Proceedings of the WILDRE-6 Workshop @LREC2020, pages 48–54
Marseille, 20 June 2022

Language Resource Building and English-to-Mizo Neural Machine Translation Encountering Tonal Words

Vanlalmuansangi Khenglawt[1], Sahinur Rahman Laskar[2], Santanu Pal[3], Partha Pakray[2], Ajoy Kumar Khan[1]
Department of Computer Engineering, Mizoram University, Mizoram, (India)[1]
Department of Computer Science and Engineering, National Institute of Technology, Silchar (India)[2]
Wipro Limited, Bengaluru (India)[3]
mzut208@mzu.edu.in, sahinurlaskar.nits@gmail.com, santanu.pal2@wipro.com,
partha@cse.nits.ac.in, ajoyiitg@gmail.com

Abstract

Multilingual country like India has an enormous linguistic diversity and has an increasing demand towards developing language resources such that it will outreach in various natural language processing applications like machine translation. Low-resource language translation possesses challenges in the field of machine translation. The challenges include the availability of corpus and differences in linguistic information. This paper investigates a low-resource language pair, English-to-Mizo exploring neural machine translation by contributing an Indian language resource, i.e., English-Mizo corpus. In this work, we explore one of the main challenges to tackling tonal words existing in the Mizo language, as they add to the complexity on top of low-resource challenges for any natural language processing task. Our approach improves translation accuracy by encountering tonal words of Mizo and achieved a state-of-the-art result in English-to-Mizo translation.

Keywords: English-Mizo, Tonal, NMT

1. Introduction

Neural machine translation (NMT) has attained a promising approach in machine translation (MT) because of its context analysis ability and deal with long-range dependency problems (Bahdanau et al., 2015; Vaswani et al., 2017). However, it needs a sufficient amount of training data, which is a challenging task for the low-resource language pair translation (Koehn and Knowles, 2017). In this work, NMT is used to deal with a low-resource language pair: English—Mizo. To the best of our knowledge, there is a lack of publicly available English–Mizo corpus suitable for MT work. Therefore, very few contributions are applicable, specifically for the English–Mizo NMT task. Mizo is popularly known as a tonal language, which means a word with various tones can express different meanings (further details available in Section 2). The distinct tone markers are used in Mizo to represent the tonal word contextually. Based on our primary investigation, the translation of English–Mizo MT suffers in handling these tonal words and their corresponding context. Table 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 this problem, we propose an approach for encountering context-specific tonal words to improve the predicted sentence during the post-processing step (see Section 5).

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

Table 1: Example of predicted sentence (tone markers are marked as bold)

The major contributions are:

- Created an Indian language resource, namely, English–Mizo corpus that covers both parallel and monolingual data of Mizo. It will be publicly available here:`https://github.com/cnlp-nits/English-Mizo-Corpus`.
- Explored different NMT models and achieved a state-of-the-art result in English–Mizo translation.
- Proposed an approach of encountering context-specific tonal words for English-to-Mizo translation. To the best of our knowledge, we are the first to tackle this problem in English–Mizo translation.

2. Mizo Tonal Language

Along with English, Mizo[1] is the official language of the Indian state of Mizoram, and it is also known

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

Types of tone	Tone Marker (e)
High tone	é
Low tone	è
Rising tone	ě
Falling tone	ê

