Bidirectional Neural Machine Translation (NMT) using Monolingual Data for Khasi-English Pair

Lavinia Nongbri [a], Gourashyam Moirangthem [a], Samarendra Salam [b] and Kishorjit Nongmeikapam [a]

[a] Computer Science and Engineering, Indian Institute of Technology, Manipur

[b] Department of Mathematics G.P. Women's College, Imphal

laviniangbri@gmail.com gourashyam@iiitmanipur.ac.in samar.crypt@gmail.com kishorjit@iiitmanipur.ac.in

Abstract

Due to a lack of parallel data, low-resource language machine translation has been unable to make the most of Neural Machine Translation. This paper investigates several approaches as to how low-resource Neural Machine Translation can be improved in a strictly low-resource setting, especially for bidirectional Khasi-English language pairs. The back-translation method is used to expand the parallel corpus using monolingual data. The work also experimented with subword tokenizers to improve the translation accuracy for new and rare words. Transformer, a cutting-edge NMT model, serves as the backbone of the bidirectional Khasi-English machine translation. The final Khasi-to-English and English-to-Khasi NMT models trained using both authentic and synthetic parallel corpora show an increase of 2.34 and 3.1 BLEU scores, respectively, when compared to the models trained using only authentic parallel dataset.

1 Introduction

1.1 Introduction of Machine Translation

Machine Translation is a sub-field of Natural Language Processing that deals with the automatic translation of human languages. The translation can be text-text, speech-speech, speech-text and text-speech. Text-based machine translation has come a long way, from Rule-based translation, example-based translation, Statistical Machine Translation (SMT), and to Neural Machine Translation (NMT).

Recurrent Neural Networks (RNN) have addressed the problems that rule-based and statistical machine translation approaches had in capturing exceptions in human languages and retaining word dependency. They are, however, slow to train and have limitations when it comes to modeling long-term dependencies. Recent advancements in NMT have demonstrated outstanding efficiency by combining the encoder-decoder architecture with attention mechanism. NMT has become more popular in academia and industry as a result of advancements in the attention mechanism. Using a self-attention mechanism, Vaswani et al., 2017 NMT model Transformer has attained a state-of-the-art BLEU score in both English-to-German and English-to-French translations.

By including NMT into the machine translation methodology, high-quality translation has been achieved. However, the performance of NMT is highly dependent on the size of the dataset. The performance of NMT on low-resource languages is marginal compared to the high-resource languages. Therefore, it is necessary to find ways to make up for the shortage of resources in order to increase the translation quality.

1.2 Khasi Language

Khasi is a language spoken by about over a million people in the north-east state of India, Meghalaya, particularly in the districts of Jaintia Hills, East Khasi Hills, and West Khasi Hills. It is a member of the Khasian languages that form the westernmost branch of the Mon-Khmer language family of the Austroasiatic language. There is no special script for the Khasi language. Early in the 19th century, a British missionary named William Carey began writing Khasi using the Bengali script. Later in 1841, Thomas Jones, a Welsh missionary, introduced the Khasi alphabet using the Latin script. Khasi language consists of 23 letters; the basic Latin alphabet's letters *c, f, q, v, x* and *z* are removed, and the diacritical letters *ï* and *ñ* are added in their place, along with the digraph *ng*, which is classified as a separate letter.

This paper aims to enhance NMT-based bilingual translation between Khasi and English by employing corpus creation and a sub-tokenization technique. It addresses the scarcity of parallel data by developing both parallel and monolingual corpora. Additionally, it utilizes subword tokenization to improve translations for words outside the vocabulary. Despite having limited resources, both parallel and monolingual, this study employs NMT for the Khasi-English language pair in both translation directions.

This work also explores the back-translation method as to how Khasi-English Neural Machine Translation can be improved under a strictly low-resource setting.

The rest of the paper is organized as follows: Section 2 describes a brief survey of the related works. Section 3 outlines the methods employed for this translation task alongside how monolingual data can be used to expand the existing parallel corpus, Section 4 discuss the experimentation setup and results of the experiments and Section 5 concludes the paper.

