Sādhanā (2023) 48:237
https://doi.org/10.1007/s12046-023-02308-8

Neural machine translation for limited resources English-Nyishi pair

NABAM KAKUM[1], SAHINUR RAHMAN LASKAR[2], KOJ SAMBYO[1] and PARTHA PAKRAY[3,*]

[1]Department of Computer Science and Engineering, National Institute of Technology, Arunachal Pradesh, Itanagar, India
[2]School of Computer Science, University of Petroleum and Energy Studies, Dehradun, India
[3]Department of Computer Science and Engineering, National Institute of Technology, Silchar, Silchar, India
e-mail: nabamkakum08@gmail.com; sahinurlaskar.nits@gmail.com; sambyo.koj@gmail.com; parthapakray@gmail.com

MS received 3 August 2021; revised 3 May 2023; accepted 22 August 2023

Abstract. Neural machine translation handles sequential data over the variable length of input and output sentences and accomplishes a state-of-the-art method for the task of machine translation. Although the neural machine translation shows good performance in both low and high-resource language pairs translation, it requires adequate parallel training data. In low-resource language sets, the preparation of the corpus is strenuous and time-consuming. Automatic translation systems like Google and Bing cover under-resourced Indian languages, but lack the support of the Nyishi language. It is due to the lack of a suitable dataset. In this work, we have contributed a parallel corpus of low-resource language pairs, English-Nyishi, and reported comparative experiments on the baseline neural machine translation systems. The results are evaluated for English to Nyishi and vice-versa via well-known automatic evaluation metrics and manual evaluation.

Keywords. English-Nyishi; NMT; low-resource; corpus.

1. Introduction

Machine translation (MT) is a subset of natural language processing (NLP) that helps to convert one natural language into another. Based on resource availability, There are three basic categories for natural languages: high, medium, and low resources. The data or resources for training a natural language model can include a corpus from various online sources, native speakers, and computational resources. A language is classified as having limited resources that have limited online resources [1, 2]. This classification can also depend on the volume of data necessary to train [3]. Hence, it is difficult to classify a language in a high, medium, or low resource without much background work. Hence, various parameters need to be considered. Thorough data collection exercises need to be undertaken to classify a language into a certain category. Moreover, the precise interpretation of the low-resource language set constitutes a challenging research query itself. According to [4], a language is regarded as having a poor number of training instances present in the corpora is below 1 million. Here, by instances, we mean the data required to train the supervised model for the concerned purpose. In our case, it is the count of parallel sentences in the source and target language. Though we say that the corpora's size must be significant, it must be diverse as well on account of both language and structure. Structurally, the sentences must comprise all categories, including very-long sentences, medium, and very short sentences. The majority of world languages are identified in the low-resource category on account of resource availability. In the northeastern regions of India, low-resourced languages are very rarely investigated in MT since the limitation of the dataset, including Mizo [5, 6], Assamese [7], Manipuri [8] and Khasi [9].

1.1 *Nyishi language*

As per the census 2011, there are 22 scheduled languages in the regional languages of India that have been used in various states of the country and 99 non-scheduled languages, mostly tribal languages.[1] The Nyishi (also known as Nishi/ Nyishing/Nissi) language[2] falls under the non-scheduled language category. The population of the Nyishi speakers is 2,08,337. The indigenous speakers of the Nyishi language are known as the Nyishi people in the Arunachal Pradesh state of India. In India's northeastern region,

*For correspondence
Published online: 02 November 2023

[1]Regular Paper https://censusindia.gov.in/nada/index.php/catalog/42458/download/46089/C-16_25062018.pdf
[2]https://bit.ly/36pOsyO

Arunachal Pradesh is considered the most prominent state where 28 significant tribes, 100 sub-tribes, and 50 distinct languages and dialects are present. The World Language Atlas by UNESCO in Danger states that Nyishi comes under the type of endangered language under the mother tongue category. Nyishi belongs to the Sino-Tibetan Family of the Tibeto-Burman, grouped as the Tani branch of Trans-Himalayan; Languages with similar phonology, lexical and grammatical systems are classified as Tani language [10, 11]. The Nyishi language is tonal in nature with three categories: rising, neutral and falling, and the meaning of the word depends on its tone. There is no grammatical gender in Nyishi but adding a suffix in a particular word changes the gender [11]. The pronominal system in English distinguishes four types of words: person, singular/plural, gender, and nominative/objective. The pronominal system in the Nyishi language contains a complete range that categorizes the three persons (first, second and third) and differentiates between the three numbers: singular, dual, and plural [10]. There are seven vowels and eighteen consonants in Nyishi [10]. Nyishi community does not have a script of their own they have adopted modified roman alphabets to represent their unique phonemes and write down their language. The Nyishi language uses subject-object-verb as its declarative word order (SOV). Still, it follows varieties that show both SOV and SVO among the native speakers of Arunachal Pradesh's central parts. Figure 1 shows an example of a Nyishi sentence with its English translation.

Unlike English, Nyishi is an isolating-agglutinating language which means a one-word element can express several grammatical categories and each word consists of a single-word element [12]. Nyishi language is also known as a tonal language. It uses three tones: rising, neutral, and falling, and applicable to all of its vowels. Words with different tones can mean differently. For example, the word "Jinam" means "to give" in a rising tone, "to beat" in a neutral tone, and "to melt" in a falling tone. Nouns are distinguished for gender. Generally, quantifiers or numerals are used to indicate the number; otherwise, the plural word is used. However, plural markers are not found in Nyishi. Masculine and feminine genders are marked with the masculine marker pu/bu and the feminine marker ne

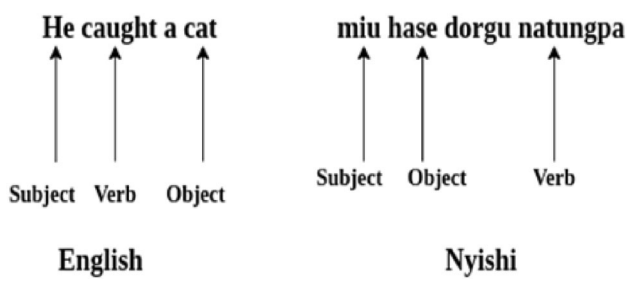

Figure 1. Example of English–Nyishi word order.

respectively. Verbs in Nyishi do not distinguish between numbers and person and the same form is used for the first-person, second-person, and third-person.

