A Neural based Bidirectional MT System to Investigate the Performance of the Low Resource Language pair English-Nepali

AMIT KUMAR ROY[1]
BIPUL SYAM PURKAYASTHA[2]
CHANAMBAM SVETA DEVI[3]
SAPTARSHI PAUL[4]

Department of Computer Science
Assam University, Silchar
Cachar, Assam, India

[1]amitroy.cs@gmail.com
[2]bipul_sh@hotmail.com
[3]chsveta91@gmail.com
[4]paulsaptarshi@yahoo.co.in

Abstract. The prime objective of an intelligent system would have been to be able to interpret and communicate reliably between natural languages. Neural Machine Translation (NMT) has demonstrated a possible solution to the machine translation problem with its recent growth. Despite NMT's high data requirements, systems using different NMT models for low-resource languages have recently shown some astounding outcomes. Given the dearth of digitally accessible text datasets, Nepali is one of these low resource languages. In our study, we explore a bidirectional neural machine translation (NMT) system for the low-resource English-Nepali language pair. A parallel text corpus with over 20,500 sentences and the open-source OpenNMT toolkit are used in the construction of the system. Our NMT system's efficacy is evaluated using the automated evaluation metrics, BLEU and METEOR. For the English to Nepali, the system scored 19.83, 53.32 BLEU and METEOR, and for the Nepali to English, that is 21.94 and 57.52.

Keywords: NLP, MT, NMT, Low Resource Language pair, Nepali-English.

(Received July 19th, 2023 / Accepted July 1st, 2024)

1 Introduction

The purpose of machine translation (MT) is to translate text across natural languages using a computational method without the need for human intervention. Although the complexity and methodology of MT systems might vary, they often fall into three primary categories: Rule-based MT, Corpus-based MT, and Hybrid MT. The SMT and NMT are included in the Corpus-based MT. Based on an end-to-end architecture, NMT Systems has made module integration and design simpler. Nonetheless, the modular foundation of a log linear model underpins SMT systems [9]. NMT systems require larger volumes of training data in order to become proficient; in environments with limited resources, this leads to lower-quality translations, while in situations with sufficient data, performance is greater [10]. NMT models the entire translation process using deep learning techniques, specifically neural networks. Because it yields translations that are more accurate and fluid, this method has risen to the top in recent years. The system is trained on a bilingual parallel corpus.

Nepali is recognized as an official language in our neighbouring country, Nepal and the northeastern state of Sikkim in India. In some places, it's also called "Khas Kura" or "Gorkhali." Nepali is a member of the Indo-Aryan language family's Eastern Pahari subbranch. There are 15,226,168 Nepali speakers living in Sikkim, Assam, Manipur, West Bengal, and Nepal. Using the Subject-Object-Verb (SOV) sentence scheme, Nepali is written in the Devanagari script [15]. There are 30 consonants, 11 independent vowels, and 10 dependent vowels in the Nepali language. Every sentence is concluded with the "purna biram" (।). Consonants are combined with dependent symbols such as matras and halanta to produce compound letters. [2].

Here, we developed a parallel dataset of approximately 20,500 sentences with the help of a native Nepali speaker, for training and testing of the MT system. This dataset was then utilized to train Neural Machine Translation (NMT) systems under different configurations. Automatic evaluation measures, such as BLEU [12], Recall [7] and METEOR [3], were used to evaluate the performance of the MT system. These automatic matrices were used to evaluate machine-generated translations and estimate the success of our method for the Nepali language.

The paper is organized as follows: Section-2 reviews previous research conducted on the Nepali language. Section-3.1 details our bilingual corpus and the pre-processing steps. Section-3.2 describes the NMT architecture we used, while Section-3.3 explains the system evaluation metrics. The experimental results and system assessment are presented in Section-4. Finally, Section-5 discusses the conclusion and future work.

2 Previous Works

Dobhase was the first machine translation project for Nepali and English, initiated in 2006. This rule-based system analyzed and parsed input strings in the source language, constructed the target language's syntax, and produced the translation. However, it was discontinued because it couldn't handle sentences with complex structures. Currently, the system has a bilingual dictionary containing 22,000 words. [4].