2 Related Works

Few works related to Khasi and English translation have been reported. Singh and Hujon, 2020 has reported findings of the effectiveness of statistical and neural machine translation systems in domain-specific English to Khasi translation. It was reported that the SMT performed better than the NMT for this language pair. However, the performance of the SMT model degraded as the sentence length increased.

Donald Jefferson Thabah and Purkayastha, 2021 reported a cross-lingual language model pretraining system for bidirectional Khasi-English machine translation. The model achieved a BLEU score of 39.63 and 32.69 for translating English–Khasi and Khasi–English respectively when tested on similar domain test sentences.

Laskar et al., 2021 has reported the development of EnKhCorp1.0, a corpus for English–Khasi pair, and implemented baseline systems for English to Khasi and Khasi to English translation based on the neural machine translation approach.

Another recent work by Hujon et al., 2023, discussed the experiments and improvement of the results of neural machine translation using transfer learning for the English-Khasi language pair. The study reported that the joint vocabulary of three languages, English, French and Khasi has contributed to the outstanding performance of NMT Transfer Learning Model as compared to the NMT Baseline model.

3 Methodology

The methodologies employed in this current work can be broadly grouped into four parts, namely : Parallel Corpus Creation, Data Preprocessing, Sub-word tokenization, Modelling of Khasi-English MT and Back-translation of Khasi monolingual sentences.

3.1 Parallel Corpus Creation

NMT benefits from ample training data which is a challenge for low-resource languages. LRLs have minimal speakers and their presence on the internet is low. Therefore, parallel text corpora for such languages are hard to find, and the development of one is expensive in terms of time and manpower. Recent research by Kocmi and Bojar, 2018,Liu et al., 2020,Platanios et al., 2018,Qi et al., 2018,Zaremoodi et al., 2018 seems to consider a language pair as low-resource (LR) or extremely LR if the available parallel corpora for the considered pair for NMT experiments are below 0.5 million and below 0.1 million, respectively. As mentioned in Ranathunga et al., 2023, even if a particular language has a large number of monolingual corpora while still having a small parallel corpus with another language, this language pair is considered as LR for the NMT task.

Since Khasi is also a LRL, creation of a parallel corpus for Khasi-English language pair is challenging. The size of the parallel corpus is minimal as there are limited sources from which parallel sentences for this language pair can be collected. The sources of the data collected are as follows:

- **WMT 23** The Shared Task: Low-Resource Indic Language Translation (LRILT), WMT 23[1] repository contains aligned Khasi and English sentences extracted from Bible. The test and validation dataset each consisting 1000 sentences are also taken from the WMT 23 repository. The domain of this parallel corpus except for the test and validation sets are based on religion.

- **PIB** Government policies, programmes, initiatives and achievements are published on the Press Information Bureau[2] website. In addition to English and Hindi, the articles on this website are available in 16 regional languages,

[1] http://www2.statmt.org/wmt23/indic-mt-task.html

[2] https://pib.gov.in/indexd.aspx

viz. Urdu, Marathi, Telegu, Tamil, Punjabi, Bengali, Kannada, Odia, Gujarati, Assamese, Malayalam, Manipuri, Mizo, Nepali, Tenyidei and Khasi. Regional news of Shillong is available on PIB website since May 2023. English is another available language at PIB Shillong. Therefore, the articles present in both the languages are scraped from the website for creation of a parallel corpus. The articles cover a wide range of subjects, so this corpus's domain is categorized as generic.

- **Glosbe** Glosbe[3] is a multi-lingual online dictionary that includes in-context translations of words for languages like English, German, Greek, Spanish, Japanese, Hindi, Khasi, etc. Khasi-to-English and English-to-Khasi translations of words are supported for the Khasi language in the online dictionary. The meaning and usage of each Khasi and English words are illustrated with example sentences in both the languages. These sentences provide a great source for the building of a parallel corpus, and therefore are scraped solely for research purposes.