1.2 *Motivation*

Every language opines a unique worldview by characterizing its philosophy, culture, and identity. Whenever a language is extinct, there is an irredeemable loss of unique information of historical, ecological, spiritual and cultural significance. It directly affects the survival of its speakers and has an indirect effect on others. The Indian north-eastern languages are gradually shifting into the endangered category. The main reason is that native speakers go with the flow of language to get socio-economic benefits or lessen the fear of discrimination. The Nyishi language is one of these low-resource languages, hence it must be preserved. In a multilingual country like India, MT attempts to eradicate the linguistic barrier. MT is one of the widely accepted digitized techniques which helps to preserve such minority languages. Also, MT attempts to diminish the linguistic barrier in communication. Since English is a sophisticated language that is generally acknowledged around the world, we have thought about using it with Nyishi. It is essential for the English-to-Nyishi automatic translation system to facilitate effective communication at the national and international levels. Although, Bing[3] and Google[4] cover 110 and 133 languages worldwide, but Nyishi is yet to be introduced. Most of the researchers have been working on low-resource pairs in the MT research community using NMT-based approaches with different techniques, and most of the experiments witnessed an improvement in translation accuracy, such as the multi-lingual model [13], cross-lingual techniques [14] and with different Indian language [15–18] Another experiment compared the two approaches using NMT and SMT [6, 19, 20]. Further, to improve the translation accuracy for limited resource sources, the pseudo-parallel corpus method is introduced with an overall improvement in accuracy [21]. The NMT needs sufficient training parallel corpus, but there is no standard parallel corpus available for the English-Nyishi pair. In this work, we have attempted to promote the Nyishi language into the machine Translation environment, where automatic translations are performed from Nyishi to English and vice-versa. As far as we are aware, this low-resource combination has never been used in MT. The following is the paper's contributions:

- Created EnNyCorp1.0: a parallel corpus of low resource English–Nyishi pair
- Evaluated baseline systems for English-Nyishi pair translation examining various NMT models.

[3]Regular Paper https://www.bing.com/translator

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

- Achieved state-of-the-art accuracy using standard evaluation metrics. The work also presents an analysis of the variety of sentences.

The rest of the sections are arranged as follows: Sect. 2 briefly narrate model background. Sect. 3 outlines the related works. The details of corpus preparation are presented in Sect. 4. Sect. 5 and Sect. 6 present the baseline systems, experimental results, and analysis. Finally, Sect. 7 concluded with future direction.

2. Model background

In this section, we will briefly examine the mathematical fundamentals of both the transformer and NMT model using a global attention-based LSTM.

2.1 *Recurrent neural machine translation (unidirectional RNN and BRNN)*

Long short-term memory (LSTM) is used by RNN to represent long-term dependencies [22, 23]. Further, an attention mechanism is collaborated to alleviate the longer sentence dependency, this allows focusing on each segment of the source sequence as based on reference [24, 25]. The encoder of RNN starts the functioning by accepting the sequence of source data as x1, x2,...,x_m, which are further transformed into vector form vv. Decoder apply the conditional probability with source and target $P(\mathrm{y} \mid \mathrm{x})$ to produce target side sequence y1,y2,...,y_n in equation 1.

$$\mathrm{P}(\mathrm{y} \mid \mathrm{x}) = \sum_{i=1}^{n} p(y_i \mid y_{<1}, vv) \tag{1}$$

LSTM layer in RNN of the source and target side contains the hidden state as h_s and h_t which are further used to produce the attention vector a_t. Finally, the softmax layer is applied to generate the translated sentence at the target side of the decoder as shown in equation 2.

$$\mathbf{p}(y_i \mid y_{<1}, vv) = \mathrm{softmax}\left(W_s h_t'\right) \tag{2}$$

Where W_s represents the source side weighted value of all the hidden states and h_t' represent the hidden vectors of the attention layer. Figure 3 shows the RNN system architecture. BRNN allows two-sided RNN for forward and backward direction context calculation, which means previous and future words are equally considered at the same time. Based on reference [18], the BRNN-based model improves the accuracy of translation for limited-resource language sets such as English–Tamil and English–Hindi. furthermore, the convolutional neural network (CNN) is introduced in NMT [26–28] using transformer-based NMT which attempts to encode each location, employ a self-attention technique to link two distinct words, and then parallelize them to quicken learning. The NMT acquires a state-of-the-art in translation on different language sets that include English–German, English–French [29] since capable of context analysis and handling long-term dependency issues.

2.2 *Transformer model*

The key component of the transformer model is the self-attention mechanism [28]. The transformer model supports parallelization and minimizes long-term dependency issues by employing an additional attention layer. The query (Q), key (K) and value (V) are input vectors that are used to calculate the single attention operation, with the dimension of the key matrix dk as follows:

$$\text{Attention } (\mathrm{Q}, \mathrm{K}, \mathrm{V}) = \mathrm{softmax}\left(\frac{QT^k}{\sqrt{d_k}}\right)\mathrm{V} \tag{3}$$

The transformers use the multiple attention head with a weighted matrix of each query, key and value as follows:

$$\mathrm{Head}_i = \text{ Attention } \left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{4}$$

In our experiment, the default configuration with a 6-layer with eight attention heads is used as follows:

$$\begin{aligned} &\text{MultiHead } (Q, K, V) \\ &\quad = \text{concat}(\text{ head }_1, \text{ head }_2, \ldots \text{ head }_8) W^0 \end{aligned} \tag{5}$$

3. Related work

Due to an absence of studies and online resources for both parallel data and monolingual, research on automatic translation employing the Nyishi language is severely constrained for now. The initial machine translation work in this language uses 30.000 parallel sentences using SMT phrased-based and the translation accuracy is measured using BLEU, NIST and human evaluation [30]. The parallel corpus used in SMT phrase-based is a manual collection comprised of several domains of different lengths. The translation was done in both the forward and backward direction with and without tuning application and the results with tuning in the uni-gram show better. We have not come across any MT work on Nyishi with any other language pair than this method in the literature review. However, there are plenty of MT works on Indian low-resource languages that include English–Mizo [6, 31], English–Tamil [32], English-Punjabi [33], Punjabi–Hindi [16], Hindi–Nepali [34], Hindi–Marathi [35], English-Hindi [34], English–Assamese [7]. Existing works show that NMT-based approaches show better translation quality than a contemporary SMT-based approach in the case of both low and high-resource languages [22–24, 29, 36].

