ISSN 2007-9737

Machine Translation for Low-Resource English-Mizo Pair Encountering Tonal Words

Vanlalmuansangi Khenglawt[1], Sahinur Rahman Laskar[2], Partha Pakray[2], Riyanka Manna[3], Ajoy Kumar Khan[1]

[1] Mizoram University,
Department of Computer Engineering,
India

[2] National Institute of Technology,
Department of Computer Science and Engineering,
India

[3] Gandhi Institute of Technology and Management,
Department of Computer Science and Engineering,
India

mzut208@mzu.edu.in, {sahinur_rs, partha}@cse.nits.ac.in,
{riyankamanna16, ajoyiitg}@gmail.com

Abstract. Machine translation is one of the most powerful natural language processing applications for preserving and upgrading low-resource language. Mizo language is considered as low-resource since there is limited availability of resources. Therefore, it is a challenging task for English-Mizo language pair translation. Moreover, Mizo is a tonal language, where a word can express different meanings depending on a variety of tones. There are four variations of tones, namely high, low, rising, and falling. A tone marker is used to represent each of the tones, which is added to the vowels to indicate tone variation. Addressing tonal words in machine translation for such a low-resource pair is another challenging issue. In this paper, the English-Mizo corpus is developed where parallel sentences having tonal words are incorporated. The different machine translation models are explored based on statistical machine translation and neural machine translation for the baseline systems. Furthermore, the proposed approach attempts to augment the train data by expanding parallel data having tonal words and achieves state-of-the-art results for both forward and backward translations encountering tonal words.

Keywords. English-Mizo, machine translation, low-resource, tonal.

1 Introduction

Language is a method of communication for individuals of varied cultures everywhere in the world. The language barrier prevents communication between different cultures. Machine translation (MT) commonly uses to address the problem and serves as a bridge for language barriers among people of divergent linguistic backgrounds.

In MT, one natural or human spoken language translates into another natural or human spoken language. Natural language is of 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 classify into the low-resource category that has restricted online resources [25, 32]. Moreover, a low-resource language pair is considered based on the minimal amount of data required for training a model [9].

The proper definition of low-resource language pair puts forward a challenging research question itself. However, if the training data is under 1 million parallel sentences, it is considered a low-resource language pair [12]. The native speakers play a vital role in different aspects of the language, including the quality and quantity of the data.

Most of the world languages are recognized under the low-resource category based on the availability of resources. The MT works are limited in India's north-eastern region, and the languages considered as low-resource languages include Assamese, Boro, Manipuri, Khasi, Kokborok, and Mizo.

1.1 Low-Resource Pair: English–Mizo

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. Mizoram is one of the states of India, situated in the northeastern parts of the country.

It shares borders with three states in northeast India: Tripura, Assam, and Manipur. Additionally, the state also shares a border with two of the neighbouring 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' [27]. It holds the second least populated state with a population of 830,846 according to the 2011 Census of India[2]. The Mizo language [24, 27] is mainly based on the *Lusei* dialect and many words are also derived from its surrounding Mizo sub-tribes and sub-clan.

The writing system of the Mizo language is based on the Roman script. The Mizo alphabet has 25 letters including 3 letters with a combination of two letters represented as one letter: *AW, CH, NG*. Among the alphabets there are six vowels which are *A, AW, E, I, O, U*. A circumflex ∧ was subsequently added to the vowels to demonstrate long vowels, which were inadequate to completely express Mizo tone. A vowel is a syllabic language unit pronounced with no stricture within the vocal tract. Each of the vowels has its meaning by itself, and they represent the tone of a word. All the other alphabets are a consonant which has no meaning by itself but can be merged to form a syllable with a vowel.

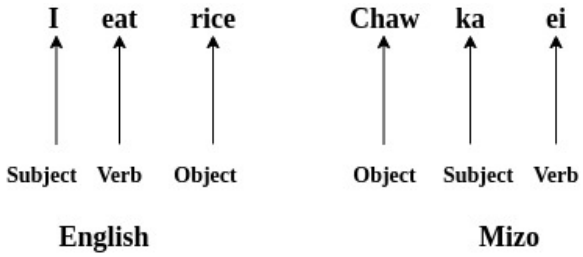

Fig. 1. Example of English–Mizo word order

A consonant is a speech tone that is articulated with a complete or partial closure of the vocal tract in articulatory phonetics. Unlike English, Mizo is a tonal language, where the lexical meaning of words is influenced by the pitch of a syllable. The structure and the word order of the Mizo Language are also different from the English language, the declarative word order of Mizo is OSV (object-subject-verb).

Fig. 1 presents an example of a Mizo sentence with its translation in English. In contrast to English, all proper names have a gender suffix in Mizo, a letter *'i'* is added at the end of every female proper name and a letter *'a'* is added at the end of every male proper name to distinguish the gender.

In terms of pronoun, there is no distinction between gender in Mizo while there is a clear distinction of gender when using the pronoun in English. However, Mizo uses the same number system like English. The English–Mizo can be considered as a low-resource pair based on the limited availability of resources, namely, parallel corpus and monolingual data of Mizo.

1.2 Motivation

Beneath every language, there is a culture involved. A language is defined by the people living in the area, their origin, traditions, custom, cuisine, and many more.

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

[2] https://www.censusindia.gov.in/2011Census/Language_MTs.html

Table 1. Variation of tone (a) in Mizo

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

Therefore, a language not only means for communication, but defines the people using them. There are many languages across the world which are extinct. It can be due to rapid change in the advancement of different technologies where there is a requirement for a dominant language like English language.

The reason can also be negligence by the native people, where their language is given less priority. It can get easily extinct when the language is not properly passed on to the younger generations. Spoken languages without written form are more likely to get extinct.