- **Opus** Opus [4] is an open source collection of parallel corpus. It assembles and aligns open-access translated texts from the web. A corpus of around 1.7k aligned Khasi-English sentence pairs is available on the website.

- **Human translation** 150 Khasi monolingual sentences from WMT 23 repository are translated to English manually. This is included since the majority of Khasi phrases obtained are translated from English and may lack the linguistic features of Khasi language.

The parallel corpus created from collection of Khasi and English sentences from the sources mentioned above is based on generic and religious domains. A breakdown of the statistics of sentences obtained from each source is provided in Table 1. The column - number of sentences specifies the sentence collected both for Khasi and English language pairs from the respective sources.

3.2 Data Preprocessing

A high quality corpus is a must for the high performance of the translation model. The collected sentences are real-world data which are often noisy and therefore, must first undergo preprocessing methods before they can be utilised in the study. The preprocessing steps conducted are Data Cleaning, Lower-casing and Tokenization.

Source	Number of sentences	Domain
WMT	26000	Religion
PIB	3159	Generic
Glosbe	2861	Generic
Opus	1755	Generic
Human Translation	150	Generic

Table 1: Data collection statistics for Khasi and English

Data Cleaning step will eliminate missing, incorrect and duplicate data from the dataset. Presence of such noise in the dataset can negatively affect the performance of the model. Therefore, it is the foremost and most crucial step of data preprocesing. The data cleaning stage in the corpus creation of Khasi-English parallel sentences entails segmenting the text paragraph into sentences, aligning the Khasi-English sentence pair, and removing duplicate sentences.

As a first step of data cleaning, the clustered texts in both the language pair are split into sentences based on delimiters such as '.', '?' and '!' using the NLTK tool by Bird, 2006. Most of the scrapped sentences are consistently aligned. However, the sentence orderings in a few PIB articles differ, which affects the sentence alignment. Since these are few in number, they are manually aligned. Thereafter, duplicate sentences are searched for and removed.

The figures in Table 2 shows the number of sentences obtained after cleaning the raw data. The table also shows the breakdown of the sentences for training, testing and validation for each language with its corresponding file name. The train.en, valid.en and test.en are the training, validation and testing file name for English language. The train.kh, valid.kh and test.kh are the training, validation and testing file name for Khasi language.

File name	Number of sentences
train.en	31708
valid.en	1000
test.en	1000
train.kh	31708
valid.kh	1000
test.kh	1000

Table 2: Statistics of Khasi and English sentences

Lowercasing is a simple text preprocessing step

[3] https://glosbe.com/

[4] https://opus.nlpl.eu/

where each single character is converted to lowercase. This step is not mandatory if the language does not differentiate between lowercase and uppercase characters. However, some languages follow the rule of capitalizing the first character of proper nouns. This can lead to data sparsity issues if two similar words occur in a dataset that differ in capitalization. Such words will be represented as different words in the vector space when they are the same. Since English and Khasi adopt the rule of capitalization, each character of the sentences is changed to lowercase.

Tokenization is the final step of data preprocessing implemented in this work. It is a method to split a sentence into tokens. A token can be an n-gram, a word, a subword, or a character. This simple step will assist the model in understanding the meaning of each of the words or tokens, as well as how they function in the larger text. Using Moses tokenizer by Koehn et al., 2007, Khasi and English phrases are split into word tokens. Using these word tokens, a vocabulary of unique tokens present in the dataset is created.

Table 3 shows the vocabulary size of unique word tokens for Khasi and English in the parallel corpus created.

Language	Vocabulary size
Khasi	11407
English	18050

Table 3: Unique word tokens for Khasi and English in the parallel corpus

3.3 Subword Tokenization

Subword tokenization is the segmentation of a word token into smaller tokens. The tokens differ with each subword tokenization technique. In a recent work by Sennrich and Zhang, 2019, a low-resource NMT attained good performance by a meaningful subword tokens vocabulary. Similarly, it has been found by Sennrich et al., 2015b that the models based on subword tokenizers achieve better accuracy for the translation of rare words than models based on large vocabulary and are able to productively generate new words that were not seen at training time. Thus, these studies suggest that low-resource NMT benefits from subword tokenizer in resolving out-of-vocabulary words. In regard to these findings, an experiment utilizing a subword tokenizer is conducted in the bidirectional Khasi-to-English translation model to study the effects on the low-resource language - Khasi.

