial Intelligence (MICAI)*, pp. 8–13, IEEE, 2016.

[94] M. V. L. Nunsanga, P. Pakray, M. Lalngaihtuaha, and L. Lolit Kumar Singh, “Stochastic based part of speech tagging in mizo language: Unigram and bigram hidden markov model,” in *Edge Analytics* (R. Patgiri, S. Bandyopadhyay, M. D. Borah, and V. Emilia Balas, eds.), (Singapore), pp. 711–722, Springer Singapore, 2022.

[95] M. V. L. Nunsanga, P. Pakray, M. Lalngaihtuaha, and L. Lolit Kumar Singh, “Part-of-speech tagging in mizo language: A preliminary study,” in *Data Intelligence and Cognitive Informatics* (I. Jeena Jacob, S. Kolandapalayam Shanmugam, S. Piramuthu, and P. Falkowski-Gilski, eds.), (Singapore), pp. 625–635, Springer Singapore, 2021.

[96] B. Lalthangliana, "History and culture of mizo in india, burma and bangladesh," 1892. Baptist Missionary Conference.

[97] P. Sarmah and C. R. Wiltshire, "A preliminary acoustic study of mizo vowels and tones," *J. Acoust. Soc. Ind*, vol. 37, no. 3, pp. 121–129, 2010.

[98] L. T. Fanai, "Some aspects of the lexical phonology of mizo and english an autosegmental approach," 1992.

[99] S. R. Laskar, A. F. U. R. Khilji, P. Pakray, and S. Bandyopadhyay, "EnAsCorp1.0: English-Assamese corpus," in *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, (Suzhou, China), pp. 62–68, Association for Computational Linguistics, Dec. 2020.

[100] C. Hogan, "Ocr for minority languages," in *Symposium on Document Image Understanding Technology*, 1999.

[101] S. R. Laskar, A. Faiz Ur Rahman Khilji Darsh Kaushik, P. Pakray, and S. Bandyopadhyay, "EnKhCorp1.0: An English–Khasi corpus," in *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021)*, (Virtual), pp. 89–95, Association for Machine Translation in the Americas, Aug. 2021.

[102] T. Mikolov, M. Karafiát, L. Burget, J. Cernocký, and S. Khudanpur, "Recurrent neural network based language model," vol. 2, pp. 1045–1048, 01 2010.

[103] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, p. 1735–1780, Nov. 1997.

[104] K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder–decoder approaches," in *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, (Doha, Qatar), pp. 103–111, Association for Computational Linguistics, Oct. 2014.

[105] M. Federico, N. Bertoldi, and M. Cettolo, "IRSTLM: an open source toolkit for handling large scale language models," in *INTERSPEECH 2008, 9th Annual Conference of the International Speech Communication Association, Brisbane, Australia, September 22-26, 2008*, pp. 1618–1621, ISCA, 2008.

[106] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, A meeting of SIGDAT, a Special Interest Group of the ACL* (A. Moschitti, B. Pang, and W. Daelemans, eds.), pp. 1532–1543, ACL, 2014.

[107] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A study of translation edit rate with targeted human annotation," in *In Proceedings of Association for Machine Translation in the Americas*, pp. 223–231, 2006.

[108] K. Imamura and E. Sumita, "NICT self-training approach to neural machine translation at NMT-2018," in *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, (Melbourne, Australia), pp. 110–115, Association for Computational Linguistics, July 2018.

[109] Y. Cheng, W. Xu, Z. He, W. He, H. Wu, M. Sun, and Y. Liu, "Semi-supervised learning for neural machine translation," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (Berlin, Germany), pp. 1965–1974, Association for Computational Linguistics, Aug. 2016.

[110] Y. Wang, Y. Xia, L. Zhao, J. Bian, T. Qin, E. Chen, and T. Liu, "Semi-supervised neural machine translation via marginal distribution estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 10, pp. 1564–1576, 2019.

[111] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (Berlin, Germany), pp. 1715–1725, Association for Computational Linguistics, Aug. 2016.