Table 2: Variation of tones with a tone marker

as Lushai which belongs to the Tibeto-Burman family of languages. According to Census-2011, there are 6,50,605[2] Mizo speakers, and they are known as Mizo/Lushai people. Although the writing system of the Mizo language is based on the Roman script like English, both languages are very different from each other. Generally, the word order of Mizo is Object–Subject–Verb (OSV), but in particular situations, it follows Subject–Verb–Object (SVO) like English. Apart from this, Mizo (Majumder et al., 2018; Pakray et al., 2015; Bentham et al., 2016) is quite different from English in linguistic aspects. Mizo language can be termed as a tonal language as the tone determines the lexical meaning of words. A total of eight tones are available in Mizo, wherein four tones are long tones and the remaining four are short tones. The use of diacritics is not standardized in Mizo tonal words. However, the tone markers or intonations are highlighted in the vowels (a, aw, e, i, o, u) with diacritics by some publisher[3] (Pakray et al., 2015). The main variation of tones in Mizo are high, low, rising and falling (Chhangte, 1993; Dutta et al., 2017; Gogoi et al., 2020). To indicate a distinct tone variation, a unique tone marker is employed, as shown in Table 2. As the tonal word alone can imply a different meaning, without the use of a tone marker, the tonal variation of a word will be determined by the context of the sentence. Therefore, an indication of proper tone marker is immensely imperative to determine the lexical denotation of the word. For example, as shown in Table 3, based on the tone, the Mizo word *'kang'* can have different connotations in English. *'Kang'* can be translated as *'fry'*, *'dried up'*, *'above the ground'* and *'burn'* with a tone of 'high', 'low', 'rising' and 'falling' respectively. The Mizo language can be categorized under the language group, which has words with diacritics (Náplava et al., 2018). Since the tonal words are represented by the tone markers (Pakray et al., 2015). However, it is observed that Mizo words with tone markers are less frequent than those without tone markers[4], unlike Vietnamese (Náplava et al., 2018), Yorùbá (Adelani et al., 2021) and Arabic language (Fadel et al., 2019).

3. Related Work

There is limited work in the area of MT on the English–Mizo language pair (Pathak et al., 2018; Lalrempuii and Soni, 2020; Lalrempuii et al., 2021). It

[2] https://bit.ly/3xA8AKj

[3] https://vanglaini.org/

[4] https://vanglaini.org/

Tone	Sentences
High	*'Káng' -'fry/fried'* *Mizo: Vawksa ka **káng** a.* *English: I **fried** a pork.*
Low	*'Kàng'- 'dried up'* *Mizo: Ruahtui a tlem avangin lui tui pawh a **kàng** ral zo tawh.* *English: Due to less rainfall, the river **dried** up.*
Rising	*'Kăng' - 'above the ground'* *Mizo: I zuang **kăng** sang thei khawp mai.* *English: You can jump **above the ground** quite high.*
Falling	*'Kâng' - 'burn'* *Mizo: Tui sa in a inti **kâng**.* *English: I **burnt** myself with hot water.*

Table 3: Example sentences of different tones

is mainly due to the lack of availability of resources, as the Mizo language is a low resource language. In (Pathak et al., 2018), a parallel corpus of English-Mizo language pairs is prepared (29,973 train data) and performed a comparison between RNN based NMT and Phrase-based MT. Also, (Lalrempuii et al., 2021; Lalrempuii and Soni, 2020) investigated English-Mizo pair using several attention-based NMT models, including RNN, BRNN and transformer. Although researchers have explored the English-Mizo pair for the MT system, there are research gaps that are identified as follows:

- There is no standard English–Mizo corpus available publicly.
- None of them have tackled the linguistic challenges like tonal words of Mizo for English-to-Mizo translation.
- Although automatic translations like Google and Microsoft cover various languages worldwide, but lack the support of the Mizo language.

In this work, we have created an English–Mizo corpus and investigated with BERT-fused NMT (Zhu et al., 2020) using a bidirectional translation approach with synthetic parallel corpus (Niu et al., 2018; Sennrich et al., 2016). Also, we proposed a post-processing step for English-to-Mizo translation by focusing on tonal words.

4. Corpus Preparation

There is no standard or publicly available corpus for English–Mizo (En-Mz) corpus. Thus, we have prepared parallel data and Mizo monolingual data from different possible online resources. Online resources

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

Table 4: Statistics for train, valid and test set

include, Bible[5], online dictionary (Glosbe)[6] and Government websites[7] [8]. We have prepared 121,235 numbers of parallel sentences that include 44,168 Mizo sentences having tonal words. The parallel corpus contains 118,895 sentences from online sources (98.06%) and manually[9] prepared 2,340 sentences (1.93%). 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 manually prepared parallel sentences cover the general domain sentences, as shown in Table 10. Moreover, monolingual Mizo data of 2,061,068 sentences are prepared from various webpages, blogs and textbooks. To collect data from online sources, we have used web crawling[10] techniques. To allow for replication over different/several web pages, each element's `xpath` is formatted/encoded with a degree of generalization. It aided in crawling and retrieving information from a vast number of web pages. Before splitting a parallel corpus, we remove duplicates and noise (web-link (URLs), too many special characters, blank lines). Also, we verified by hiring a linguistic expert who possesses linguistic knowledge of both languages. The data statistics of the train, valid and test set, are shown in Table 4. During the split, we have considered parallel sentences having tonal words 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 corpus covers domains: Bible, Government notices/messages, dictionaries, and general domains. The percentage of tonal words presents in the train, validation, and test set are 11.20%, 10.50%, and 10.30%.