Given the limited number of sentence pair in the Khasi-English training set, a subword tokenization method that can handle new and rare words must be selected. **Byte Pair Encoding** Gage, 1994 is known for handling such new and uncommon words. For this purpose, BPE is implemented as the subword tokenizer for this current work.

The BPE tokenizer starts by computing the unique set of words used in the training corpus after the word tokenization step, then builds a vocabulary which is a set of all individual characters. Thereafter, the vocabulary size is gradually increased in the following ways.

1. Two most frequently adjacent occurring symbols, say 'i' and 'j' are selected.

2. Then, the two symbols in step 1 are merged and every adjacent 'i' and 'j' in the corpus is replaced with the new symbol 'ij'.

3. Step 1 and 2 is continued for n times. This will create n novel tokens.

4. The resulting vocabulary consists of the original set of characters plus n new symbols.

3.4 Modelling of bidirectional Khasi-English NMT

The bidirectional Khasi-English machine translation model is implemented using the state-of-the-art Neural Machine Translation method, Transformer Vaswani et al., 2017. Transformer is a NMT model based on the encoder and decoder architecture together with the self attention mechanism and Feed Forward Neural Network. The model architecture of Transformer is illustrated in Figure 1.

The input to a Transformer model is a sequence of tokens, such as word tokens or subword tokens (from a subword tokenizer). Each token is represented as an embedding vector in the vector space. Since the model lacks information about the order of the tokens, positional embeddings are added to these vector embeddings to provide information about the positions of tokens in the sequence.

The encoding and decoding components are composed of stacks of encoders and decoders of the same number. Each encoder layer is composed of two sub-layers which are multi-head self-attention mechanism and fully connected Feed Forward Neural Network. The embedded tokens of the source language passes through the self-attention layer of the encoder where the model determine the score of each token in the input sequence with respect to

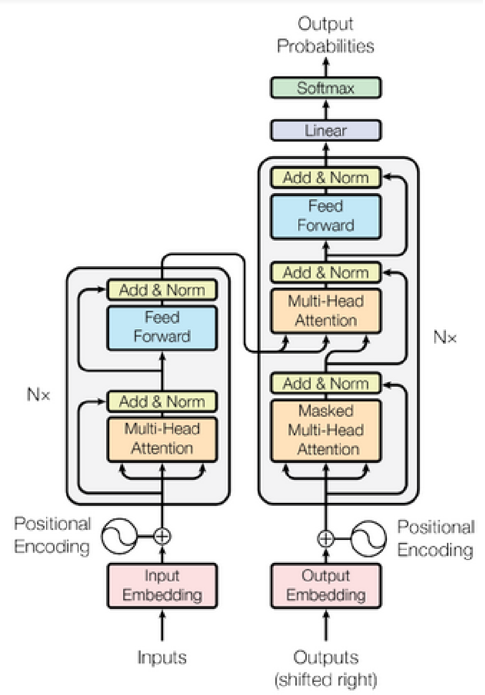

Figure 1: The Transformer - model architecture

every other token. The score is obtained by taking the dot product of the query vector and the key vector of the word being scored.

$$Attention = softmax(\frac{Q \times K^T}{\sqrt{d_k}}) \times V \quad (1)$$

where Q, K, d_k, V are the query vector, key vector, dimension of the key vectors and value vectors.

Ultimately, the weights of each embedded tokens of the sequence with respect to a token is computed and a weighted sum is obtained for each token. These tokens then pass through the Feed-Forward Neural Network. Layer normalization is applied after each sub-layer to stabilize training.

