Deep Fusion of Bi-LSTM Attention Mechanism for the Enchantment of Machine Translation Performance

BY

Nusrat Jahan Prottasha ID: 171-15-8933

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Dr. Fizar Ahmed
Assistant Professor

Department of CSE

Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY DHAKA, BANGLADESH

DECEMBER 2021

APPROVAL

This Project titled "Deep Fusion of LSTM-Attention Mechanism for the Enchantment of Machine Translation Performance", submitted by Nusrat Jahan Prottasha 171-15-8933 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents.

BOARD OF EXAMINERS

Chairman

Dr. Sheak Rashed Haider Noori

Associate Professor and Associate Head

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University

Internal Examiner

Abdus Sattar

Assistant Professor

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University

Internal Examiner

Saiful Islam

Senior Lecturer

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University

Farid

External Examiner

Dr. Dewan Md. Farid

Professor

Department of Computer Science and Engineering United International University

DECLARATION

We hereby declare that this project has been done by us under the supervision of

Dr. Fizar Ahmed, Assistant Professor, Department of CSE Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

ফিজার আহমেদ

Dr. Fizar Ahmed
Assistant Professor
Department of CSE
Daffodil International University

Submitted by:

Nusrat Jahan Prottasha
Department of CSE

ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to Almighty God for His divine blessing making us possible to complete the final year project/internship successfully.

We are grateful and wish our profound indebtedness to **Dr.Fizar Ahmed, Assistant Professor,** Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "*Data Science, IOT*" to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting them at all stages have made it possible to complete this project.

We would like to express our heartiest gratitude to the Head**,** Department of CSE, for his kind help to finish our project and also to other faculty members and the staff of the CSE department of Daffodil International University.

We would like to thank our entire coursemate in Daffodil International University, who took part in this discussion while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

Language translation is in high demand for a variety of reasons, including business travel and news comprehension, in this modern-day when people are increasingly reliant on technology. Bengali is the world's seventh most widely spoken language; yet, when compared to a language with more resources, such as English, the work done on Bengali machine translation falls short. The accuracy could be increased utilizing a state-of-the-art technique. We were inspired to investigate English to Bengali machine translation after finding research gaps. We utilized Neural Machine Translation namely BiLSTM. The encoder and decoder both employed BiLSTM; the encoder and decoder map the English sentences to Bengali sentences. In the decoder, the attention mechanism was implemented for better mapping. We found that this method works well for accurate machine translation. Because the sequences of the input language of the BiLSTM, are taken from both the left and the right, the mapping becomes more accurate. We compared our approaches to other standard results of English-Bengali translation. And finally, we showed that our result outperformed other claimed results.

TABLE OF CONTENTS

FIGURES **PAGE NO**

LIST OF

LIST OF TABLES

TABLE **PAGE NO**

FIGURES

CHAPTER 1 Introduction

1.1 Introduction

Language is the basic form of communication used by a human being. There are thousands of languages in the world. In order to effectively communicate with the people of the other parts of the world, language translation is a big concern. Bengali, although being the world's seventh most frequently spoken language, has stayed a resource-constrained language, for applying machine translation. Only a few parallel corpora for the Bengali language are publically available at the moment [25], and those, too, suffer from inadequate sentence segmentation and alignments. They also contain a lot of noise, which degrades translation quality [12]. In contrast, English, French, and German are resource-riched languages, providing great success in machine translation.

Machine translation was the first step for the development of Natural Language Processing (NLP) as a distinct field of Artificial Intelligence. The research in NLP got underway in the 1950s, after Booth and Richens' investigations, and also Weaver's seminal memorandum on translation from 1949 [16]. Even though most of the work was done with punched cards and batch processing, the late 1940s to late 1960s were marked by the enthusiasm and optimism of Machine Translation. Rule-based machine translation was the first approach for translation using the artificial agent. However, the accuracy of translation was not up to the mark. Then, the researcher came up with a new approach called statistical machine translation. The MT came into a new era when the neural network was used in MT. After 2014 the neural machine translation is being extensively used by the research community. A sequence-to-sequence (seq2seq) [14] model is most widely utilized in the neural machine translation for machine translation. However, the Bengali language remains a resource-constrained language, although Tahmid Hasan et al. claimed that Bengali was not resource-constrained anymore. However, the accuracy of MT in other language are far higher than in Bengali. We, therefore, firmly believe that there is much work that needs to be done in the near future.