5. Approach

Our approach consists of three phases, as shown in Figure 1. Initially, for the first phase, we extracted Mz tonal sentences from the monolingual data of Mz. Then, the extracted Mz tonal sentences are used to generate En synthetic sentences by utilizing the backward NMT model (Mz-to-En). In this case, we used the conventional transformer model (Vaswani et al., 2017). We removed blank lines, under-translated sentences (single or double words) from En synthetic sentences, and the corresponding Mz tonal sentences. Thus, we prepared 33,021 synthetic parallel sentences, as given in Table 5. In the second phase, the synthetic parallel corpus is augmented with the original parallel corpus. Then we followed the technique of (Niu et al., 2018) by augmenting the swapped sentences (Mz-to-En). We added artificial tokens 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 (Zhu et al., 2020) for the forward (En-to-Mz) translation. BERT-fused NMT is utilized for leveraging the pre-trained model of English. We investigated different configurations, 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 the 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, to improve the translation accuracy of encountering tonal words, we propose an example-based post-processing step.

Example-based post-processing: For the post-processing step, we created an example-based dictionary by following these steps.

- We extracted keywords having tonal words from monolingual data of Mizo using a language-independent keyword extraction tool known as YAKE (Campos et al., 2020), considering maximum n-gram size= 3.
- We discarded the uni-gram words from the extracted keywords. Since, the uni-gram words are not able to represent the context-specific tonal words.
- We created an example-based dictionary $(K_z||K_y)$. 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. We searched each keyword of K_z in the predicted sentences and if it is found then replace it with the keyword of K_y. The reason behind using the post-processing step is that if

[5] `https://www.bible.com/`

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

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

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

[9] consumes manual effort and then verified by the hired linguistic expert

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

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

Table 5: Synthetic parallel data statistics

the trained model is unable to 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. We used an example-based dictionary because the tonal word is contextually dependent on the pre-or post-word of the concerned tonal word. In summary, 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.

6. Experiment and Result and Analysis

We performed preliminary experiments for both En-to-Mz and Mz-to-En translations using RNN (Bahdanau et al., 2015), transformer model (Vaswani et al., 2017) with sub-word segmentation technique i.e., byte pair encoding (BPE) (considered 32k merge operations). The results of the preliminary experiment are reported in Table 6. The quantitative results are evaluated in terms of automatic evaluation metric, bilingual evaluation understudy (BLEU)[11] (Papineni et al., 2002) and also with human evaluation (HE) (Pathak et al., 2018) on randomly selected 100 sample sentences by hiring a linguistic expert. We followed default configurations of OpenNMT-py[12] toolkit to implement RNN and transformer model. The Adam optimizer with a learning rate of 0.001, drop-outs of 0.3 (in case of RNN) and 0.1 (in case of transformer) are used in the training process. Also, followed default configurations of Fairseq[13] toolkit to implement BERT-fused NMT (Zhu et al., 2020). For En-to-Mz translation, the comparative results are reported in Table 7 and 8, where our approach (M8) attains a higher score. To examine the effectiveness of our approach in terms of encountering tonal words, a comparative analysis is presented in Figure 2. Although our approach encounters a higher frequency of tonal words than conventional transformer (Vaswani et al., 2017) and BERT-fused transformer (Zhu et al., 2020) models, far away from the frequency of tonal words in reference test sentences. Further, to inspect qualitative analysis of encountering tonal words, a few examples are presented in Table 9. It is observed that the conventional transformer (M1) and BERT-fused transformer (M2) models are unable to 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

[11] Utilized multi-bleu.perl script

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

[13] https://github.com/bert-nmt/bert-nmt

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

Table 6: BLEU scores of preliminary experiments

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 7: 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

tonal markers, which are marked as 'bold.' By capturing tone markers in tonal words, our approach significantly represents the contextual meaning of the sentences as compared to other models.