However, it is also endangered as a low-resource language where a minority of the population uses the language. As the language becomes extinct, the culture dies along with it. Therefore, the preservation of language from extinction is highly necessary, especially for low-resource languages.

With Mizo language being a low-resource language, it is imperative for preservation. Machine translation is capable of preservation of low resource language as it breaks language barriers. Since English language is considered to be the most dominant language, English-Mizo machine translation can enhance the limitations of the Mizo language in today's digital world.

Therefore, a low-resource language like the Mizo language has a chance for survival and is capable of encountering technological advances with Machine Translation. There are very limited machine translation works on English–Mizo pair [30, 15], that lags in encountering tonal words of Mizo. Apart from this, automatic translations like Google and Bing cover 109 and 70 languages across the globe, but the Mizo language lags in.

This is due to the lack of standard corpus. In this paper, we have considered machine translation work of English–Mizo pair by encountering challenges of Mizo tonal words. From the best of our knowledge, no prior work available that encounters Mizo tonal words in such low-resource English–Mizo pair translation. The contributions of this work are as follows:

— Detailed survey of linguistic challenges in English–Mizo machine translation.

— Created EnMzCorp1.0:English–Mizo corpus.

— Evaluated baseline systems for low-resource English–Mizo pair, encountering tonal words through different machine translation models.

— Proposed approach investigates with data augmentation technique and achieved state-of-the-art results for English–Mizo pair translation.

— Analysis is reported for inspecting errors on predicted translation.

2 Challenges of English–Mizo Machine Translation

Translation of a language is not a simple task. There are several challenges to be dealt with when translating one language to another. Like many other languages, the Mizo language deals with several challenges. This section has surveyed linguistic challenges.

2.1 Tonal Words

A language is treated as a tonal language when its tone influences the meaning of the word. Mizo language is undoubtedly a tonal language, which can lead to certain challenges for machine translation. Variation in tones and contour tones can alter the meaning of particular words.

The type of pitch used is capable of automatically determining 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 ways of vowel sound: long vowel and short vowel.

Table 2. Example of different meaning of the word *Buk* in Mizo

Mizo Word	Tone	English Meaning	Mizo	English
buk	High tone *(búk)*	Hut	Kan ramah **búk** sak ka duh.	I want to build hut in 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 weight 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 chŏ.	They dig up the ground to plant seeds.
	Rising tone *(lěi)*	Buy	Thil ka **lěi**.	I am buying something.
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.

However, the Mizo tone framework accepts four tones: High (H), Low (L), Rising (R), and Falling (F) [6]. The tones are also named in Mizo as *'Ri sang'*, *'Ri hniam'*, *'Ri lawn'* and *'Ri kuai'* respectively. Linguist had created a tone-marker for each of the tone to indicate the tone variation in the Mizo Language, which are listed in the following Table 1.

The four different tones used in Mizo words can indicate different meanings in the English word, as shown in Table 2. For 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. The Mizo is undeniably a tonal language where a change in tone will completely alter a word's meaning.

However, in the writing system, the indication of tonal words is neglected and not correctly considered. Most of the writings in Mizo use only circumflex $\wedge$ for indication of tone. Furthermore, it is also an understudied language with a limited resource in terms of tones. Based on the four tones applicable to the five vowels (a, e, i, o, u), we have identified $4 \times 5 = 20$ possible types that exist in Mizo.

2.2 Tonal-Polysemy Words

Mizo language is also rich in polysemy words where the intonation is the same, yet its meaning is different. Polysemy is a side of linguistics ambiguity that considerations the multiplicity of word meanings. Table 3 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 natural language processing task. As tonal languages go, the Mizo language is one of the most complicated languages. It is a tonal language where not only a particular word has several tones, but also it is a language in which the pitch of the word 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 pitch of the word alone cannot define the word. Therefore, a complete understanding of the nearby word or understanding the whole sentence's context is necessary. A few polysemy words in the Mizo language can also act as both noun and verb. For example:

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

— Ruah a sur dawn sia, **púk** ah hian awm mai ang u **(Noun)**
I pawisa ka lo **púk** ang e, I phal em? **(Verb)**.

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

— **lěi - Buy** (Raising tone)

  – Thil ka **lěi** → I am buying something. (Raising tone)
  – Khawiah nge I **lêi?** → where did you buy? (sound as falling tone)

— **áng – will** (High tone)

Table 3. Example of tonal-polysemy words in Mizo

Tonal-Ploysemy	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?

- 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)

— **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).

2.3 Symbolic Words

Apart from the tone, a few symbols are used in the writings of the Mizo sentence. In many places, – (hyphen) is found, used for continuing English (non-Mizo) word with Mizo word to appear as one word. It is used after figures.

Another famous symbol is *’n*, which is used after the noun to show possession with the noun. It works as putting (apostrophe) in the English sentence. Table 4 demonstrates symbolic words in Mizo. Moreover, there are a few words that 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).

We have tackled the above challenges in two ways. First, we have extracted Mizo tonal and symbolic words from the monolingual corpus of Mizo. Then, manually translated into corresponding English words.

Secondly, Mizo tonal sentences are extracted from monolingual data. Then, the best-trained baseline model (Mizo to English) is applied to generate pseudo-English sentences.

To improve the Mizo tonal word’s translation quality, we have augmented the parallel train data by injecting more tonal word information. The data statistics and proposed approach are described in Sect. 5 and 7.

3 Machine Translation

Machine translation removes human intervention from a translation of one natural language to another using automatic translation, thereby resolving linguistically ambiguous problems. It is divided into two broad categories: rule-based and corpus-based approaches. The knowledge-driven approach is another name for a rule-based approach based on the linguistic information of the language.