1.2 Motivation

Statistical Machine Translation (SMT) has been popular in the community for the past decade, but as more features are added, the entire pipeline becomes more complex, saturating the translation quality. Because of the limitations of SMT and the success of deep learning, the MT community is focusing on NMT techniques for machine translation.

The advantage of NMT is that it learns an end-to-end mapping from the input to the output in a single large neural network. In order to maximize the performance of the translation output, the model simultaneously learns the parameters [12,14,23], which also requires minimal domain knowledge. Furthermore, unlike Statistical Machine Translation (SMT), NMT does not require the tuning and storage of several models such as the translation language and reordering models.

In the twentieth century, neural networks were investigated for machine translation [25]. Yet, it has lately achieved state-of-the-art performance[11]. Fortunately, With the rapid advancement of Neural Networks and their application in Machine Translation, researchers in the NLP domain started using Neural Networks-based Machine Translation also called neural machine translation (NMT). Deep learning breakthroughs [1,24,26] have greatly helped in the creation of NMT models, obtaining standardized acceptable outcomes in a variety of language pairs. However, in order to train these models properly, a huge number of high-quality sentence pairs must be fed into them [13]; in fact, the lack of such a corpus has a significant impact on their performance.

Cho et al. [4] found that NMT models only require a fraction of the memory required by standard SMT models. NMT has been offering cutting-edge performance for a variety of language pairings since its inception; yet, the literature also acknowledges its limits, such as coping with large sentences [2]. To address the aforementioned difficulties, Attention-based methods have been implemented, in which the model learns to align and translate in tandem. Several attention-based methods have also been proposed by the authors [16], [24], although the transformer design [24] is the most well-known, having self-attention.

With the advancement of the bidirectional encoder, pre-trained model such as BERT plays a vital role in Neural Machine Translation. Haoran Xu et al claimed that state-of-the-art machine translation could be achieved if a pre-trained language model such as dubbed BIBERT, and the output of the contextual embeddings were employed to the input of the Neural Machine Translation model [27]. Furthermore, they introduced the concept of the stochastic layer selection approach, assuring proper utilization of the contextual embedding. They obtained an accuracy of 30.45 (in the case of English to German translation), 38.61 in the case of German to English translation for the IWSLT14 dataset. None of these works have been addressed on English-Bengali machine translation in earlier work performed in Bengali. However, the success of the researchers working solely on Bengali to other Langauge translations or vice versa was not up to the mark, compared to other resource-rich languages. As Bengali is a resource-constrained language lacking English to Bengali translation, and new approaches, we feel the necessity to investigate this field using a stateof-the-art technique.

1.3 Research Challenges

There are several challenges including the standard corpus which can be compared to other established approaches. Parallel corpora are the main obstacle for using a state-of-the-art model such as Neural Machine Translation because these types of methods required a big corpus for training. Due to the lack of a big corpus, a new term OOV, out of vocabulary, arises. Furthermore, training the NMT model is not cost-effective: hardware requirements are the fundamental issue to be addressed here; training time is required several hours to even days. As a result power consumption remain a concerning issue in addition to having a shortage of big corpus. However, the corpus issue can be solved using data augmentation often used in resource-constrained languages such as Bengali. Leveraging monolingual data, and Exploring multi-lingual machine translation for resource-constrained language, semisupervised and unsupervised techniques can be extensively used. Parameters optimization such as learning rate, batch size, number of epochs, are often unstable to small size corpora. The optimization should be handled professionally, utilizing domain knowledge. Increasing dense layer in the seq2seq model is a way to upgrade test accuracy. Designing model architecture is a concern in MT; moreover, picking the best model is also a challenging task to be identified. One to many or many to one sentence alignment can not be properly handled using LASER (language Agnostic Sentence Representations).