[112] W. Adouane and J. Bernardy, "When is multi-task learning beneficial for low-resource noisy code-switched user-generated algerian texts?," in *Proceedings of the The 4th Workshop on Computational Approaches to Code Switching, CodeSwitch@LREC 2020, May, 2020*, (Marseille, France), pp. 17–25, European Language Resources Association, 2020.

[113] V. C. D. Hoang, P. Koehn, G. Haffari, and T. Cohn, "Iterative back-translation for neural machine translation," in *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, (Melbourne, Australia), pp. 18–24, Association for Computational Linguistics, July 2018.

[114] F. Guzmán, P. Chen, M. Ott, J. M. Pino, G. Lample, P. Koehn, V. Chaudhary, and M. Ranzato, "Two new evaluation datasets for low-resource machine translation: Nepali-english and sinhala-english," *CoRR*, vol. abs/1902.01382, 2019.

[115] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013.

[116] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[117] J. Pfeiffer, A. Kamath, A. Rücklé, K. Cho, and I. Gurevych, "AdapterFusion: Non-destructive task composition for transfer learning," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (P. Merlo, J. Tiedemann, and R. Tsarfaty,

eds.), (Online), pp. 487–503, Association for Computational Linguistics, Apr. 2021.

[118] G. Lample, L. Denoyer, and M. Ranzato, “Unsupervised machine translation using monolingual corpora only,” *CoRR*, vol. abs/1711.00043, 2017.

[119] M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Viégas, M. Wattenberg, G. Corrado, M. Hughes, and J. Dean, “Google’s multilingual neural machine translation system: Enabling zero-shot translation,” *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 339–351, 2017.

[120] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer, “Multilingual denoising pre-training for neural machine translation,” *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 726–742, 2020.

[121] K. Song, X. Tan, T. Qin, J. Lu, and T. Liu, “MASS: masked sequence to sequence pre-training for language generation,” *CoRR*, vol. abs/1905.02450, 2019.

[122] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, “Unsupervised cross-lingual representation learning at scale,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, eds.), (Online), pp. 8440–8451, Association for Computational Linguistics, July 2020.

[123] C. Jia, Y. Yang, Y. Xia, Y. Chen, Z. Parekh, H. Pham, Q. V. Le, Y. Sung, Z. Li, and T. Duerig, “Scaling up visual and vision-language representation learning with noisy text supervision,” *CoRR*, vol. abs/2102.05918, 2021.

[124] Y.-A. Chung and J. Glass, “Speech2vec: A sequence-to-sequence framework for learning word embeddings from speech,” 03 2018.

[125] X. Niu, M. J. Denkowski, and M. Carpuat, “Bi-directional neural machine translation with synthetic parallel data,” in *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation, NMT@ACL 2018, Melbourne, Australia, July 20, 2018* (A. Birch, A. M. Finch, M. Luong, G. Neubig, and Y. Oda, eds.), pp. 84–91, Association for Computational Linguistics, 2018.

[126] J. Zhu, Y. Xia, L. Wu, D. He, T. Qin, W. Zhou, H. Li, and T. Liu, “Incorporating BERT into neural machine translation,” in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, OpenReview.net, 2020.

[127] R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes, and A. Jatowt, “Yake! keyword extraction from single documents using multiple local features,” *Inf. Sci.*, vol. 509, pp. 257–289, 2020.

[128] A. Pathak and D. P. Pakray, “English-mizo machine translation using neural and statistical approaches,” *Neural Computing and Applications*, vol. 31, 11 2019.

[129] Z. Thihlum, V. Khenglawt, and S. Debnath, “Machine translation of english language to mizo language,” in *2020 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM)*, pp. 92–97, 2020.

[130] I. Calixto and Q. Liu, “Incorporating Global Visual Features into Attention-Based Neural Machine Translation,” in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, (Copenhagen, Denmark), 2017.