The rule-based translation system is built using a set of grammatical rules and linguistic experts. 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

Table 4. Example of symbolic words in Mizo

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

structures. 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 in NLP techniques. Example-based Machine translation (EBMT), statistical machine translation (SMT), and neural machine translation (NMT) are the three methods of corpus-based machine translation.

The EBMT requires a parallel corpus, and the central concept is text similarity. It identifies the approximately matching sentences (i.e., examples) using a 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.

For a given input sentence, the retrieval module finds identical parallel sentences from the corpus.

The adaptation module determines the parts of translation to be reused from the retrieval module.

The relevant match concerning the source language is used in case it does not match. The two most common corpus-based MT are SMT and NMT, which are described in the following subsections.

3.1 SMT

In the corpus-based approach, the main drawback of EBMT is that in real-time scenarios, we can not cover various types of sentences by examples only. To encounter this issue, statistical machine translation (SMT) is introduced [14, 13].

In this approach, a statistical model in which 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.

Out of which phrase-based translation is the most widely used. Before NMT, phrase-based SMT achieves a state-of-the-art approach. 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. 1:

$$m_{best} = \arg\max_{m} P(m \mid e). \tag{1}$$

To estimate $P(m \mid e)$ for the given source-target sentences, the probability distribution of all possible target sentences is required, which is achieved by understanding what makes a good translation. Any good translation should possess two aspects: adequacy and fluency.

The target sentence should keep the same meaning as the source sentence, which is known as adequacy, and the target sentence should be fluent. Both adequacy and fluency factors must be balanced to yield a good translation. This can be formulated based on Bayes Theorem by the extension of Eq.1 as shown in Eq.2:

$$\begin{aligned} m_{\text{best}} &= \arg\max_{m} \ \text{adequacy}(e \mid m) \times \text{fluency}(m), \\ &= \arg\max_{m} P(e \mid m) \times P(m). \end{aligned} \tag{2}$$

In the SMT, TM and LM are used to 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 used to estimate the probable target words/phrases as shown in the Eq.3:

$$P(e \mid m) = \frac{\text{count}(e,\ m)}{\sum_e \text{count}(e,\ m)}. \tag{3}$$

The LM reorders the obtained target words/phrases from TM to predict syntactically correct target sentences for ensuring fluency of translation.

The LM is estimated from monolingual target data, where the target sentence is modelled by the conditional probability of each word given the previous words in the sentence. This modelling is also known as n-gram LM. Lastly, the decoder utilizes a beam search strategy to find out the best possible translation. The abstract pictorial representation of SMT is shown in Fig. 2.

3.2 NMT

In the MT task, the NMT approach attains state-of-the-art for both high and low resource pair translations [1, 30, 29, 18, 21]. NMT can learn the model in an end-to-end manner by mapping the source and target sentence.

The main problem with SMT is that SMT creates a model context by considering a set of phrases of limited size. As the phrase size increases, the data sparsity will reduce the quality.

Likewise, feed-forward based NMT calculates the phrase pairs score by considering the length of the fixed phrases. But in real-time translation, the phrase length of both source and target are not fixed. Therefore, recurrent neural networks (RNN) based NMT [5, 4] 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. Also, RNN updates and maintains a memory known as a state during the processing of each word.

The Eq. 4 represents probability of a sentence S and $S_1, S_2, ..S_n$ denotes a sequence of n words. The RNN based LM [26] can be represented by considering the Equation 5, where next word S_{t+1} is predicted for the given current word S_t and previous words $S_1....S_{t-1}$:

$$\begin{aligned} P(S) &= P(S_1, S_2....Sn), \\ &= \prod_{t=1}^{n} P(S_t \mid S_1....S_{t-1}), \end{aligned} \tag{4}$$

$$\begin{aligned} S_{t+1} &= \arg\max_{t+1}(P(S_{t+1} \mid S_1....S_t)), \qquad (5) \\ &= \arg\max_{t+1}(p_t), \\ p_t &= \text{softmax}(y_t), \qquad (6) \\ y_t &= h_t^1 W_0, \qquad (7) \\ h_t^1 &= \tanh([h_t^0; h_{t-1}^1]W_h), \qquad (8) \\ h_t^0 &= S_t E. \qquad (9) \end{aligned}$$

The RNN based LM processes each word in a sentence at every time step t to predict the next word. Here, consider the vocabulary size and all hidden layers as 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 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, and for all other positions is denoted by a "0".

This helps to create the embedding through the multiplication of the one-hot vector having size $1 \times V$ with the embedding matrix of size $V \times M$. The RNN based LM maintains memory using a hidden state, h_{t-1}^1 called as the previous state. When the first word encounters in the sequence, the previous state is set to all zeros' vector.

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, softmax function is used to transform the vector y_t (also knows 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) [10] or gated recurrent unit (GRU) [3] 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 unable to capture all important information if the sequence is too long.

Therefore, the attention mechanism comes into existence [1, 23] 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. 3 depicts attention-based RNN, where the input Mizo sentence *"Thil ka lei˘"* is translated into the target English sentence *"I am buying something"*. The drawback of RNN is that input processing follows in a strict temporal order, which means it computes context in one direction based on preceding words, not on future words. RNN impotent to look ahead into future words. BRNN (Bidirectional RNN) [1] resolves this issue by utilizing two distinct RNNs, one for the forward direction and another for the backward direction.

In [33], a BRNN based model improves translation accuracy on low-resource pairs like English–Hindi, English–Tamil. Moreover, the convolutional neural network (CNN) based NMT is introduced [11, 8] by taking advantage of parallelizing operation and considering relative positions of the tokens instead of the temporal dependency among the tokens of the sequence.