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

7. Conclusion and Future Work

In this work, our goal is to prepare an Indian language resource, i.e., English–Mizo corpus and investigate En-to-Mz translation by encountering tonal words by exploring different NMT models on the developed dataset. We will release the English-Mizo corpus, to be publicly available. Our approach is based on BERT-fused NMT with bidirectional data augmentation with synthetic parallel corpus and an example-based post-processing step. We attained better translation accuracy than a conventional transformer and BERT-fused NMT. In the future, we will increase the dataset size, domain-wise translation, and do more experiments to improve the translational accuracy of encountering tonal words.

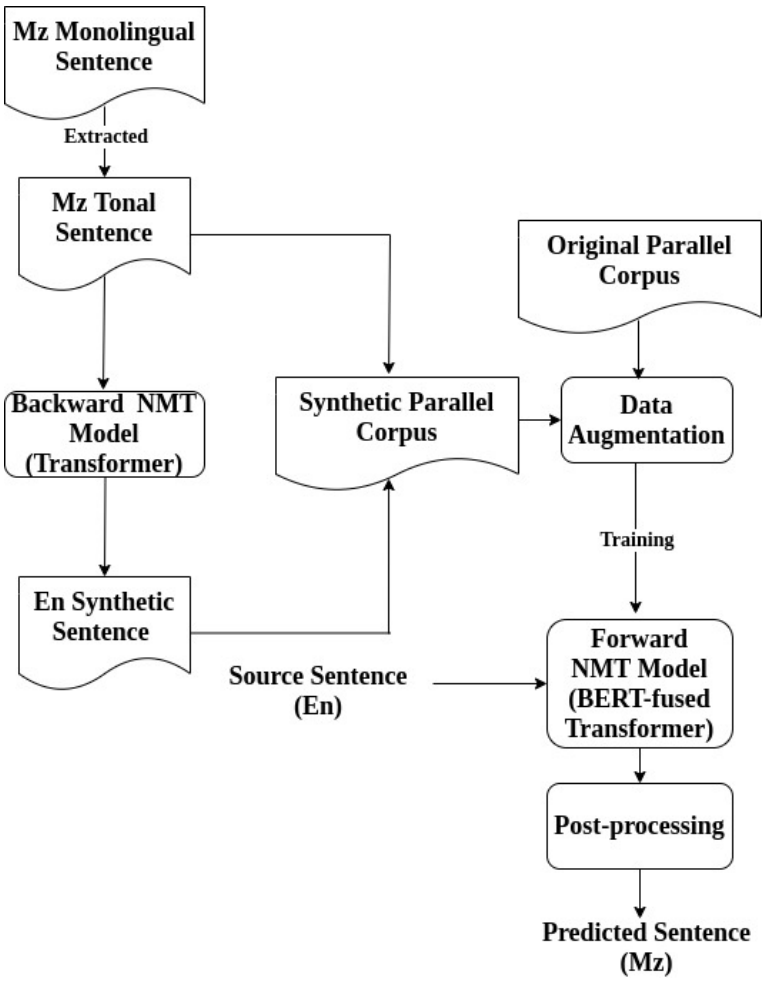

Figure 1: English-to-Mizo NMT System

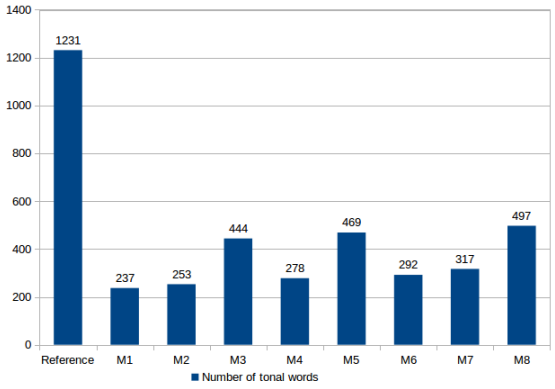

Figure 2: Comparative analysis on tonal frequency of words. Reference: Mizo target sentences (test data)