[131] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[132] S. Parida and O. Bojar, “Hindi visual genome 1.1,” 2020. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

[133] S. Parida, O. Bojar, and S. R. Dash, “Hindi Visual Genome: A Dataset for Multimodal English-to-Hindi Machine Translation,” *Computación y Sistemas*, vol. 23, no. 4, pp. 1499–1505, 2019. Presented at CICLing 2019, La Rochelle, France.

[134] G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. Rush, “Opennmt: Open-source toolkit for neural machine translation,” in *Proceedings of ACL 2017, System Demonstrations*, (Vancouver, Canada), pp. 67–72, Association for Computational Linguistics, July 2017.

[135] H. Isozaki, T. Hirao, K. Duh, K. Sudoh, and H. Tsukada, “Automatic evaluation of translation quality for distant language pairs,” in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, (Cambridge, MA), pp. 944–952, Association for Computational Linguistics, Oct. 2010.

[136] V. Khenglawt, S. R. Laskar, D. P. Pakray, R. Manna, and A. Khan, “Machine translation for low-resource english-mizo pair encountering tonal words,” *Computación y Sistemas*, vol. 26, 09 2022.

[137] S. Sen, M. Hasanuzzaman, A. Ekbal, P. Bhattacharyya, and A. Way, “Neural machine translation of low-resource languages using smt phrase pair injection,” *Natural Language Engineering*, vol. 27, no. 3, p. 271–292, 2021.

[138] S. R. Laskar, A. F. Ur Rahman 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, 2022.

[139] C. Baziotis, B. Haddow, and A. Birch, “Language model prior for low-resource neural machine translation,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (Online), pp. 7622–7634, Association for Computational Linguistics, Nov. 2020.

[140] M. Junczys-Dowmunt, R. Grundkiewicz, T. Dwojak, H. Hoang, K. Heafield, T. Neckermann, F. Seide, U. Germann, A. F. Aji, N. Bogoychev, A. F. T. Martins, and A. Birch, “Marian: Fast neural machine translation in C++,” in *Proceedings of ACL 2018, System Demonstrations*, (Melbourne, Australia), pp. 116–121, Association for Computational Linguistics, July 2018.

[141] M. R. et al., “No language left behind: Scaling human-centered machine translation,” 2022.

BIO-DATA OF THE CANDIDATE

Name of Candidate	:	Vanlalmuansangi Khenglawt
Date of Birth	:	19/07/1988
Phone	:	8794480408
email	:	mzut208@mzu.edu.in
Permanent Address	:	H.No: S-5, Dawrpui, Bazar Bungkawn, Aizawl, Mizoram Pin : 796001
Married	:	Yes
Educational Details		
(a) M.Tech	:	IIT Guwahati
(b) Ph.D Course Work	:	Mizoram University
Present Occupation	:	Assistant Professor Department of Information Technology
Organization	:	Mizoram University

LIST OF PAPERS BASED ON THESIS

Journals:

1. **Vanlalmuansangi Khenglawt**, Sahinur Rahman Laskar, Partha Pakray, Riyanka Manna, Ajoy Kumar Khan "Machine translation for low-resource English-Mizo pair encountering tonal words", *Computación y Sistemas*, vol. 26, no. 3, pp. 1377-1398, 2022, DOI: 10.13053/cys-26-3-4358. (ESCI/Scopus)
2. **Vanlalmuansangi Khenglawt**, Sahinur Rahman Laskar, Partha Pakray, Ajoy Kumar 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, 2024, DOI: 10.3233/JIFS-235740. (SCI)
3. **Vanlalmuansangi Khenglawt**, SR Laskar, Partha Pakray, Ajoy Kumar Khan, "System For Preparing And Investigating An English–Mizo Corpus", *German Patent*, IPC: G06F 40/47, 2022

Conference:

1. **Vanlalmuansangi Khenglawt**, Sahinur Rahman Laskar, Santanu Pal, Partha Pakray, Ajoy Kumar Khan,"Language resource building and English-to-mizo neural machine translation encountering tonal words", *In Proceedings of the WILDRE-6 Workshop within the 13th Language Resources and Evaluation Conference*, Marseille, France. European Language Resources Association, 2022, pp 48–54.
2. **Vanlalmuansangi Khenglawt**, Sahinur Rahman Laskar, Riyanka Manna, Partha Pakray, Ajoy Kumar Khan, "Mizo Visual Genome 1.0: A Dataset for

English-Mizo Multimodal Neural Machine Translation", *2022 IEEE Silchar Subsection Conference (SILCON)*, Silchar, India, 2022, pp. 1-6, doi: 10.1109/SILCON55242.2022.10028882.