But it lags behind features of RNN to enhance the encoding of the source sentence. The demerits of CNN-based approaches require many layers to hold long-term dependency, making the network large or complex without ever succeeding, which seems to be impractical. To handle such an issue, a transformer-based NMT comes in [38].

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, which is known as multi-head attention.

However, both SMT and NMT require minimal training data to provide a promising result, 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.

4 Related Work

This section focuses on existing MT work done on English–Mizo and other low-resource pairs. In MT, there are limited existing works available for English–Mizo pair [30, 15, 16]. A comparative study [30] in English to Mizo translation was performed between SMT and NMT, where NMT outperforms SMT.

In [15, 16], various attention-based NMT models, RNN and BRNN, have been examined in English to Mizo translation with parallel data only. Monolingual data is not incorporated to improve such low-resource pair translation. Furthermore, no previous work is found that focuses on tonal words of Mizo in MT in both directions of translation, i.e., English to Mizo and vice versa.

Besides, MT related works include recognizing named entity classes [2], Multi-word Expressions (MWE) [24], and resource building and POS tagging for Mizo language [27]. The NMT has been investigated with RNN for low-resource pairs like English to Punjabi, English to Tamil, and English

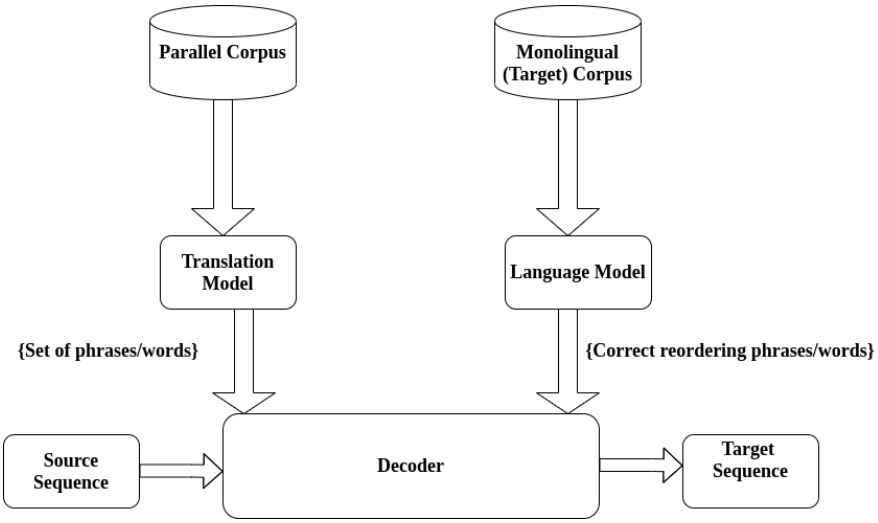

Fig. 2. Abstract diagram of phrase-based SMT

to Hindi and observed that performance increases with an increase in parallel train data [29].

In [34, 18], English to Hindi translation on the benchmark dataset, the NMT shows promising results. 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 [41].

Moreover, among similar language pair translations in WMT19, NMT systems attained remarkable performance on Hindi-Nepali [20]. With monolingual data to address the low-resource language problem, a filtering approach for the pseudo-parallel corpus is proposed to increase the parallel training corpus.

Despite achieving state-of-the-art performance in various language pairs, the NMT demands parallel corpus, which is a big challenge in low-resource pairs. To address this issue, a monolingual data-based NMT has been introduced without modifying system architecture [35].

By applying BackTranslation (BT) on low resource language monolingual data, the low-resource target sentences can be generated 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 with original parallel data [40]. In the context of low resource tonal language like Burmese with English pair, NMT with BT strategy shows remarkable performance [39].

Moreover, unsupervised pre-train based NMT is introduced [37, 17], where monolingual data of both source and target sentences are pre-trained and then fine-tuned the trained model with original parallel data.

5 EnMzCorp1.0: English–Mizo Corpus

The low-resource English–Mizo (En-Mz) pair has limited available options for parallel and monolingual data of Mizo. We have explored different viable resources to prepare the corpus, which discusses in the following subsections.

5.1 Corpus Details

We have prepared an En-Mz parallel corpus that contains a total of 130,441 sentences. Also, monolingual data of Mizo is prepared. The parallel corpus is collected from various online sources

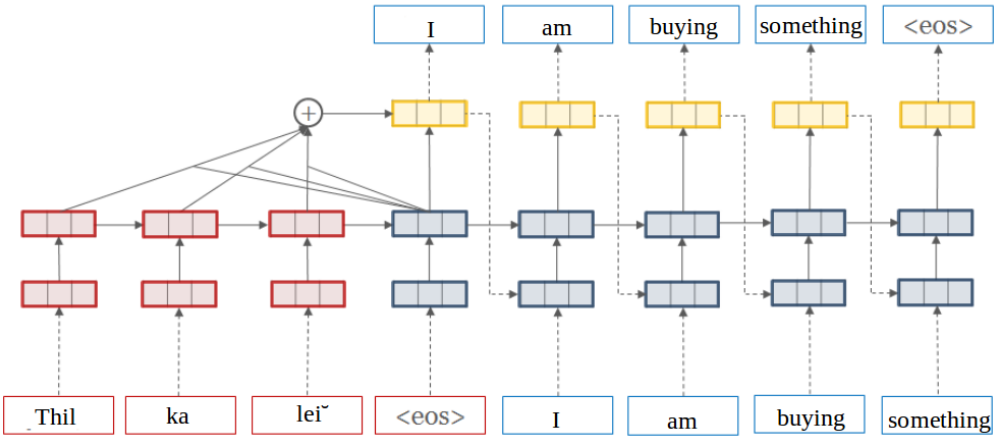

Fig. 3. English–Mizo NMT system (attention-based RNN)

namely, Bible[3], online dictionary (Glosbe)[4], Government websites[5] [6] and different web pages / blogs. Table 6 presents the corpus of sources with statistics, and Table 5 demonstrates example sentences collected from various sources. In Table 5, tonal words in the sentences are marked as bold.