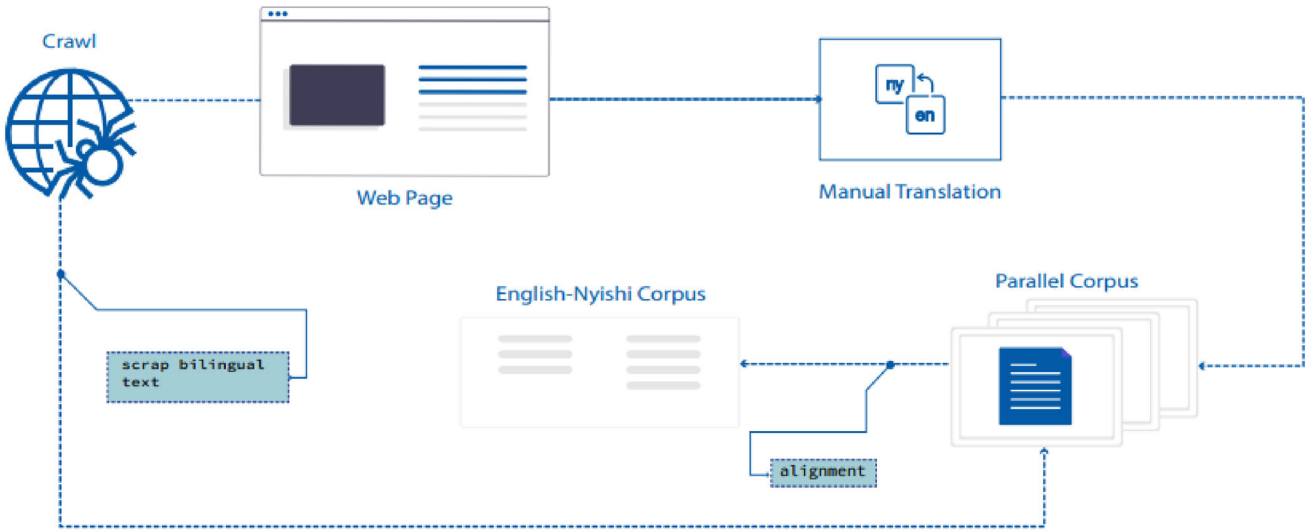

Figure 2. Corpus acquisition for English-Nyishi parallel corpus.

Because the NMT system can represent the data in continuous space, such representation effectively captures the essential properties. Moreover, NMT can handle the problem of long-term dependency and can analyze the context. Besides this, NMT gives end-to-end solutions, unlike SMT. There are many research scopes in Nyishi for the development of MT since it is in the outset stage. Inspired by the performance of NMT and the widely accepted approach by most of the researchers for low-resource pair translation [13, 14, 17, 21], we have investigated NMT on a self-curated dataset.

4. EnNyCorp1.0: English-Nyishi parallel corpus

The low-resource English–Nyishi (en-ny) pair has limited available options for parallel corpus. We have explored two viable resources to collect a total of 62,474 parallel sentences. First, extracted 26,184 parallel sentences from the Bible[5] by utilizing the web crawling technique[6] Secondly, English (en) sentences of 34,092 collected from various web pages/blogs and then manually translated to corresponding Nyishi sentences. Moreover, parallel dictionary words and sentences of 2198 are manually collected from the available online sources [10]. This manual translation process took the first author 3-4 months. The author preparing the dataset is an indigenous speaker of the Nyishi language and has a good command of English. Figure 2 presents the overall corpus acquisition process. The Nyishi sentences have been verified by hiring a linguistic expert who is native and possesses linguistic knowledge of the Nyishi language. The parallel source-target sentences are divided into a train set, validation set and test set, the details of which are described in table 1 During the training process, training data is required to learn the parameters and validation data to check the model's performance and select the best model. After the training process is over, using the test results, the model's performance is verified. We have considered two test sets, i.e., test set-1 and test set-2. Test set-2 is the subset of test set-1 and consists of 100 sentences, mainly for human evaluation.

Table 1. Training, validation, and test set data statistics.

Type	Sentences	English tokens	Nyishi tokens
Train	58,474	703,718	662,245
Validation	3,000	61,912	52,288
Test set-1	1,000	21,752	18,785
Test set-2	100	1,867	1,638

5. Baseline system

The OpenNMT-py toolkit is used to carry out the experiments [37] which is freely available.[7] We have built three NMT baseline systems (NMT-1, NMT-2, NMT-3) separately for each direction, i.e., English to Nyishi and Nyishi to English translation. NMT-1 and NMT-2 are unidirectional recurrent neural networks (RNN) and bidirectional RNN (BRNN) with attention mechanism [24] NMT-3 is based on the transformer model [28].

[5]Regular Paper https://www.bible.com/

[6]https://scrapy.org/

[7]Regular Paper https://github.com/OpenNMT/OpenNMT-py

5.1 *Data prepossessing*

The procedure for organizing data before it is put into training is referred as data preprocessing. This is done by creating dictionaries that link the source and target vocabularies to the respective indexes. The parallel data set from the prior experiment phrase-based SMT system [30] was coupled with newly gathered data to improve translation accuracy because there aren't enough corpora for Nyishi. After that, the data were divided into three groups: training, testing, and validation. In this step, the target and source sentences are tokenized using the OpenNMT-py toolkit to construct a dictionary used for indexing each word during the training procedure. The size of the vocabulary is 50,002 and 37,938 for the source (English) and target (Nyishi) sentences.