3. **Vanlalmuansangi Khenglawt**, Sahinur Rahman Laskar, Riyanka Manna, Partha Pakray and Ajoy Kumar Khan, "Recent Trends on Low-Resource Neural Machine Translation and Research Scope for English-Mizo Pair" *International Conference on Intelligent Computing Systems and Applications*, National Institute of Technology Silchar Conference, 2022. (Accepted)

4. Santanu Pal, Partha Pakray, Sahinur Rahman Laskar, Lenin Laitonjam, **Vanlalmuansangi Khenglawt**, Sunita Warjri, Pankaj Kundan Dadure and Sandeep Kumar Dash, "Findings of the WMT 2023 Shared Task on Low-Resource Indic Language Translation" *Proceedings of the Eighth Conference on Machine Translation*, Association for Computational Linguistics, Singapore, pp 682–694.

PARTICULARS OF THE CANDIDATE

NAME OF CANDIDATE	: VANLALMUANSANGI KHENGLAWT
DEGREE	: Ph.D
DEPARTMENT	: COMPUTER ENGINEERING
TITLE OF THE THESIS	: ENGLISH MIZO LANGUAGE PAIRS: AUTOMATIC MACHINE TRANSLATION
DATE OF ADMISSION	: 31.07.2019
APPROVAL OF RESESARCH PROPOSAL	
1. DRC	: 28.10.2020
2. BOS	: 29.10.2020
3. SCHOOL BOARD	: 05.11.2020
MZU REGISTRATION NO	: 1906313
PH.D REGISTRATION NO	: MZU/Ph.d./1585 of 31.07.2019
EXTENSION	: NA

(Dr. V.D.AMBETH KUMAR)

Head

Department of Computer Engineering

ABSTRACT

ENGLISH MIZO LANGUAGE PAIRS : AUTOMATIC MACHINE TRANSLATION

AN ABSTRACT SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

VANLALMUANSANGI KHENGLAWT

MZU REGN NO : 1906313

Ph. D REGN NO : MZU/Ph.D./1585 OF 31.07.2019

DEPARTMENT OF COMPUTER ENGINEERING
SCHOOL OF ENGINEERING AND
TECHNOLOGY

JULY, 2024

ENGLISH MIZO LANGUAGE PAIRS : AUTOMATIC MACHINE TRANSLATION

BY

VANLALMUANSANGI KHENGLAWT
DEPARTMENT OF COMPUTER ENGINEERING

Supervisor : Prof. (Dr.) Ajoy Kumar Khan
Joint-Supervisor : Dr. Partha Pakray

Submitted

In partial fulfillment of the requirement of the Degree of Doctor of Philosophy in Computer Engineering of Mizoram University, Aizawl.

ABSTRACT

With technological advances, Machine Translation (MT) significantly influences today's society, bridging the gap between languages. Language is a medium of communication for varied cultures around the world. The language barrier prevents communication between different cultures. The advent of MT systems has functioned as a substitute for professional human translators, offering instant and instantaneous translations. It removes human intervention from translating one natural language to another using automatic translation, thereby resolving linguistically ambiguous problems. MT uses computer software to translate text or speech from one natural language into another. Languages are frequently classified as high-resource or low-resource in MT, depending on the availability of linguistic data and resources for training translation models. Even though it substantially impacts high-resource languages, it is especially beneficial for low-resource languages since it resolves barriers to communication. MT can also help preserve low-resource languages by enhancing automatic translation capabilities. Through continuous improvement and developing translation models tailored to specific languages, MT can contribute to their requirements, applications, and survival in the digital age. Although conventional human translations remain unrivaled in accuracy, MT systems have significantly improved translation accuracy and fluency, offering reduced costs and faster turnaround times. Ongoing advancements and the ability to incorporate a multilingual system for cross-language communication further enhance the appeal of MT.