H. K. Shrestha [16] conducted further research in Nepali machine translation. In this study, the sentence's structural features in the target language were aligned with the root word and its related features in the source language (Nepali) following syllable segmentation and tokenization of Nepali text.

In 2018, a new machine translation (MT) system was created using the Statistical Machine Translation (SMT) method to translate sentences in English into their most likely Nepali. This system was trained on approximately 5000 parallel sentences, and 100 English sentences were used to test its performance, with the system evaluated based on fluency and sufficiency criteria via manual assessment. The examination resulted in an accuracy score of 2.7 out of a maximum of 4 [13].

In the subsequent study, P. Acharya [1] examined two primary techniques for the Nepali-English language pair, namely Neural Machine Translation (NMT) and Statistical Machine Translation. They used a small parallel corpus collected from the Nepali National Corpus (NNC), which contained 6535 sentences. Interestingly, SMT received a higher BLEU score than NMT in their examination. SMT obtained a maximum BLEU score of 5.27, whereas NMT received a score of 3.28.

In 2019, S. R. Laskar's NMT paper [11] explored the application of an attention mechanism to enhance the translation process between closely related languages Hindi-Nepali, which was presented at WMT19. The NMT systems were trained on a parallel corpus and assessed for translation quality in both the Hindi-to-Nepali and Nepali-to-Hindi directions. According to the official results, the system obtained remarkable BLEU scores of 53.7 (Hindi to Nepali) and 49.1 (Nepali to Hindi) in the contrastive system type.

Another system was developed in 2020 with a corpus of 15,000 sentences that were manually aligned or categorized for the language pair Tamang-Nepali. These sentence pairs were used to train an attention-based Transformer Neural Machine Translation (NMT) architecture. The results showed that the BLEU scores for the translations from Tamang to Nepali and from Nepali to Tamang were 27.74 and 23.74, respectively. [5].

3 System Description

The three primary phases of system architecture are preprocessing of data, training the system, and testing. The following steps have been thoroughly discussed in the subsections:

3.1 Data & its Pre-processing

The monolingual Bible corpus [18] and the agriculture domain corpus gathered from TDIL (Indian Language Technology Proliferation and Deployment Centre) were the sources of 20,727 English sentences that made up the bilingual corpus. The table shows statistics of the monolingual corpus. The following sentences were thoroughly translated from English into Nepali by native speakers and saved in two different text files that formed the parallel corpus. The corpus text files are shown in Figure 1. Pre-processing processes include cleaning, truecaseing, and tokenization of sentences and the NMT system is then trained using these pre-processed corpus files.

The first phase of this process is called tokenization, and it entails dividing the text into smaller chunks called tokens that are separated by whitespace. This facilitates context understanding and guarantees that the NMT system can correctly interpret the data's meaning for translation.

Further, truecasing is used for the tokenized data to lessen sparsity by translating uppercase to lowercase letters. However since Nepali does not have capital and lowercase letters, truecasing is meaningless.

Cleaning the data entails eliminating punctuation, non-printable letters, and long or empty sentences that can interfere with the training process and lead to alignment problems in translation.

Table 1: The Statistics of Monolingual Corpus

Cospus	*No.of Sentences*
Bible Text	17,500
Agriculture Domain Text	3227
Total	**20,727**