5.2 *System training*

In the heart of NMT architecture, the encoder and decoder are the key ingredients. During NMT-1, NMT-2 the training process, parallel source-target sentences are fed into the seq2seq model. The main difference between NMT-1 and NMT-2 is that NMT-2 uses two independent encoders: one for the forward sequence and the other for the backward sequence. On the other hand, NMT-1 uses only one encoder used for the forward sequence. We have used two-layer LSTM-based seq2seq modeling with attention in NMT-1 following [23] and NMT-2 following [24]. The learning rate Adam optimizer is 0.001 and a default drop out 0.3 are used. For illustration, the source Nyishi sentence "ngo sam param" is shown in figure 3 RNN model predicts the target English sentence *"I won't cut it"*

Although RNN-based NMT improves the translation quality, it has parallelization limitation issues. The transformer model offers parallelization and long-term dependence [28]. In the transformer model powerful mechanism of self-attention is used to interact with the words each other ("self"). Figure 4 the transformer-based encoder and decoder model is presented with the predicted sentence *"I won't cut it"* and source sentence "ngo sam param". The transformer model contains stacked self-attention and completely interconnected layers: a feed-forward network and a multi-head self-attention mechanism are two sub-layers. The transformer model computes numerous attention blocks in relation to the source instead of single computing attention at a time as given in figure 5.

The default configuration with a 6-layer with eight attention heads, and 0.1 drop-outs, and It employs the Adam optimizer with a 0.001 learning rate by default. The

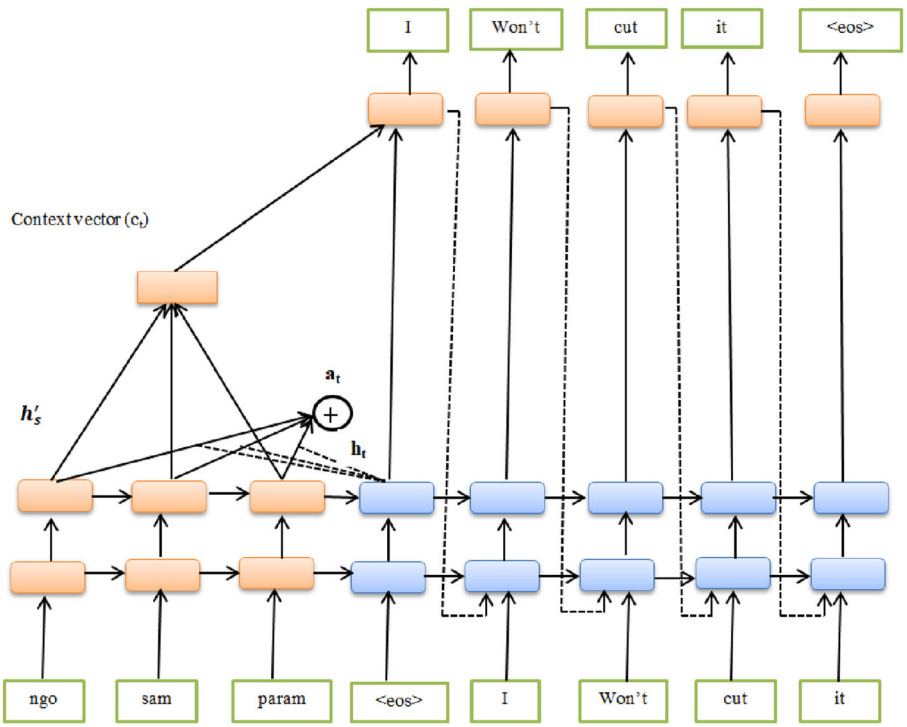

Figure 3. RNN model-based NMT system architecture.

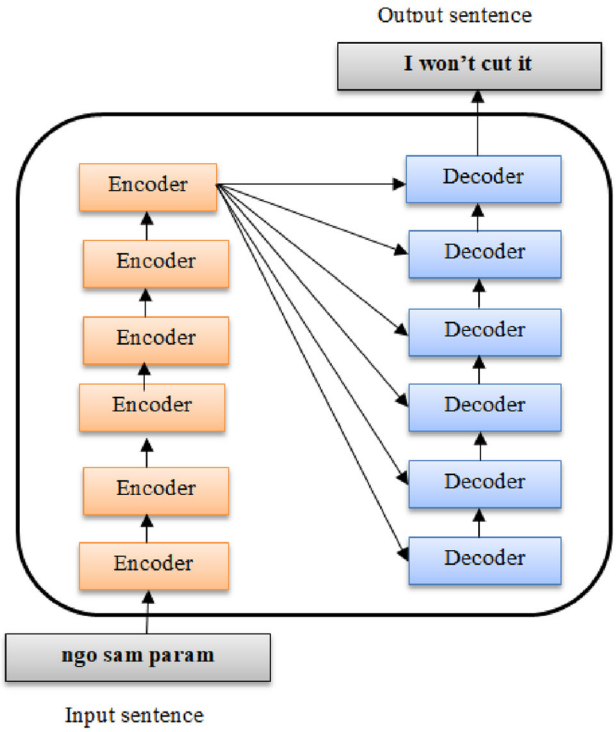

Figure 4. Transformer Architecture.

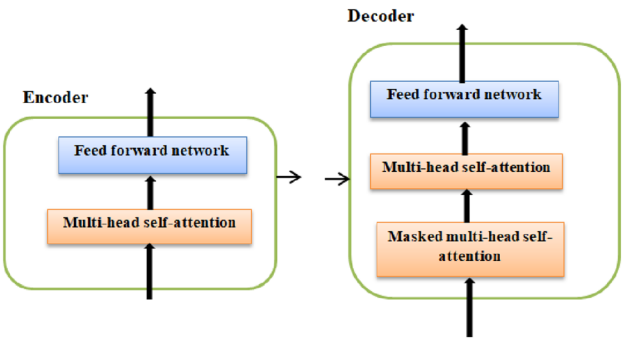

Figure 5. Component of Transformer model.

models are trained for up to 50,000 epochs on one NVIDIA Quadro P2000 GPU.