Throughout history, numerous languages have become extinct for various reasons. One of the factors can be due to rapid changes in the advancement of different technologies, where globalization favors dominant languages like English. Another

cause can also be negligence by the native people, where their language is given less priority. When younger generations do not correctly inherit the language, it can quickly become extinct. Spoken languages without written forms are more likely to go extinct. However, low-resource language becomes endangered of extinction when a minority uses it. As the language becomes extinct, the culture dies along with it. Extinction erases unique cultural knowledge and traditions. Therefore, preserving language from extinction is highly necessary, especially for low-resource languages.

Mizo language is a Tibeto-Burman language spoken primarily by the Mizo people in northeastern India, particularly in the state of Mizoram. It is considered a low-resource language since there is limited availability of resources. Therefore, preservation is imperative. With MT as one of the most powerful natural language processing (NLP) applications for preserving and upgrading low-resource language, an English↔Mizo language pair translation will significantly impact the Mizo language, as English is considered the most dominant language. Additionally, the difference in the linguistic information between the language pair also substantially impacts the dataset creation for improving translation accuracy. MT techniques yield better results when translating closely related language pairs than those with significant structural diversity. In the case of the English↔Mizo language pair, differences in language origin, word order, and gender distinction present significant complications due to their distinct linguistic characteristics. Furthermore, Mizo is a tonal language, where a word can express different meanings depending on various tones. There are four variations of tones, namely high, low, rising, and falling. A tone marker represents each tone, added to the vowels to indicate tone variation. In comparison, English is a non-tonal language. Addressing tonal words in MT for such a low-resource pair is another challenging issue.

Various techniques exist to handle the problem of low-resource languages in MT. Linguistic constraints of a language are one of the critical considerations. Therefore, identifying appropriate methodologies and tackling specific language issues is necessary to improve low-resource pair translation. Hence, the linguistic challenges of the low-resource English↔Mizo pair and various existing low-resource NMT ap-

proaches are surveyed. Here, four techniques have been proposed that benefit the advanced automatic MT of the English↔Mizo language pair.

In the first approach, the MT system for English↔Mizo pairs is developed by encountering tonal words, as Mizo is a tonal language. A few studies in MT for English↔Mizo pairs have been explored. However, no prior work is available that encounters Mizo tonal words in low-resource English↔Mizo pair translations. Translating low-resource tonal languages presents unique challenges in the realm of MT. Accurate translation for tonal languages requires handling tonal distinctions and understanding contextual cues. Therefore, addressing low-resource and tonal complexities requires unique strategies to improve translation quality. The English↔Mizo corpus consists of parallel sentences with tonal words. For baseline systems, different MT models are explored, such as Phrase-Based Statistical Machine Translation (PBSMT), Recurrent Neural Network (RNN), and Bidirectional Recurrent Neural Network (BRNN). The proposed approach generates a synthetic parallel corpus from the Mizo tonal sentence extracted from the Mizo monolingual sentence. To improve the Mizo tonal word's translation quality, the synthetic parallel corpus is augmented with the original parallel corpus, injecting more tonal word information into the corpus. The proposed approach attempts to train the augmented data using the best-trained baseline model (BRNN). Automatic evaluation metrics and human evaluation (HE) are considered to evaluate the predicted sentence of the proposed approach. For En-to-Mz translation, the BLEU score attains 20.21, and the overall rating for HE is 32.24%. For Mz-to-En translation, the BLEU score attains 20.31, and the overall rating for HE is 33.48%. The proposed approach outperforms the baseline systems for both forward (En-to-Mz) and backward (Mz-to-En) translations by encountering tonal words.