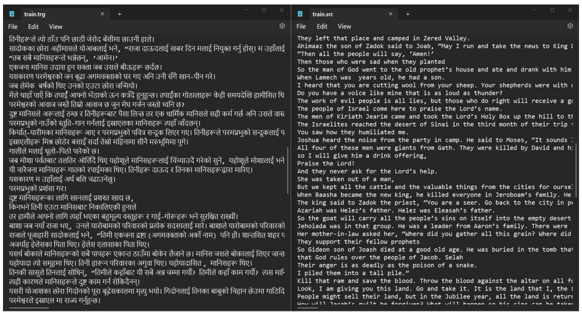


Figure 1: A view of the parallel corpus files

3.2 System Training and Testing

Typically modelling full sentences in a single integrated model, neural machine translation (NMT) is a machine translation technique that uses an artificial neural network to estimate the likelihood of a word sequence[17]. Here large neural networks are trained to translate natural languages end-to-end manner. The core components of an NMT system consist of an Encoder, which processes the input sentence and encodes it into a fixed-length context vector, and a Decoder, which takes the context vector produced by the encoder and generates the translated output sentence, one word at a time [6]. Here the Encoder and the Decoder are both 2-layer Long Short-Term Memory (LSTM) Recurrent Neural Networks (RNN) with 500 hidden units. The Long Short-Term Memory (LSTM) networks handle sequences of varying lengths and capture long-range dependencies. the architecture of the model is given below in Figure 2.

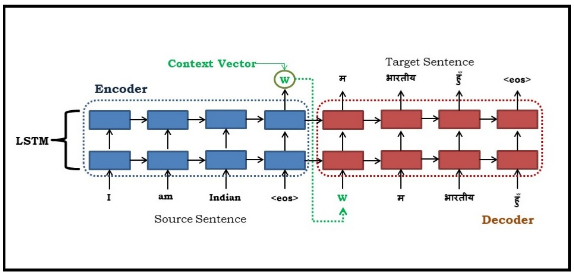


Figure 2: The NMT system architecture.

Here the source language sentences are fed into the encoder LSTM, which processes the sequence and produces hidden states. The final hidden state of the encoder LSTM is used as the context vector, representing the entire input sequence. The decoder LSTM takes the context vector and the start token of the target sequence as input. At each step, the decoder predicts the next word in the target sequence, using the previous word and the context vector. This process continues until the end-of-sequence token is generated.

3.3 System Evaluation Matrices

We use automatic evaluation techniques for system evaluation: BLEU, and METEOR.

3.3.1 BLEU

The BLEU (Bilingual Evaluation Understudy) score measures the quality of text that has been machine-translated from one language to another. It evaluates the machine-translated text (candidate translation) against one or more reference translations. The BLEU score goes from 0 to 1, with 1 representing a perfect match between the candidate and reference translations. In practice, it is frequently stated as a percentage (e.g., 0.75 = 75%) [12]. The following formula is used to calculate the BLEU score:

$$\mathbf{BLEU} = BP.exp(\sum_{N}^{n=1} w_n log p_n) \quad (1)$$

Here,

- p_n is the precision for n-gram.

- w_n, is the weight for each n-gram precision, typically set to $\frac{1}{N}$ for N n-grams, implying equal weight for each n-gram length.

- exp denotes the exponential function.

- BP is the Brevity Penalty, that avoids favoring shorter translations, given by,

$$BP = \begin{cases} 1, & if c > r \\ e^{(1-\frac{r}{c})}, & if c \leq r \end{cases} \quad (2)$$

The c is the length of the candidate translation and r is the length of the reference translation.

3.3.2 METEOR

The METEOR (Metric for Evaluation of Translation with Explicit ORdering) is a metric used to evaluate the quality of machine translation by comparing the machine-generated translation to a set of human-generated reference translations. Unlike BLEU, which relies heavily on precision, METEOR aims to improve correlation with human judgment by considering additional factors such as recall, stemming, synonymy, and word order. [3]. The following steps are used to calculate the METEOR score:

$$METEOR = F_{mean}.(1 - Penalty) \quad (3)$$

Here,

- $$F_{mean} = \frac{10.P.R}{R + 9P} \quad (4)$$

  Here, P(Precision) is the fraction of matched words in the candidate translation, and R (Recall) is the fraction of matched words in the reference translation.

- $$Penalty = 0.5 \left(\frac{Total\_Chunks}{Total\_Matches} \right)^3 \quad (5)$$

  Here, $Total\_Chunks$ is the number of contiguous matched word sequences, and $Total\_Matches$ is the total number of matched words.

4 Implementation and Result Analysis

After pre-processing, our corpus contains 20,347 sentences, which are used for training, validation, and testing. These sentences are then separated into three corresponding files. Table 2 depicts the distribution of the corpus files.