5.3 *System testing*

The most effective training model discovered during the training phase is used to construct the prediction sentence based on the test data. To select the optimal translation, we employed a beam search technique with a default size of 5.

Table 2. English-to-Nyishi automatic evaluation on Test set-1.

Metric	NMT-1	NMT-2	NMT-3
BLEU	9.47	9.99	10.24
TER	89.9	84.1	83.2
METEOR	0.1481	0.1492	0.1513
F-Measure	0.3379	0.3512	0.3566

Table 3. English to Nyishi automatic evaluation on Test set-2.

Metric	NMT-1	NMT-2	NMT-3
BLEU	9.73	10.18	11.67
TER	92.1	82.4	81.1
METEOR	0.1468	0.1503	0.1559
F-Measure	0.3322	0.3473	0.3621

Table 4. Nyishi to English automatic evaluation on Test set-1.

Metric	NMT-1	NMT-2	NMT-3
BLEU	15.41	15.43	15.60
TER	83.4	79.6	75.3
METEOR	0.1898	0.1920	0.1934
F-Measure	0.4286	0.4396	0.4448

6. Experimental results and analysis

The predicted translations obtained from the baseline systems, have been evaluated based on Metrics for automatic evaluation and human evaluation. The metrics used for automatic evaluation include, bilingual evaluation under study (BLEU) [38], translation edit rate (TER) [39], metric for evaluation of translation with explicit ordering (METEOR) [40] and F-measure. In tables 2, 3, 4, 5 we have reported experimental results based on automatic metrics for English to Nyishi and vice-versa translation. Also, human evaluations are presented in tables 6, 7, 8.

6.1 *Automatic evaluation metrics*

6.1.1 *BLEU* The BLEU metric uses the modified precision approach (n-gram) to calculate the score between the predicted translation and the reference translation. The number of $n-grams$ in the candidate translation and the reference translation are compared. The formula for the calculation of the BLEU score is presented

Table 5. Nyishi to English automatic evaluation on Test set-2.

Metric	NMT-1	NMT-2	NMT-3
BLEU	14.66	15.47	15.61
TER	89.6	81.8	79.1
METEOR	0.1943	0.1961	0.1971
F-Measure	0.4194	0.4399	0.4490

Table 6. Scores of human evaluation using NMT-1 on Test set-2.

Translation	Adequacy	Fluency	Overall rating
English to Nyishi	27.10	50.20	34.78
Nyishi to English	29.80	50.74	35.18

Table 7. Human evaluation scores of NMT-2 on Test set-2.

Translation	Adequacy	Fluency	Overall rating
English to Nyishi	27.34	38.50	32.92
Nyishi to English	28.10	43.50	35.80

Table 8. Human evaluation scores of NMT-3 on Test set-2.

Translation	Adequacy	Fluency	Overall rating
English to Nyishi	24.10	47.20	35.65
Nyishi to English	27.12	48.50	37.81

in Equation 6. Here, PL, RL, Pre_i represent the length of the predicted, reference translation and i^{th}, gram matching precision score. From the 6, it is realized that the net BLEU score will be 0 if Pre_i is 0.

$$BLEU = \min\left(1, \frac{PL}{RL}\right)\left(\sum_{i=1}^{n} \mathrm{Pre}_i\right)^{\frac{1}{n}} \tag{6}$$

6.1.2 *TER* TER metric measures the required minimum editing effort on a predicted translation of system output that matches a reference translation. Mathematically, it can be represented using Eq. 7

$$TER = \frac{N_e}{N_{rw}} \tag{7}$$

Here, N_e, N_{rw} represent number of edits and reference words. Consider the example: Reference sentence: John is a good boy. Predicted sentence: John a is boy. In this example, "a" and "is" are shifted in the predicted sentence concerning the given reference sentence. Also, *" good"* is missing in the predicted sentence. So, we can apply TER on the predicted and reference sentence, the number of edits is 3 (2 shifts and 1 insertion), which gives TER score of $\frac{3}{5} = 60\%$

6.1.3 *METEOR and F-measure* Unlike BLEU, METEOR and F-measure scores consider recall by matching the n-gram count among candidate and reference translation since recall is an important parameter to check translation quality. The METEOR score is calculated based on matching exact, stemmed, and synonyms between the candidate and reference sentences. The F-measure score is generated by the harmonic means of recall and precision. The range of both METEOR and F-measure lies from 0 to 1. The difference between F-measure and METEOR is that the former takes into account unigram matching, whereas the latter considers higher n-gram matching. The absence of matching lowers the scores. Consider, r_t and p_t are the number of unigrams in reference, predicted translation, and T_m be the total number of matched unigrams in between reference and predicted translations. Then, unigram precision p_u recall r_u and F-measure can be represented by the given Eq. 8, Eq. 9, Eq. 10 and Eq. 11.

$$P_u = \frac{T_m}{p_t} \tag{8}$$

$$R_u = \frac{T_m}{r_t} \tag{9}$$

$$F - measure = \frac{2 \times P_u \times R_u}{P_u + R_u} \tag{10}$$

The METEOR score can be calculated using Equation

$$METEOR = (1 - Pen) \times F_{mean} \tag{11}$$

Here, F_{mean} is the extension of Equation 10 parametric harmonic mean of precision P_u, recall R_u. The pen represents the penalty which is used to ensure the word order of predicted and reference sentences. It is calculated by Equation 12. Also, Frag denotes fragmentation, and it is obtained by the division of chunks (collection of matched unigrams) with T_m.

$$Pen = \gamma \times Frag \tag{12}$$

6.2 *Human evaluation*

Human evaluation (HE), also known as manual evaluation, is another option to evaluate the predicted sentence machine translation quality by manual judgment. Since automatic evaluation metrics fail to assess all important aspects of translation quality. In the HE, the human evaluator is hired and familiar with the concerned languages and various evaluation aspects. There are three main aspects of human evaluation, adequacy, fluency, and overall rating. The adequacy aspect refers to measuring predicted translation quality based on the amount of meaning corresponding to the reference translation. Fluency evaluates how accurately the predicted sentence is formed. It is not relevant to the source sentence. The overall rating considers both adequacy and fluency by the average computing score [6] Let us consider, given reference translation: *"He wakes up early in the morning,"* and the predicted translation: *"He is flying to Delhi"* is inadequate, as it represents the different contextual meaning of reference translation. But, the translation is fluent since it is a well-formed sentence of reference language. The rating score is evaluated on a scale of 0-5, with a higher value signifying better for 100 test sentences in human evaluation.