Source / Target	Predicted
It is nice. (En) A thà lutùk. (Mz)	A tha lutuk. (M1) A tha lutuk. (M2) **A thà lutùk.** (M3) A thà khawp mai . (M4) **A thà khawp mài .**(M5) A tha lutuk. (M6) A tha lutuk. (M7) **A thà lutùk .**(M8)
Don't tell lie. (En) Dáwt sáwi suh. (Mz)	Dawt sawi suh. (M1) Dawt sawi duh suh . (M2) **Dáwt sáwi duh suh .** (M3) Dawt sawi suh . (M4) **Dáwt sáwi suh .** (M5) Dáwt sawi suh . (M6) Dáwt sawi suh .(M7) **Dáwt sáwi suh .** (M8)

Table 9: Output examples of different models for En-to-Mz translation

Acknowledgement We want to thank the Center for Natural Language Processing (CNLP), the Artificial Intelligence (AI) Lab, and the Department of Computer Science and Engineering at the National Institute of Technology, Silchar, India, for providing the requisite support and infrastructure to execute this work.

8. Bibliographical References

Adelani, D. I., Ruiter, D., Alabi, J. O., Adebonojo, D., Ayeni, A., Adeyemi, M., Awokoya, A., and España-Bonet, C. (2021). The effect of domain and diacrit-

Table 10: Example of parallel sentences

English	Mizo	Source
Every grain offering of a priest shall be wholly burned.	Puithiam chhangphut thilhlan apiang chu **hâl** ral vek **tûr** a ni.	Glosbe
What burdens can advanced age impose on a person?	Kum upatnain mi chungah eng phurrit nge a thlen theih?	Glosbe
Joseph was already in Egypt.	Josefa chu Egypt ramah chuan lo awm tawh a.	Bible
Each with his household go to Jacob.	Mi tin mahni **chhûngte** ṭheuh **nên** Jakoba hnênah chuan an kal a.	Bible
Farmers are the backbone of our economy and our state.	Anni hi kan economy innghahna an ni a.	Government Website
This is a day for all of us to celebrate and honour our nation and our sovereignty.	He ni hi sawrkar ropui, mipui rorelna sawrkar kan neih theihna ni a ni a.	Government Website
I will be with you no more.	In hnênah hian ka **áwm dâwn** tawh lo a ni.	Manually
Now therefore you are cursed.	**Chuvângin**, **ânchhedawng** in lo nih **tâk** hi.	Manually

ics 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*, pages 61–75. Association for Machine Translation in the Americas.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In Yoshua Bengio et al., editors, *3rd International Conference on Learning Representations, ICLR 2015, May 7-9, 2015, Conference Track Proceedings*, pages 1–15, San Diego, CA, USA. arXiv.

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

Campos, R., Mangaravite, V., Pasquali, A., Jorge, A., Nunes, C., and Jatowt, A. (2020). Yake! keyword extraction from single documents using multiple local features. *Inf. Sci.*, 509:257–289.

Chhangte, L. (1993). *Mizo syntax*. Ph.D. thesis, University of Oregon.

Dutta, I., S., I., Gogoi, P., and Sarmah, P. (2017). Nature of Contrast and Coarticulation: Evidence from Mizo Tones and Assamese Vowel Harmony. In *Proc. Interspeech 2017*, pages 224–228.

Fadel, A., Tuffaha, I., Al-Jawarneh, B., and Al-Ayyoub, M. (2019). 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*, pages 215–225. Association for Computational Linguistics.

Gogoi, P., Dey, A., Lalhminghlui, W., Sarmah, P., and Prasanna, S. R. M. (2020). Lexical tone recognition in mizo using acoustic-prosodic features. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6458–6461, Marseille, France, May. European Language Resources Association.

Koehn, P. and Knowles, R. (2017). Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver, August. Association for Computational Linguistics.

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

Lalrempuii, C., Soni, B., and Pakray, P. (2021). An improved english-to-mizo neural machine translation. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 20(4), May.

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

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

Niu, X., Denkowski, M. J., and Carpuat, M. (2018). Bi-directional neural machine translation with synthetic parallel data. In Alexandra Birch, et al., editors, *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation, NMT@ACL 2018, Melbourne, Australia, July 20, 2018*, pages 84–91. Association for Computational Linguistics.

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

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

USA. Association for Computational Linguistics.

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

Sennrich, R., Haddow, B., and Birch, A. (2016). 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)*, pages 86–96, Berlin, Germany, August. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *CoRR*, abs/1706.03762.

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