Table 2: Corpus Statistics

$Language$	$Type$	$Size(No. of sentences)$	$Size\ in\ MB$
$Nepali$	Training	16,278	5.12
	Validation	2,035	0.74
	Testing	2,034	0.74
$English$	Training	16,278	1.89
	Validation	2,035	0.27
	Testing	2,034	0.27

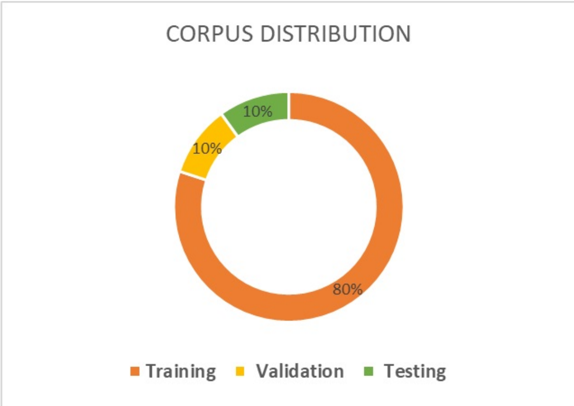


Figure 3: Distribution Of Corpus In Training, Validation, Testing Dataset.

Using the aforementioned corpus, our NMT system is trained to predict translations from English to Nepali and from Nepali to English. We used the default parameter settings of the open-source tool OpenNMT [8] to construct the system. These parameters include a word vector size of 500, 50,000 training epochs, and LSTM RNNs as the default kind of encoder and decoder. Although different combinations of parameter values might produce different outcomes, those combinations have not been investigated in this work. We employed default parameter settings, which are regarded as some of the most widely used ideal values. Our trained NMT system's source, reference, and target are displayed in Table 3.

Table 3: Translations

Nepali to English	
Sentence No. : 1	
Source Sentence (297)	तब फिरऊनले यूसुफलाई भने, "तिमा बाबु अनि दाज्यू-भाइहरू तिमी कहाँ आएका छन्।
Reference Translation	then Pharaoh said to Joseph , " Your father and your brothers have come to you .
Machine Translation	then Pharaoh said to Joseph , " If your father and your brothers have come to you .
Sentence No. : 2	
Source Sentence (401)	परमेश्वरका जनले बताए, "हो, म नै हुँ।"
Reference Translation	the man of god answered , " yes , i am . "
Machine Translation	the man of god said , " yes , i am . "
Sentence No. : 3	
Source Sentence (1106)	ती मानिसहरूले केहेलाताबाट यात्रा गरी पर्वत शेपेरमा छाउनी हाले।
Reference Translation	they left kehelathah and camped at mount shepher .
Machine Translation	they left केहेलाताबाट mountain and camped at शेपेरमा .
English to Nepali	
Sentence No. : 1	
Source Sentence (99)	praise him who alone does wonderful miracles
Reference Translation	परमेश्वरको प्रशंसा गरौं ! जसले एक्लैले नै आश्चर्यकामहरू गर्नुभयो ।
Machine Translation	प्रशंसा गरौं ! जसले एक्लैले नै आश्चर्यकामहरू गर्नुभयो ।
Sentence No. : 2	
Source Sentence (608)	then the people said to Joshua , " No , we will serve the Lord ! "
Reference Translation	तब मानिसहरूले यहोशूलाइ भने , " अँहँ ! हामी परमप्रभुको सेवा गर्नेछौं । "
Machine Translation	तब मानिसहरूले यहोशूलाई भने , " होइन , हामी परमप्रभुको सेवा गर्नेछन् । "
Sentence No. : 3	
Source Sentence (1334)	you must also give male goat as a sin offering , this must be in addition to the daily sacrifice and its grain offerings and drink offerings .
Reference Translation	पापबलिको निम्ति एउटा बोको पनि चढाउनुपर्छ । यो बलि तिमीहरूले चढाउने दैनिकबलि र त्यससित चढाइने अन्नबलि तथा अर्घबलिको अतिरिक्त हो ।
Machine Translation	तिमीहरूले पापबलिको निम्ति एउटा बोको पनि चढाउनु पर्छ । यो बलि तिमीहरूले चढाउने दैनिकबलि र अन्नबलि तथा अर्घबलिको अतिरिक्त हो ।