5.2 Corpus Extraction Approaches

We have used a web crawling technique, namely Scrapy, which is an open-source framework. 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/image files, Google OCR[7] tool is used.

It is mainly used to extract Mizo data from text book[8] (Government website) and Table 8 presents the extracted data statistics for the same. Fig. 4 depicts the overall data acquisition. Moreover, we have used manual effort to prepare parallel data, mainly government websites extracted data.

[3] https://www.bible.com/
[4] https://glosbe.com/en/lus
[5] https://finance.mizoram.gov.in/
[6] https://dipr.mizoram.gov.in/
[7] https://cloud.google.com/vision/
[8] https://scert.mizoram.gov.in/page/english-medium

From the monolingual data of Mizo, tonal and symbolic words are extracted and translated manually to their corresponding English words. The manual process alignment took a period of 2 to 3 months by the first author. Moreover, the Mizo sentences are cross-verified by hiring a linguistic expert of Mizo, who is a native speaker and possesses linguistic knowledge of Mizo.

5.3 Data Cleaning and Split

The prepared corpus contains noise like too many special characters, web-link (URLs), blank lines, and duplicates. Therefore, we have removed noise and the duplicate sentences, the total number of parallel sentences reduced to 118,449.

During data cleaning, conversion of lower-case and removal of punctuation is not performed as in [19] to maintain the semantic contextual meaning. Table 7 presents the split data for the train, validation, and test data.

During the partition of validation and test data, we have considered those sentences which have tonal words. We have also considered two test sets, namely Test Set-1 for in-domain data from the split data and Test Set-2 for out-domain data that includes different types of tonal words having maximum length of 15 words, which we have prepared manually.

Table 5. Example of parallel and monolingual sentences

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?	Glosbe
	What is humility?	**Inngaihtlâwmna** chu eng nge ni?	
	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.	Government Website
	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.	
	advance	**hmasâwn**	Tonal Word (Manually Prepared)
	punch	**hnék**	
	At Famous	'Famous'-ah	Symbolic word Manually Prepared)
	God for ever	kumkhua-in—Pathian	
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 6. Corpus sources and statistics

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	402,90	653,65
	Tonal and Symbolic words (Manually Prepared)	2,341	2,341	2,341
	Total	130,441	2,165,169	2,608,458
Monolingual	Web Pages/Blogs/Text Book	1,943,023	-	25,813,315

Following [19], we have considered small test data in comparison to training data because it is used for the baseline system. In the train data, out of 115,249 Mizo sentences, 44,604 sentences have tonal words.

5.4 Domain Coverage

Our corpus EnMzCorp1.0 covers various domains: Bible, daily usage, Government messages/notices, elementary textbook, dictionary, and general-domains.

6 Baseline System

We have considered phrase-based SMT [14] and sequence-to-sequence model-based NMT (recurrent neural network (RNN), bidirectional RNN (BRNN)) for baseline systems to provide benchmark translation accuracy for both the directions of translations in English-Mizo pair. We have utilized our EnMzCorp1.0 dataset and monolingual data of English (3 million sentences) from WMT16[9].

6.1 Experimental Setup

We have followed SMT and NMT setup by employing Moses[10] and OpenNMT-py[11]toolkit respectively. The SMT and NMT setup is used for

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

[10] http://www.statmt.org/moses/

[11] https://github.com/OpenNMT/OpenNMT-py

Table 7. Statistics for train, valid and test set

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

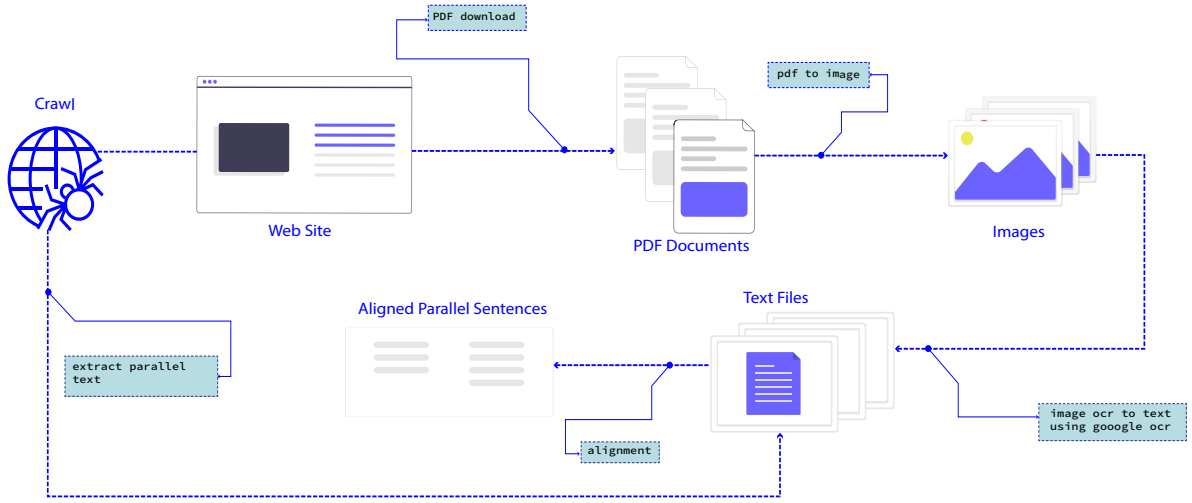

Fig. 4. Data acquisition

Table 8. Extracted Mz data using Google OCR

Monolingual Data	Sentences	Tokens
Mz	14,957	243,657

building phrase-based (PBSMT), RNN and BRNN based NMT systems.