1.4 Research Question

To introduce something, a new or upgrade version of an existing one, we need to ask some questions before modeling or implementing it in the real world. For Bengali to English machine translation, we first need to ask these questions, then these questions should be addressed. The flowing is research questions are listed below: What is Human Translation

1. What is Machine Translation?
2. What is the difference between human translation and machine translation
3. Can Machine Translation replace Human-based Translation?
4. What are the challenges of Machine Translation, and how to solve these?
5. What is the obstacle of resource-constrained such as Bengali to other languages machine translation, and how to tackle these issues?
6. What are the approaches used by the research community for Machine translation?

7. What is the state-of-the-art technique for Machine Translation?
8. What are the approaches taken by the researchers for Bengali-English or EnglishBengali translation?
9. What is the dataset for English-Bengali Machine Translation?
10. How to preprocess Bengali-English text data?
11. What will be the BiLSTM model architecture?
12. How to train attention-based BiLSTM?
13. How to perform BiLSTM optimization?
14. Is this model is implementable in real-world problem-solving?
15. What is the Future of Machine Translation?

1.5: Research Methodology

Dataset for the MT is collected from [10]. We preprocessed the dataset using the Bengali language Toolkit. We explored different models to choose the best one. After doing several experiments such as SMT, and NMT, we chose Bi-directional Long Short-term Memory, in short BiLSTM, an NMT-based model with an attention mechanism, as our proposed method. We will discuss the methodology in great detail in section 3.

1.6: Research Objective

Finding the research gap, we were motivated to explore Bengali-English machine translation. There are many reasons to do this research works, however, in the flowing, some main points are listed.

- Develop a system that can effectively translate from English to the Bengali language
- Utilize existing corpus to develop a system that can solve real-world problems
- Improve existing methods to increase translation accuracy
- Use attention mechanism to improve accuracy

1.7 Research contribution

Our contribution includes: introducing BiLSTM with attention proposed by DzmitryBahdanau et al.[2]. We used the dataset introduced by Tahmid Hasan et al. [10] and obtained an accuracy score that outperformed the baseline. The model we used had an attention mechanism that was not addressed in the Bengali language earlier. Furthermore, the giant work such as machine translation of Google and Microsoft is not good enough for the Bengali language compared to other resource-rich languages.

1.8 Report Organization

The report was prepared according to the flowing structure, section 1 introduction, section 2 literature review; methodology section was discussed in section 3. We kept place results and discussion in section 4. In section 5, we mentioned the conclusion and future scope. And finally, the references were added at the end of this report i.e. section 6.

Chapter 2 Literature Review 2.1 Statistical Machine Translation (SMT)

Although machine translation was being started in the 1940s, English to Bengali machine translation was unexplored until 1991 when Naskar et al. used Indian languages including EnglishBengali language pair for machine translation[19]. Flowing this study, rule-based approaches were explored, taking into consideration whether sentence structure during model training [20,21] or exploring grammar [6,7]. Then Statistical Machine Translation (SMT) such as word and phrase-based translation approaches were investigated. However, for English-Bengali machine translation, phrase-based translating approaches were widely used including Banerjee et al.'s study [3]. Another notable work worth mentioning here is that Mumin et al. claimed a BLEU score of 17.43, utilizing a log-linear phrase-based SMT approach [17].