6.3 *Analysis*

From the automatic and human evaluations, it is observed that NMT-3 slightly outperforms NMT-1 and NMT-2. We have considered samples of predicted sentences from various perspectives to inspect all the systems' translation accuracy. We have used the following notations in the examples.

- English Test (ET): English Test sentence
- Nyishi Gold (NG): Reference/Gold Nyishi sentence.
- Nyishi Predicted (NP): Predicted sentence in Nyishi.
- Nyishi Test (NT): Test sentence in Nyishi language.
- English predicted (EP): Predicted English sentence.
- English Gold (EG): Reference/Gold sentence in the English language.

1. Example of **inadequacy but goog fluency** (English-to-Nyishi)

ET: *"Any questions?"*
NG: *"Hoq tas dopayv?"*
NP1 (NMT-1): *"Hiyvkam donu manum."*
NP2 (NMT-2): *"Nyi anyio go cengconam."*
NP3 (NMT-3): *"golo Akin num,"*
2. Example of **inadequacy but good fluency** (Nyishi-to-English)
NT: *"Hoq tas dopayv?"*
EG: *"Any questions?"*
EP1 (NMT-1): *"Have any loom?"*
EP2 (NMT-2): *"Is there a message?"*
EP3 (NMT-3) : *"Is that accident here?"*
Discussion: For the above two examples, predicted translations of all the NMT systems inadequately represent reference translation. However, predicted translations are fluent since these are the well-formed sentence of Nyishi and the English language in Example 1 and 2, respectively.
3. Example of **partially adequate and perfectly fluent** (English-to-Nyishi)
ET: *"Tom seemed smart."*
NG: *"Tom nyilaq nyipaq bo hvb kaado."*
NP1: *"Tom nyiq pan hvb kaado."*
NP2: *"Tom nyaa pan hvb kaado."*
NP3: *"Tom kaadwb nyaapaku."*
4. Example of **partially adequate and perfectly fluent** (Nyishi-to-English)
NT: *"Tom nyilaq nyipaq bo hvb kaado."*
EG: *"Tom seemed smart."*
EP1: *"Tom seemed innovative."*
EP2: *"Tom seems innovative."*
EP3: *"Tom seems dazed."*
Discussion: For the above two examples, all the NMT systems' predicted translations are partially close to the reference sentences. Also, perfectly fluent in both the example sentences.
5. Example of **incorrect named-entity prediction**(English-to-Nyishi)
ET: *"And Isaac s servants digged in the valley, and found there a well of springing water."*
NG: *"Hvbmabde hwdlo Aisakwg nyera tulungv lywpa ho iss buqlin bo sokum mam dukoql dupa yin,"*
NP1: *"Hoo Aisak gv nyera tulungv lywpa uraho doonum."*
NP2: *"Hoo Aisak mwwg nyera tulungam Al Sa-svlo ho, hoo iss hingkum bo ak ham kaapanum."*
NP3: *"Hoo Aisaknvn Jehoas mwwg nyera-paqn tulungam iss subuq buqlin koob nyinum."*
6. Example of **incorrect named-entity prediction** (Nyishi-to-English)
NT: *"Hvbmabde hwdlo Aisakwg nyera tulungv lywpa ho iss buqlin bo sokum mam dukoql dupa yin,"*
EG: *"And Isaac s servants digged in the valley, and found there a well of springing water."*
EP1: *"But it shall come to pass, when there shall come to pass, that, when the servants of repentance is come to the mountain,"*
EP2: *"And it is written, that the servants which are upon the valley of servant is upon the plain."*
EP3: *"And it shall come to pass, when our servants know that which is in the valley."*
Discussion: For the above two examples, all the systems cannot predict the named-entity words/noun in both the direction of translations. The translations yield inadequacy and poor fluency of translation.

7. Example of **over-prediction** (English-to-Nyishi)
ET: *"Laughed the Queen with delight."*
NG: *"hvl lvpw atwg nywxw ak hv hempwl nywrnum."*
NP1: *"Lwyanam hv hemp mvdwn."*
NP2: *"Poaq lvgab vcjaqb baqmin gvnum."*
NP3: *"Vm nyij-muje haang doogdvn."*
8. Example of **over-prediction** (Nyishi-to-English)
NT: *"hvl lvpw atwg nywxw ak hv hempwl nywrnum."*
EG: *"Laughed the Queen with delight."*
EP1: *"As the king s wife was going to take his wife."*
EP2: *"It pleased the king s wife."*
EP3: *"And the king s boon of the king s wife."*
Discussion: For the above two examples, predicted translations of all the systems are over-predicted with different meanings (inadequacy) in both translations' directions. However, fluency is good.
9. Example of **under-prediction** (English-to-Nyishi)
ET: *"And the damsel ran, and told them of her mothers house these things."*
NG: *"Ho nyijwr akv mwnwg annwg nammwb xarbnv soq hog mwlw sam betamnum."*
EP1: *"Nyem nyem akv nyebia jaqb xarbnv bulam soq bon tulu sam bekinnum."*
EP2: *"Hoo nyem akv mwnwg ann hoqg nyem ko tulungam betamnum,"*
EP3: *"Hoo nyem akv mwnwn betamnum, Soq tulu sam bulv atwg nam tuluho betam kunum."*
10. Example of **under-prediction** (Nyishi-to-English)
NT: *"Ho nyijwr akv mwnwg annwg nammwb xarbnv soq hog mwlw sam betamnum."*
EG: *"And the damsel ran, and told them of her mothers house these things."*
EP1: *"Then the young woman took her back in his mother s house."*
EP2: *"And the woman took to go into her house, and told all it."*
EP3: *"And the woman took the house and the mother of her mother s house."*
Discussion: In the above two examples, all the systems suffer from under-translation, which means some parts of the source sentence are predicted only tables 9 and 10 present English-to-Nyishi and vice-versa translations by considering worst examples in three types of short, medium, and long sentences. The predicted sentences suffer from under-translation, over-translation, inadequacy, completely different contextual translation, and poor fluency. However, it is noted that NMT-3 shows better translation over NMT-1 and NMT-2, as shown in table 11 Moreover, it is found that the accuracy of translation from Nyishi-to-English is better than English-to-Nyishi. It is because the vocabulary contains fewer Nyishi words than English words, as mentioned in Sect. 5.1. the outcome shows that NMT systems encoded more English words than Nyishi words, and the decoder could generate better Nyishi-to-English translation. Apart from this, it is noticed that few predicted sentences achieve very good human evaluation scores in comparison to automatic evaluation (BLEU) scores as shown in table 12 This remark that standard evaluation metrics like BLEU needs improvement in the adequacy and fluency factor of