For PBSMT, GIZA++ and IRSTLM [7] are utilized to produce phrase pairs and language models following the default settings of Moses. For RNN and BRNN, a 2-layer long short term memory (LSTM) network of encoder-decoder architecture with attention is used [1].

The LSTM contains 500 units at each layer. The Adam optimizer with a learning rate of 0.001 and drop-outs 0.3 is used in RNN and BRNN models. We have used unsupervised pre-trained word vectors of monolingual data using Glove[12] [31] and pre-trained up to 100 iterations with embedding vector size 200.

6.2 Results

To evaluate predicted sentences, automatic evaluation metrics and human evaluation are considered. The automatic evaluation metrics viz. bilingual evaluation understudy (BLEU) [28], translation edit rate (TER) [36], metric for evaluation of translation with explicit ordering (METEOR) [22] and F-measure.

6.2.1 BLEU

It utilizes the modified precision of n-gram by comparing the n-grams of the candidate

[12] https://github.com/stanfordnlp/GloVe

Table 9. BLEU scores of baseline systems

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

(predicted) translation with the n-grams of the reference translation. Eq. 10 represents the formula for the computation of the BLEU score. Here, P_l, R_l denote the length of the predicted and reference translation, respectively.

Pd_i represents precision score of i^{th} gram. To deals with too short translation, a brevity penalty is 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 [30]. In this work, $n = 3$ is considered because the BLEU score tends to zero while crossing the tri-gram score. Table 9 presents the BLEU scores for both the directions of translation:

$$\text{BLEU} = \min\left(1,\ \frac{P_l}{R_l}\right)\left(\sum_{i=1}^{n} Pd_i\right)^{\frac{1}{n}}. \qquad (10)$$

6.2.2 TER

It is an automatic metric used to calculate the number of actions required to update a candidate translation to align with the reference translation. It is a technique used in MT for measuring the amount of post-editing effort needed for the output of machine translation. TER is computed in Eq. 11 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}):

$$\text{TER} = \frac{N_{ed}}{L_{rw}}. \qquad (11)$$

Several possible edits include insertion, deletion, the substitution of single words, and shifts of word sequences. The cost of all the edits is the same. Consider the following scenario of candidate translation and reference translation where the mismatch is highlighted by italics:

— 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, TER score becomes $\frac{4}{11} = 36\%$. Lower the value of the TER score, accuracy will be good. Table 10 presents TER scores.

6.2.3 METEOR and F-measure

Meteor 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. 12 and 13:

$$P_{ug} = \frac{T_m}{p_n}, \tag{12}$$

$$R_{ug} = \frac{T_m}{r_n}. \tag{13}$$

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

Table 10. TER (%) scores of 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 11. METEOR scores of 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 12. F-measure 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

During the computation of METEOR score, F-measure score is calculated, which is the harmonic mean of precision P_{ug} and recall R_{ug} as shown in Eq. 14.

$$\mathrm{F-measure} = \frac{2 \times P_{ug} \times R_{ug}}{P_{ug} + R_{ug}}. \tag{14}$$

Also, F-mean is calculated by the parameterized harmonic mean of the precision P_{ug} and recall R_{ug}. Then, METEOR is computed using Eq. 15:

$$\mathrm{METEOR} = (1 - \mathrm{Pen}) \times F_{\mathrm{mean}}. \tag{15}$$

Here, fragmentation penalty (Pen) is calculated by fragmentation fraction (frag) and γ in Eq. 16 to account for the degree to which the uni-grams in both translations are in the same order. γ is the maximum penalty which is determined by the value ranges 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 the translations, is divided by the number of matches (m) as given in Eq. 17. METEOR and F-measure are assigned, ranging from 0 to 1 in each segment. Table 11 and 12 present METEOR and F-measure scores:

$$\mathrm{Pen} = \gamma \times \mathrm{Frag}, \tag{16}$$

$$\mathrm{Frag} = \frac{ch}{m}. \tag{17}$$

6.2.4 HE

Human evaluation (HE) is a manual evaluation metric that is used for evaluating the predicted sentence of the machine translation systems [30].

As automated evaluation metrics fail to assess all critical aspects of translation accuracy, the human evaluator with a linguistic expert has evaluated the predicted translation. The linguistic expert engaged in human evaluation is acquainted with both the Mizo and English language.

The expert is well-versed with the complexities and challenges of the Mizo language. Based on adequacy, fluency, and overall rating, a human evaluator evaluates the predicted translations. 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. By computing an average score of both adequacy and fluency, the overall rating is measured. Considering 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"*.

Here, the predicted translation is considered inadequate since it contains a different contextual meaning with the corresponding reference translation. The predicted sentence is also fluent; even though the meaning is entirely different from the reference translation, it is a well-formed sentence in the target language. The overall rating[13] considers the average of the adequacy as well as fluency.

Table 13. HE (Overall Rating (%)) scores of 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 14. Augmented train data statistics

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

The assessment criteria are measured on a scale of 1-5, with higher values indicating better performance [30]. The rating score is assigned for 50 predicted test sentences (randomly chosen). Table 13 reports human evaluation scores which are calculated using Eq. 18.