The output of one encoder acts as an input to the next encoder. Finally, the output of the top-most encoder is then transformed into a set of attention vectors K and V and are passed to the decoder stack.

Each decoder layer has both the sub-layers as encoder with the addition of a third sub-layer that performs multi-head self attention over the encoder stack's output.Similarly, for decoder also, the output of one decoder is passed to the next decoder and cumulate the decoding results. Positional embeddings are also added to the inputs of decoder. The final linear layer which is followed by a Softmax Layer generates the prediction of the input sentence.

3.5 Taking advantage of Monolingual Data

The parallel corpora size also plays an important factor in the quality of translation generated by the model. The absence of massive volume of parallel corpora acts as a barrier for Neural Machine Translation in low-resource languages. The creation of a parallel corpus is expensive; therefore, ways to expand the existing parallel corpus must be developed and adopted. One way of expanding the existing parallel corpora is by using monolingual data to create a synthetic parallel corpus.

Back translation Sennrich et al., 2015a is a technique for producing synthetic parallel corpus using target→source machine translation. Through reverse translation of monolingual Khasi sentences, synthetic English sentences are obtained. This creates a synthetic parallel corpus which can be collated with the actual sentence pairs to expand the parallel corpus.

To expand the parallel corpus developed in this work, monolingual Khasi sentences are extracted from school textbooks in Meghalaya enscribed in Khasi dialect. These textbooks are available at Internet archive [5] in OCR extracted text format. These monolingual sentences also undergoes through a data preprocessing step as discussed in Section 3.2. Spelling mistakes were in abundant because the OCR misread some of the characters such as e as c, h as b and many more. As a spell checker for Khasi language is not available, most of the editing is done manually and therefore, the number of monolingual sentences obtained are few in number.

Source	Language	Number of Sentences
School text-books	Khasi	3408

Table 4: Monolingual Khasi sentences

The monolingual Khasi sentences are then translated to English sentences using Khasi-to-English translation model.

The English sentences are post edited and few sentences that are irrelevant with the source sentences are removed. Thereafter, 3364 synthetic parallel sentences remain. These sentences are added to the the authentic parallel sentences corpus. The

[5] https://archive.org/

sentences are randomized to ensure a proper combination with the synthetic and original sentences.

Authentic	Synthetic	Combined
31708	3364	35072

Table 5: Statistics of combined authentic and synthetic Khasi and English sentences

4 Experimentation

4.1 Experiment setup

The base model of Transformer Vaswani et al., 2017 use 8 attention heads, 512 dimensions of model and 2048 dimensions of Feed Forward Neural Network. The same configurations of the base model is used for this study. The Khasi-to-English and English-to-Khasi translation models are trained using Jupyter notebook with OpenNMT Klein et al., 2018. The training was done for 100k steps which lasted for 24 hours approximately.

To test if BPE outperforms word-level tokenization for this task, an experiment on two models ***T_{base}*** and ***T_{base}+bpe*** is conducted for both directions. These two models are tuned to the base hyperparameters of Transformer and is trained with Khasi-English parallel corpus consisting of 31708 sentences.

For Back-translation of the 3408 monolingual sentences, ***T_{base}+bpe*** model in Khasi-to-English direction is used. After completing this stage, synthetic parallel corpus are produced, which are utilized to train model ***T_{base}+bpe+back*** alongside authentic parallel corpus.

Model ***T_{base}+bpe+back*** is trained on the same hyperparameters as the above models in both directions.

4.2 Experimental Results and Analysis

The models are tested using the test data mentioned in Table 2. The results are evaluated using the BLEU score Papineni et al., 2002.

Model	Khasi-English	English-Khasi
T_{base}	13.79	15.16
T_{base}+bpe	15.81	18.13
T_{base}+bpe+back	18.15	21.23

Table 6: Experiment results evaluated using BLEU score

- **Comparative study between word tokenization and subword tokenization method**

  The BLEU score of the model *T_{base}+bpe* increased by 2.02 and 2.97 in Khasi-to-English and English-to-Khasi, respectively as compared to model *T_{base}*. This increment shows that the subword tokenization method performs better than the word tokenization method. Considering the minimal training corpus size, the subword tokenization technique shows promising results, especially when low resource language is considered.