In the second approach, addressing low resource challenges in English↔Mizo MT necessitates the development of resource languages. Building a resource language in MT provides essential training data and is the foundation for developing accurate and reliable translations across diverse domains and languages. A multilingual country like India has an enormous linguistic diversity. The increasing demand

for developing language resources extends to various NLP applications, including MT. This approach investigates a low-resource English-to-Mizo language pair by building a language resource, i.e., the English↔Mizo corpus, thereby contributing to an Indian language resource. The corpus consists of both parallel and monolingual data of Mizo. An English-to-Mizo NMT system was proposed, utilizing a synthetic parallel corpus alongside the original dataset, both enhanced through training with a BERT-fused transformer NMT model. Various post-processing steps are employed to handle tonal words in the Mizo language as they add to the complexity on top of low-resource challenges for any NLP task. Combining the Bert-fused transformer model with bidirectional and synthetic parallel corpus with the post-processing step attains the best BLEU score of 28.59. This approach enhances translation accuracy by building language resources and effectively addressing the tonal words in Mizo, achieving state-of-the-art results in English-to-Mizo translation.

In the third approach, the multimodal notion, which combines textual and visual aspects, has been introduced. Multimodal Machine Translation (MMT) handles extracting information from several modalities, considering the presumption that the extra modalities will include beneficial alternative perspectives of the input data. Traditional MT systems primarily focus on text-to-text translation, but multimodal MT integrates various data sources to provide more context, enriching low-resource language with different modalities. Regardless of its significant benefits, it is challenging to implement an MMT system for several languages, mainly due to the scarcity of the availability of multimodal datasets. As for the low-resource English↔Mizo pair, the standard multimodal corpus is not available. As a result, the Mizo Visual Genome 1.0 (MVG 1.0) dataset has been developed for English↔Mizo multimodal machine translation (MMT). It comprises images paired with corresponding bilingual textual descriptions. With the BRNN model, the En-to-Mz translation achieves a BLEU score of 7.39 in the multimodal system and a BLEU score of 6.18 in the text-only system. The Mz-to-En translation achieves a BLEU score of 10.03 in the multimodal system and a BLEU score of 8.84 in the text-only system. Automated assessment metrics indicate that multimodal neural machine translation (MNMT) outperforms traditional text-only NMT. To the best of current knowledge, the English↔Mizo MMT system is the pioneering work in this approach, serving as a baseline for future studies in MMT for the low-resource English↔Mizo language pair.

In the fourth approach, the English↔Mizo NMT System is designed to tackle the limitation of the language pair by addressing the data scarcity issue and handling the tonality of the Mizo language in MT. Low-resource language in MT systems poses multiple complications regarding accuracy in translation due to insufficient incorporation of linguistic information. The system is designed to increase the training amount of data and provide token alignment information via phrase pair augmentation. The augmentation of synthetic parallel data and phrase pairs handles the data scarcity problems. Integrating a pre-trained language model (LM) has also enhanced English↔Mizo translation. Different transformer-based NMT models were explored to address data scarcity issues and the tonal word problem in English↔Mizo pair translation. The model with a combination of the original parallel sentence with the synthetic sentence and phrase pair with pre-trained LM attains the best BLEU score for both directions, i.e., BLEU score of 32.54 for En-to-Mz and BLEU score of 30.26 for Mz-to-En translation. This approach yields the best translation accuracy compared to the former experiments. State-of-the-art results have been achieved on the English-to-Mizo and Mizo-to-English translation test data.

In conclusion, improving English↔Mizo MT entails overcoming limited data and linguistic complexities through innovative approaches like transformer-based models and synthetic data augmentation. Tackling the limitation of a specific language in MT results in better translation quality and improves overall performance. Additionally, collaboration with native speakers and linguists aids in capturing cultural and contextual nuances and refining translation outputs. Finally, iterative refinement and adaptation to evolving linguistic patterns in both languages are crucial for achieving and sustaining high-quality translation in the English↔Mizo MT domain.