Where n_{TAR} is the total average rating scores of adequacy, and fluency. Here, n_{TBR} is calculated by multiplying best rating score with total number of questions, i.e., $5 \times 50 = 250$:

$$\mathrm{HE}(\mathrm{Overall\ Rating}) = \frac{n_{\mathrm{TAR}}}{n_{\mathrm{TBR}}} \times 100\%, \qquad (18)$$

7 Proposed Approach

Our proposed approach is based on BT [35] strategy without modifying the model architecture. It consists of three operations. First, extraction of Mizo sentences having tonal words from monolingual data of Mizo. Secondly, extracted Mizo tonal sentences are used to generate the English synthetic sentences via the best translation model (BRNN) of Mizo 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 for training the NMT model (BRNN) independently for each direction of translations. Fig. 5 depicts the pictorial diagram of the proposed approach. Since the original train data contains only 44,604 Mizo tonal sentences, we have extracted 44,000 Mizo tonal sentences by considering a maximum word length of 10.

However, we have removed blank lines, single word sentences from the synthetic English sentences and their corresponding Mizo sentences. Thus, the synthetic parallel corpus contains a total of 33,229 sentences as shown in Table 14. The evaluated results are reported in Table 16, 15.

8 Analysis

Among the automatic evaluation scores in Table 9, 10, 11, 12 and 16 on both test data, the proposed approach attains higher accuracy than baseline systems. The accuracy on Test Set-1 (in-domain) is better than Test Set-2 (out-domain).

It is noticed that Mizo to English translational evaluation scores outperforms English to Mizo. Because train data contains more Mizo tokens in comparison to English tokens, as mentioned in Table 7. Therefore the model encoded more Mizo word frequency, and the decoder can produce a better translation for 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, impotent in named-entity prediction and out-of-vocabulary issues. By the following notations, the samples of predicted sentences are considered below to inspect the errors.

[13] https://nlp.amrita.edu/mtil_cen/\#results

— 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)

Table 15. Human evaluation results of the 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 16. Automatic evaluation results of the proposed approach

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

— 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. But it is unable to detect the tone marker *è*. The word *"dig"* in the source sentence is predicted as *"laih"* which is correct and is also having a similar meaning as the tonal word *"chŏ"* in the reference sentence.

The Engliah 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. As compared to baseline systems translation, the proposed approach has the best-predicted sentence since most of the words are correctly predicted. Therefore, it attains partial adequacy but 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 similar meaning with *"group"* in the reference translation.

But PD2, PD3 and PD3 do not recognize the tonal word. They have incorrectly predicted the sentence by only one word which is completely inadequate and also not fluent.

However, in the PD1, the contextual meaning of the predicted sentence is partially adequate as 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)

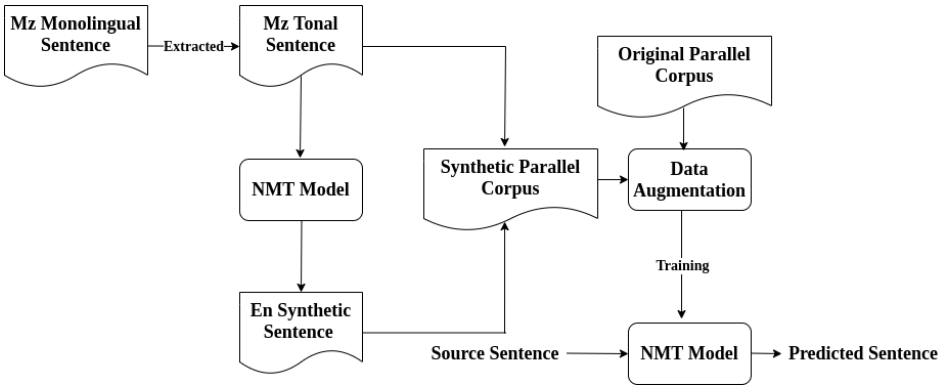

Fig. 5. Proposed Approach

— 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.

For all the systems, the predicted translations are contextually inadequate with a completely different meaning as compared to the reference translation. But 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 are able to 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 a well-formed and proper sentence of the target language, it is considered to be good in fluency.

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.*

ISSN 2007-9737

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 are unable to correctly generate the tonal word in their predicted sentence.

There are multiple errors in the named entity as the word *"goal"* appears in both source text as well as reference text, but none to the systems have correctly generated in their predicted sentence. Therefore, due to huge errors in named entities and contextually different predictions, the predicted sentences of all the systems are inadequate.

In terms of 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 are able to 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 both predicted a contextually different sentences but 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"*. But 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, however, all the systems have predicted a sentence which 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 but 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 are able to 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 but 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 but 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 it is not a well formed Mizo sentence.

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 are able to 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 contextual meaning of the reference translation is different from the predicted sentence of the systems. In terms of fluency, it is not a well-formed Mizo sentence.

9 Conclusion and Future Work

In this article, we have developed EnMzCorp1.0 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 dataset will be available here[14]. Moreover, the proposed approach based on the data augmentation technique 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. In the future, we will increase the size of the dataset and explore the knowledge-transfer-based NMT approach for improvement.

[14] https://github.com/cnlp-nits/EnMzCorp1.0

Acknowledgments

The authors are thankful to the Department of Computer Science and Engineering and Center for Natural Language Processing (CNLP) at the National Institute of Technology, Silchar for providing the requisite support and infrastructure to execute this work.