- **Effect of Back-translation**

  In the case of back translation, the *T_{base}+bpe+back* model outperforms the model *T_{base}+bpe* by 2.34 and 3.1 BLEU score for Khasi-to-English and English-to-Khasi, respectively. The result has improved tremendously considering the minimal monolingual Khasi sentences. There is a potential the score could increase if the quantity of monolingual sentences is increased. This can be investigated further by determining the ratio of synthesised and real parallel datasets that yields the maximum score of the translation model.

Table 7,8,9,10 show the output of each source sentence to its target sentence produced by the models mentioned in Section 4.1.

Khasi-sen-1	La sam ia ka jingai jingiarap bai seng kam sha palat shi lak ki dkhot SHG
English-ref-1	Disburses Seed Capital Assistance to over one lakh SHG members
T_{base}	Sweets were distributed towards providing development to more than shgs
T_{base}+bpe	Distributes essential support in being issued to more than one lakh members
T_{base}+bpe+back	Distributes essential support to more than one lakh shgs members

Table 7: Translation of source Khasi sentence *Khasi-sen-1* to English language by various Khasi-to-English translation models.

Source sentence *Khasi-sen-1* in the Khasi language is not adequately translated by any of the three Khasi-to-English translation models. When comparing the output, it can be seen that model *T_{base}+bpe+back* translation is quite similar to the reference sentence *English-ref-1*. Model *T_{base}+bpe*

Khasi-sen-2	Kane ka jaka kaba itynnad bha
English-sen-2	This place looks beautiful
$\boldsymbol{T_{base}}$	This point is a success of attraction
$\boldsymbol{T_{base}}$***+bpe***	This most beautiful room
$\boldsymbol{T_{base}}$***+bpe+back***	This is a beautiful place

Table 8: Translation of source Khasi sentence *Khasi-sen-2* to English language by various Khasi-to-English translation models.

translation also conveys some of the sentence's meaning but has fluency issues.

Model T_{base}*+bpe+back* produced the most accurate translation for the source text *Khasi-sen-2*, except for the sentence structuring. Model T_{base}*+bpe* also captured the meaning of the sentence except, it misunderstood 'jaka' as room rather than place.

English-sen-1	The fish cannot live without water.
Khasi-ref-1	Ka dokha kam lah im khlem ka um.
$\boldsymbol{T_{base}}$	Ki dohkha kim lah ban im shabar ka um .
$\boldsymbol{T_{base}}$***+bpe***	Ki dohkha kim lah ban im khlem um .
$\boldsymbol{T_{base}}$***+bpe+back***	Ki dohkha kim lah ban im khlem um .

Table 9: Translation of source English sentence *English-sen-1* to Khasi language by various English-to-Khasi translation models.

The word 'fish' in the source English sentence *English-sen-1* is interpreted as plural by the translation models and therefore, it is translated to 'ki' rather than 'ka', which stands for single noun. Model T_{base} translate 'without' as shabar which means outside. All the models have generated a close version of *Khasi-ref-2* for source sentence *English-sen-1*.

However, the models performed poorly in translation of source sentence *English-sen-2*. The translated word 'shnong' indicates that Meghalaya has been referred to as a village by model T_{base}. The model also outputs words that are not at all related with the source sentence, such as 'jingdon' which means wealth. This is the same case with the output of the other two models, T_{base}*+bpe* and T_{base}*+bpe+back*.