Table 9. Worst example for English to Nyishi translation.

Type	Source	Reference	Predicted
Short	"Come in."	"hato wurab."	"sob in."
Medium	"And Laban said to him, Surely thou art my bone and my flesh. And he abode with him the space of a month."	"hoo Laban mwam benum, Jvqtw jaqb no ngoqg soodin soqgv lengn bov! Hoo mwv pool bargwb mwam doomin gvnum."	"Hoo mwv mwam benum, Hvbmabde no ngoqg loobungv hoo ngoqg ajin tulungv dooba do."
Long	"And before I had done speaking in mine heart, behold, Rebekah came forth with her pitcher on her shoulder; and she went down unto the well, and drew water: and I said unto her, Let me drink, I pray thee."	"Ngoqg hang uraho hoqhvb goodunamv goonya maatab, kaato, Rebeka mwnwg uppum mam mwnwg gorbw ho joolayil haanum, hoo mwn sokumb haabnv iss suqnum. Ngo mwnwn benum, 'Anya mwwpa gvdwl ngam iss loq twwm tvb."	"Hoo ngoqg hang bekwb goonam mam ngo tapanum, kaato, mwnwg pitcher haaknum. Mwnwg shoulder; haanam mam mwnwg lvpia ho haanum. Hoo mwn mwnwn benum, Ngam wwm tvb."

Table 10. Worst example for Nyishi to English translation.

Type	Source	Reference	Predicted
Short			
	"Mwv tvbab gwwiyam hoonum,"	"He offered the milk to the snake and then,"	"He used to play bread."
Medium			
	"Nenibi gvla Kala pengko ho Resenam mvnum; hoqhv koibo bopam gob nyinum."	"And Resen between Nineveh and Calah: the same is a great city."	"And the battle and the midst of the midst of the midst of the great."
Long			
	"Lwwx la ak vnywg otuq poolwg otuq aalho swcw ao hoqgv issiv xootwr kunum. Hoo Noa hulungam lwqko tvl hoo kaanum, hoo kaanamv, ked aongv pwtw kunum.	"And it came to pass in the six hundredth and first year, in the first month, the first day of the month, the waters were dried up from off the earth: and Noah removed the covering of the ark, and looked, and, behold, the face of the ground was dry."	"And in the first day of the first month, in the first day of the first month, by the first day of the first month, in the first day of the first month, and the looked, and looked, and looked, and looked, and looked, and looked, and looked, and looked, and looked, and looked, and"

Table 11. Best example of NMT-3 over NMT-1, NMT-2 translation.

Translation	Source	Reference	System	Predicted
Nyishi-to-English	"Aal loonyi kokwso"	"Two days later"	NMT-1	"After two days came"
			NMT-2	One day
			NMT-3	"After the second day"
English-to-Nyishi	"he thought"	"mwv mwwgamv"	NMT-1	"bonam"
			NMT-2	"Mwwe mwwgab pa"
			NMT-3	"Mwwe mwwpa"

Table 12. Example sentence for comparison of human evaluation with BLEU Score.

Translation	Reference	Predicted	Automatic evaluation (BLEU)	HE (Overall rating)
English to Nyishi	"Ngo mwlwngam loomin gvma"	"Ngo al gab loomin gvma"	28.33	5
Nyishi to English	"I don't quite agree"	"I can't quite agree"	36.10	5

translation especially for Indian low-resource tonal languages.

7. Conclusion and future work

In our work, we have performed a challenging task of low-resource language set translation, English- Nyishi, using the state-of-the-art MT approach. We have prepared EnNyCorp1.0, the parallel corpus of English-Nyishi, a new language pair in the machine translation task. The corpus will be publicly available for the non-commercial and academic research purposes. Using the three main NMT models-RNN, BRNN, and Transformer models-we have assessed baseline systems, both forward and backward directions of translation. Moreover, we have analysed the predicted translations by considering different perspectives. In future work, we aim to add more data and comparatively analysis the accuracy of predicted result. Also, we investigate the word-segmentation and the knowledge-transfer-based NMT approach, encountering low-resource challenges for further research.