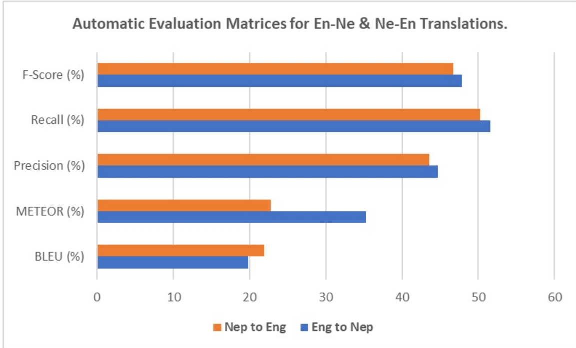


Figure 4: The scores of evaluation matrices.

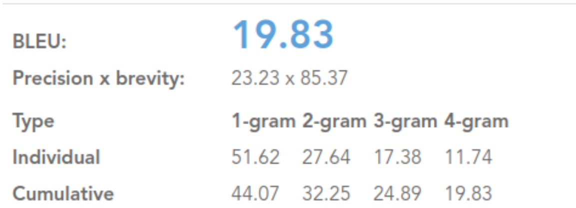


Figure 5: BLEU score of Eng-Nep translation.

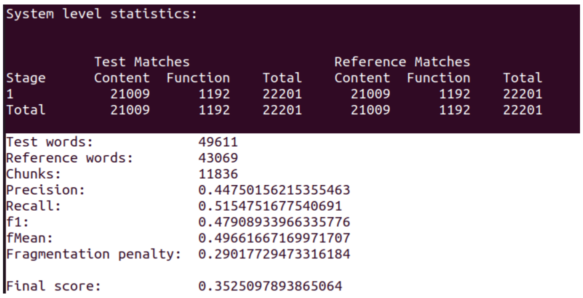


Figure 6: METEOR, Precision, Recall and F-Score of Eng-Nep translation.

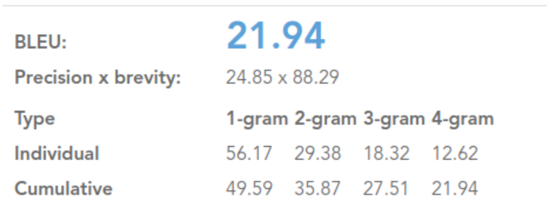
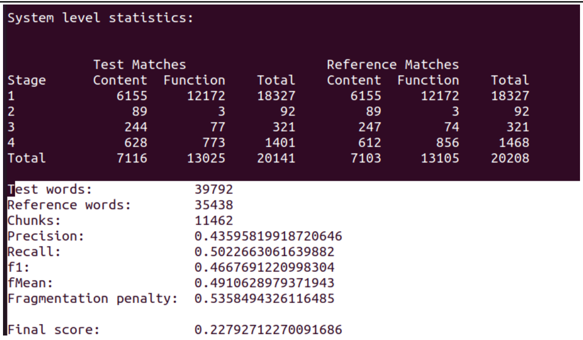


Figure 7: BLEU score of Nep-Eng translation.

The system's experimental results have been evaluated using standardly used automatic evaluation metrics for translations between Nepali and English and between English and Nepali. Table 4 provides a summary of these findings, which are further illustrated in Figures 4, 5, 6, 7, and 8.

Throughout our experiments, we made a few observations. It was discovered that the translation quality is enhanced when the corpus contains more parallel sentences. The machine translations' quality of our system



```
System level statistics:

        Test Matches                    Reference Matches
Stage   Content  Function  Total       Content  Function  Total
1          6155     12172  18327          6155     12172  18327
2            89         3     92            89         3     92
3           244        77    321           247        74    321
4           628       773   1401           612       856   1468
Total      7116     13025  20141          7103     13105  20208

Test words:               39792
Reference words:          35438
Chunks:                   11462
Precision:                0.43595819918720646
Recall:                   0.5022663061639882
f1:                       0.4667691220998304
fMean:                    0.4910628979371943
Fragmentation penalty:    0.5358494326116485

Final score:              0.22792712270091686
```

Figure 8: METEOR, Precision, Recall and F-Score of Nep-Eng translation.