English-sen-2	Meghalaya village council files fir against scribe patricia mukhim for social media post on assault case.
Khasi-ref-2	Ka dorbar shnong ha Meghalaya ka ai fir ia patricia mukhim na bynta ki jingthoh ha social media halor ki case ba leh donbor.
$\boldsymbol{T_{base}}$	Ki shnong meghalaya ki rim ia ki jingdon jingem kiba kordor pyrshah ia ka jingpynpoi ia ki lad social media katkum ka juk mynta .
$\boldsymbol{T_{base}}$***+bpe***	Ka jylla meghalaya ka peit bniah ia u high commissioner uba ki social media ki dang pyrshang ban kurup ia ka rynsan social media .
$\boldsymbol{T_{base}}$***+bpe+back***	Ka jylla meghalaya ka peit bniah ia u high commissioner uba ki social media ki dang pyrshang ban kurup ia ka rynsan social media .

Table 10: Translation of source English sentence *English-sen-2* to Khasi language by various English-to-Khasi translation models.

5 Conclusion

Approaches to improving Khasi-English bidirectional machine translation are discussed in the paper. The effectiveness of techniques including subword tokenization and back-translation in LRL NMT is being investigated. Experiments have shown that subword tokenization and back-translation methods are promising methods to enhance the translation quality of Khasi-English bidirectional machine translation.

Expanding the machine translation model's effectiveness involves training it with an increased volume of parallel sentences. Additionally, exploring augmentation techniques for generating parallel sentences could be beneficial, especially in scenarios with limited resources. Notably, the sentence structures found in the Bible differ significantly from contemporary sentence structures. To adeptly translate sentences from the current generation, the model must also be trained on up-to-date sentences, consequently expanding its vocabulary.

6 Acknowledgement

This research work was supported by grant from Ministry of Electronics & Information Technology (MeitY), Govt. of India vide approval no. 11(1)/2022-HCC(TDIL)-Part(4).

References

Steven Bird. 2006. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 Interactive*

Presentation Sessions, pages 69–72.

N Donald Jefferson Thabah and Bipul Syam Purkayastha. 2021. Low resource neural machine translation from english to khasi: A transformer-based approach. In *Proceedings of the International Conference on Computing and Communication Systems: I3CS 2020, NEHU, Shillong, India*, pages 3–13. Springer.

Philip Gage. 1994. A new algorithm for data compression. *C Users Journal*, 12(2):23–38.

Aiusha V Hujon, Thoudam Doren Singh, and Khwairakpam Amitab. 2023. Transfer learning based neural machine translation of english-khasi on low-resource settings. *Procedia Computer Science*, 218:1–8.

Guillaume Klein, Yoon Kim, Yuntian Deng, Vincent Nguyen, Jean Senellart, and Alexander M Rush. 2018. Opennmt: Neural machine translation toolkit. *arXiv preprint arXiv:1805.11462*.

Tom Kocmi and Ondřej Bojar. 2018. Trivial transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1809.00357*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pages 177–180.

Sahinur Rahman Laskar, Abdullah Faiz Ur Rahman Khilji Darsh, Partha Pakray, Sivaji Bandyopadhyay, et al. 2021. Enkhcorp1. 0: An english–khasi corpus. In *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021)*, pages 89–95.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Emmanouil Antonios Platanios, Mrinmaya Sachan, Graham Neubig, and Tom Mitchell. 2018. Contextual parameter generation for universal neural machine translation. *arXiv preprint arXiv:1808.08493*.

Ye Qi, Devendra Singh Sachan, Matthieu Felix, Sarguna Janani Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? *arXiv preprint arXiv:1804.06323*.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural machine translation for low-resource languages: A survey. *ACM Computing Surveys*, 55(11):1–37.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015a. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015b. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Rico Sennrich and Biao Zhang. 2019. Revisiting low-resource neural machine translation: A case study. *arXiv preprint arXiv:1905.11901*.

Thoudam Doren Singh and Aiusha Vellintihun Hujon. 2020. Low resource and domain specific english to khasi smt and nmt systems. In *2020 International Conference on Computational Performance Evaluation (ComPE)*, pages 733–737. IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Poorya Zaremoodi, Wray Buntine, and Gholamreza Haffari. 2018. Adaptive knowledge sharing in multi-task learning: Improving low-resource neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 656–661.