References

[1] Karine M and Dan P 2008 Low-density language bootstrapping: the case of Tajiki Persian. In: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pp. 3293–3298

[2] Katharina P, Ralf D B, Jaime G C, Alon L, Lori L and Erik P 2001 Design and implementation of controlled elicitation for machine translation of low-density languages. In: *Workshop on MT2010: Towards a Road Map for MT*

[3] Jiatao G, Hany H, Jacob D and Victor O L 2018 Universal neural machine translation for extremely low resource languages. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, pp. 344–354

[4] Tom K 2020 Exploring benefits of transfer learning in neural machine translation. in: *Computation and Language (cs.CL)*, pp. 1–150

[5] Candy L, Badal S and Partha P 2021 An improved English-to-Mizo neural machine translation. *Transactions on Asian and Low-Resource Language Information Processing* 20(4): 1–21

[6] Amarnath P, Partha P and Jereemi B 2019 English-mizo machine translation using neural and statistical approaches. *Neural Computing and Applications* 31(11): 7615–7631

[7] Sahinur RL, Abdullah FURK, Partha P and Sivaji B 2020 Enascorp1. 0:English-assamese corpus. In: *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pp. 62–68

[8] Salam MS and Thoudam DS 2020 Unsupervised neural machine translation for english and manipuri. In: *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pp. 69–78

[9] NDonald JT and Bipul SP 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*, vol. 170, p. 3

[10] Pierre T A 2005 A Grammar of Nyishi Language. Farsight Publishers and Distributers, Delhi, pp 1–134

[11] Mark WP 2015 Tones in northeast indian languages, with a focus on tani: A fieldworker's guide. In: *Language and culture in Northeast India and beyond: In honour of Robbins Burling*, pp. 182–210

[12] Moumita D 2018 Negation in Nyishi. *NEHU Publication*, pp. 80–100

[13] Xinyi W, Yulia T and Graham N 2020 Balancing training for multilingual neural machine translation. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8526–8537

[14] Guillaume L and Alexis C 2019 Cross-lingual language model pertaining. In: *NIPS'19: Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pp. 7059–7069

[15] Himanshu C, Shivansh R and Rajesh R 2020 Neural machine translation for low-resourced Indian languages. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association*, pp. 3610–3615

[16] Karthik R, Kaushik T and Shrisha R. 2017 Neural machine translation of Indian languages. In: *Proceedings of the 10th Annual ACM India Compute Conference*, pp. 11–20

[17] Surafel M L, Matteo N and Marco T 2020 Low resource neural machine translation: A benchmark for five African languages. *Africa NLP workshop at ICLR* 2020: 1–10

[18] Sree H R and Krishna P S 2018 Neural machine translation for low resource languages using bilingual lexicon induced from comparable corpora. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pp. 112–119

[19] Sukanta S, Mohammed H, Asif E, Pushpak B and Andy W 2021 Neural machine translation of low-resource languages using smt phrase pair injection. *Natural Language Engineering* 27(3): 271–292

[20] Vikrant G, Sourav K, and Dipti MS 2020 Efficient neural machine translation for low-resource languages via exploiting related languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pp. 162–168

[21] Aizhan I, Takayuki S and Mamoru K 2019 Filtered pseudo-parallel corpus improves low-resource neural machine translation. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)* 19(2): 1–16

[22] Kyunghyun C, Bart VM, Caglar G, Dzmitry B, Fethi B, Holger S and Bengio Y 2014 Learning phrase representations using rnn encoder-decoder for statistical machine translation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734

[23] Ilya S, Oriol V and Quoc V L 2014 Sequence to sequence learning with neural networks. In: *NIPS'14: Proceedings of the 27th International Conference on Neural Information Processing Systems*, vol. 2, pp. 3104–3112

[24] Dzmitry B, Kyunghyun C and Yoshua B 2014 Neural machine translation by jointly learning to align and translate. In: *3rd International Conference on Learning Representations ICLR 2015*, pp. 1–15

[25] Minh-Thang L, Hieu P and Christopher D M 2015 Effective approaches to attention-based neural machine translation. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1412–1421

[26] Nal K and Phil B 2013 Recurrent continuous translation models. In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1700–1709

[27] Jonas G, Michael A, David G and Yann D 2016 A convolutional encoder model for neural machine translation. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, vol. 1, pp. 123–135

[28] Ashish V, Noam S, Niki P, Jakob U, Llion J, Aidan NG, Łukasz K and Illia P 2017 Attention is all you need. In: *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 6000–6010

[29] Myle O, Michael A, David G and Marc'AR 2018 Analyzing uncertainty in neural machine translation. In: *International Conference on Machine Learning, PMLR*, pp. 3956–3965

[30] Kakum N and Sambyo K 2022 Phrase-based English-Nyishi machine translation. In: *Pattern Recognition and Data*

Analysis with Applications, Springer Nature Singapore, Singapore, vol. 888, pp. 467–477

[31] Amarnath P and Partha P 2019 Neural machine translation for Indian languages. *Journal of Intelligent Systems* 28(3): 465–477

[32] Himanshu C, Aditya KP, Rajiv RS and Ponnurangam K 2018 Neural machine translation for English-Tamil. In: *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pp. 770–775

[33] Shivkaran S, Anand Kumar M and Soman K P 2018 Attention-based English to Punjabi neural machine translation. *Journal of Intelligent & Fuzzy Systems* 34(3): 1551–1559

[34] Sahinur RL, Abinash D, Partha P and Sivaji B 2019 Neural machine translation: English to Hindi. In: *IEEE Conference on Information and Communication Technology*, pp. 1–6

[35] Sahinur RL, Abdullah Faiz Ur RK, Partha P and Sivaji B 2020 Hindi-Marathi cross lingual model. In: *Proceedings of the Fifth Conference on Machine Translation*, pp. 396–401

[36] Kyunghyun C, Bart VM, Dzmitry B and Yoshua B 2014 On the properties of neural machine translation: Encoder-decoder approaches. in: *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pp. 103–111

[37] Guillaume K, Yoon K, Yuntian D, Jean S and Alexander MR 2017 Opennmt: Open-source toolkit for neural machine translation. In: *Proceedings of ACL 2017, System Demonstrations*, pp. 67–72

[38] Kishore P, Salim R, Todd W and Wei-Jing Z 2002 Bleu: a method for automatic evaluation of machine translation. In: *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318

[39] Matthew S, Bonnie D, Richard S, Linnea M and John M 2006 A study of translation edit rate with targeted human annotation. In: *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pp. 223–231

[40] Alon L and Michael J D 2009 The meteor metric for automatic evaluation of machine translation. *Machine Translation* 23(2): 105–115