Table 4: Automatic Evaluation Matrices for En-Ne & Ne-En Translations.

$Translation$ $Language$	$BLEU$ (%)	$METEOR$ (%)	$Precision$ (%)	$Recall$ (%)	$F - Score$ (%)
$Eng\ to\ Nep$	19.83	**35.25**	**44.75**	**51.55**	**47.91**
$Nep\ to\ Eng$	**21.94**	22.79	43.59	50.23	46.68

improved as the corpus grew in size. Additionally, incorporating a transliteration system could improve the overall accuracy of our system, as it would convert out-of-vocabulary words that cannot be translated by our system into target language words [14].

5 Conclusion and Future Scope

This paper explores and tests a bidirectional NMT system for the low-resource language pair of English and Nepali. Owing to the fact that not much research has been done on this language pair yet, the ratings derived from automatic assessment measures are ideal for translations in both directions. Yet, analysis of the translated sentences suggests that in order to translate sentences more accurately, the current NMT system needs to be improved. Changes are needed to the traditional NMT approach in order to handle the difficulties posed by low-resource languages. Further studies could concentrate on applying cutting-edge strategies like deep learning and other neural machine translation approaches to improve the system's efficiency and produce translations with higher accuracy.

References

[1] Acharya, P. and Bal, B. K. A comparative study of smt and nmt: Case study of english-nepali language pair. In *SLTU*, pages 90–93, 2018.

[2] Bal, B. K. Structure of nepali grammar. *PAN Localization, Madan Puraskar Pustakalaya, Kathmandu, Nepal*, pages 332–396, 2004.

[3] Banerjee, S. and Lavie, A. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

[4] Bista, S., Keshari, B., Bhatta, J., and Parajuli, K. Dobhase: online english to nepali machine translation system. In *The proceedings of the 26th Annual conference of the Linguistic Society of Nepal*, 2005.

[5] Chaudhary, B. K., Bal, B. K., and Baidar, R. Efforts towards developing a tamang nepali machine translation system. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 281–286, 2020.

[6] Cho, K., Van Merriënboer, B., Bahdanau, D., and Bengio, Y. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014.

[7] Goutte, C. and Gaussier, E. A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In *European conference on information retrieval*, pages 345–359. Springer, 2005.

[8] Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. M. Opennmt: Open-source toolkit for neural machine translation. *arXiv preprint arXiv:1701.02810*, 2017.

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

[10] Koehn, P. and Knowles, R. Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*, 2017.

[11] Laskar, S. R., Pakray, P., and Bandyopadhyay, S. Neural machine translation: Hindi-nepali. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 202–207, 2019.

[12] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. 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, 2002.

[13] Paul, A. and Purkayastha, B. S. English to nepali statistical machine translation system. In *Proceedings of the International Conference on Computing and Communication Systems: I3CS 2016, NEHU, Shillong, India*, pages 423–431. Springer, 2018.

[14] Roy, A. K., Paul, A., and Purkayastha, B. S. Statistical and syllabification based model for nepali machine transliteration. In *International Conference on Computational Intelligence in Communications and Business Analytics*, pages 19–27. Springer, 2022.

[15] Roy, A. K. and Purkayastha, B. S. Machine translation systems for official languages of north-eastern india: A review. In *International Conference on Computational Intelligence in Communications and Business Analytics*, pages 301–315. Springer, 2023.

[16] Shrestha, H. K. Rule based machine translation system in the context of nepali text to english text. *Department of Computer Science, University of Oklahoma*, 2005.

[17] Sutskever, I., Vinyals, O., and Le, Q. V. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.

[18] www.bible.com. RSV Bible: Revised standard version: Youversion. Last accessed 23 October 2023.