References

1. **Bahdanau, D., Cho, K., Bengio, Y. (2015).** Neural machine translation by jointly learning to align and translate. 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, pp. 1–15.
2. **Bentham, J., Pakray, P., Majumder, G., Lalbiaknia, S., Gelbukh, A. (2016).** Identification of rules for recognition of named entity classes in Mizo language. 2016 Fifteenth Mexican International Conference on Artificial Intelligence (MICAI), IEEE, pp. 8–13.
3. **Cho, K., van Merriënboer, B., Bahdanau, D., Bengio, Y. (2014).** On the properties of neural machine translation: Encoder–decoder approaches. Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, Association for Computational Linguistics, Doha, Qatar, pp. 103–111.
4. **Cho, K., van Merrienboer, B., Bahdanau, D., Bengio, Y. (2014).** On the properties of neural machine translation: Encoder-decoder approaches. **Wu, D., Carpuat, M., Carreras, X., Vecchi, E. M.**, editors, Proceedings of SSST@EMNLP 2014, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, Doha, Qatar, 25 October 2014, Association for Computational Linguistics, pp. 103–111.
5. **Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y. (2014).** Learning phrase representations using RNN encoder–decoder for statistical machine translation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, pp. 1724–1734.
6. **Fanai, L. T. (1992).** Some aspects of the lexical phonology of Mizo and English an autosegmental approach.
7. **Federico, M., Bertoldi, N., Cettolo, M. (2008).** IRSTLM: an open source toolkit for handling large scale language models. INTERSPEECH 2008, 9th Annual Conference of the International Speech Communication Association, Brisbane, Australia, September 22-26, 2008, ISCA, pp. 1618–1621.
8. **Gehring, J., Auli, M., Grangier, D., Dauphin, Y. (2017).** A convolutional encoder model for neural machine translation. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vancouver, Canada, pp. 123–135.
9. **Gu, J., Hassan, H., Devlin, J., Li, V. O. (2018).** Universal neural machine translation for extremely low resource languages. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, pp. 344–354.
10. **Hochreiter, S., Schmidhuber, J. (1997).** Long short-term memory. Neural Comput., Vol. 9, No. 8, pp. 1735–1780.
11. **Kalchbrenner, N., Blunsom, P. (2013).** Recurrent continuous translation models. Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Seattle, Washington, USA, pp. 1700–1709.
12. **Kocmi, T. (2020).** Exploring benefits of transfer learning in neural machine translation. CoRR, Vol. abs/2001.01622.
13. **Koehn, P. (2010).** Statistical Machine Translation. Cambridge University Press, USA, 1st edition.
14. **Koehn, P., Och, F. J., Marcu, D. (2003).** Statistical phrase-based translation. Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, pp. 127–133.
15. **Lalrempuii, C., Soni, B. (2020).** Attention-based english to Mizo neural machine translation. Machine Learning, Image Processing, Network Security and Data Sciences, Springer Singapore, Singapore, pp. 193–203.

16. **Lalrempuii, C., Soni, B., Pakray, P. (2021).** An improved English-to-Mizo neural machine translation. ACM Trans. Asian Low-Resour. Lang. Inf. Process., Vol. 20, No. 4.

17. **Lample, G., Conneau, A. (2019).** Cross-lingual language model pretraining. Advances in Neural Information Processing Systems (NeurIPS).

18. **Laskar, S. R., Dutta, A., Pakray, P., Bandyopadhyay, S. (2019).** Neural machine translation: English to Hindi. 2019 IEEE Conference on Information and Communication Technology, pp. 1–6.

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

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

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

22. **Lavie, A., Denkowski, M. J. (2009).** The meteor metric for automatic evaluation of machine translation. Machine Translation, Vol. 23, No. 2–3, pp. 105–115.

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

24. **Majumder, G., Pakray, P., Khiangte, Z., Gelbukh, A. (2018).** Multiword expressions (mwe) for Mizo language: Literature survey. **Gelbukh, A.**, editor, Computational Linguistics and Intelligent Text Processing, Springer International Publishing, Cham, pp. 623–635.

25. **Megerdoomian, K., Parvaz, D. (2008).** Low-density language bootstrapping: the case of Tajiki Persian. Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco, European Language Resources Association, pp. 3293–3298.

26. **Mikolov, T., Karafiát, M., Burget, L., Cernocký, J., Khudanpur, S. (2010).** Recurrent neural network based language model. volume 2, pp. 1045–1048.

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

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

29. **Pathak, A., Pakray, P. (2018).** Neural machine translation for Indian languages. Journal of Intelligent Systems, pp. 1–13.

30. **Pathak, A., Pakray, P., Bentham, J. (2018).** English–Mizo machine translation using neural and statistical approaches. Neural Computing and Applications, Vol. 30, pp. 1–17.

31. **Pennington, J., Socher, R., Manning, C. D. (2014).** Glove: Global vectors for word representation. **Moschitti, A., Pang, B., Daelemans, W.**, editors, 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, ACL, pp. 1532–1543.

32. **Probst, K., Brown, R. D., Carbonell, J. G., Lavie, A., Levin, L., Peterson, E. (2003).** Design and implementation of controlled elicitation for machine translation of low-density languages.

33. **Ramesh, S. H., Sankaranarayanan, K. P. (2018).** Neural machine translation for low resource languages using bilingual lexicon induced from comparable corpora. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop, Association for Computational Linguistics, New Orleans, Louisiana, USA, pp. 112–119.

34. **Saini, S., Sahula, V. (2018).** Neural machine translation for english to Hindi. 2018 Fourth International Conference on Information Retrieval and Knowledge Management (CAMP), pp. 1–6.

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

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

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

38. **Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., Polosukhin, I. (2017).** Attention is all you need. In **Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R.**, editors, Advances in Neural Information Processing Systems 30. Curran Associates, Inc., pp. 5998–6008.

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

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

41. **Zaremoodi, P., Buntine, W., Haffari, G. (2018).** Adaptive knowledge sharing in multi-task learning: Improving low-resource neural machine translation. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Melbourne, Australia, pp. 656–661.

Article received on 04/02/2022; accepted on 20/05/2022.
Corresponding author is Partha Pakray.