2.2 Neural Machine Translation (NMT)

With the development of the neural network, Neural Machine Translation, a-state-fo-the-art approach, was emerged. However, For Bengali, a low-resourced language, there have been investigated very few works. Among them, Liu et al. used an NMT approach named LSTM with an attention mechanism [15] and claimed a BLEU score of 10.92. Dandapat and Lewis developed a general NMT-based model for Bengali-English corpus[5]. They compensate for the inadequacy of training data, utilizing data augmentation[9,22]. Another limited training data was introduced by Mumin et al in 2019, yet for the Bengali-English pair, they claimed improved translation accuracy[18]. Besides Hasan et al. introduced the NMT-based model [9], securing a BLEU score of BERT, a pretrained transformer model, can be used for machine translation. Sentence relationships can be extracted using BERT. Furthermore, this pre-trained model can be used to predict the next words which are very useful for machine translation. Utilizing the pre-trained model, ZhiyuGuo and Minh Le Nguyen introduced document-level machine translation which outperformed the baseline BLEU score [8]. Zhebin Zhang proposed BERT-JAM in which they represent the BERT model in such a way that can mostly be utilized for neural machine translation [28]. They introduced fusion modules in the BERT-JAM.

2.3 Research Scope

However, training transformer models such as BERT required much time compared to Recurrent Neural networks. Raúl Vázquez et al. claimed, by reviewing that BERT works better for other language tasks, however, for translation tasks, an NMT-based encoder works well. By motivating their work, and finding their result we decided to investigate BiLSTM with attention to machine translation.

Chapter 3

Research Methodology

The encoder-decoder approach is a state-of-the-art technique is in the Neural Machine Translation along with the Transformer model such as BERT. We chose Bi-directional Long Short-term Memory, or BiLSTM, an NMT-based model with an attention mechanism, as our desired method after conducting numerous trials such as SMT and NMT. See Figure 1.

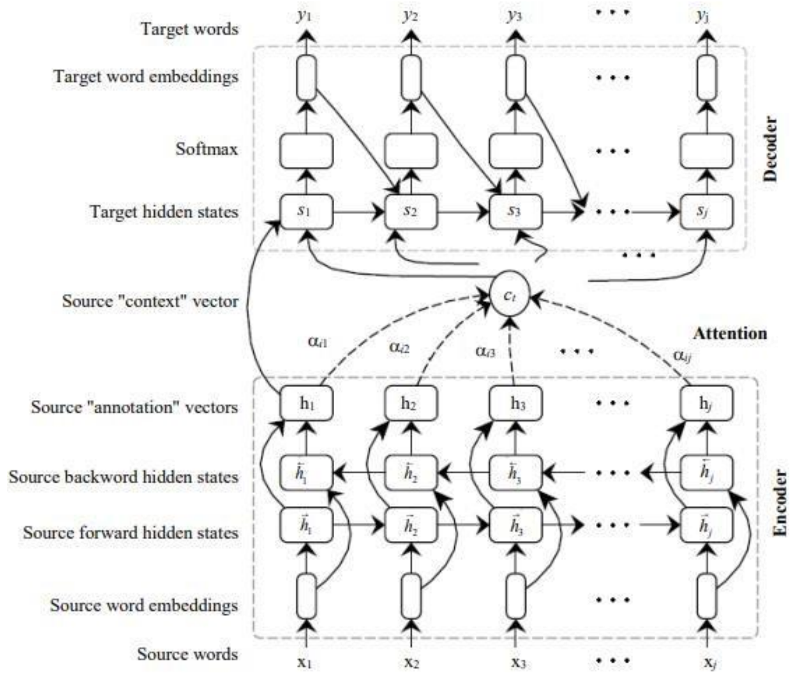

Target words
Target word embeddings
Softmax
Target hidden states
Source "context" vector
Source "annotation" vectors
Source backword hidden states
Source forward hidden states
Source word embeddings
Source words
Decoder
Encoder
Attention
y_1 y_2 y_3 y_j
s_1 s_2 s_3 s_j
c_t
α_{i1} α_{i2} α_{i3} α_{ij}
h_1 h_2 h_3 h_j
$\overleftarrow{h}_1$ $\overleftarrow{h}_2$ $\overleftarrow{h}_3$ $\overleftarrow{h}_j$
$\overrightarrow{h}_1$ $\overrightarrow{h}_2$ $\overrightarrow{h}_3$ $\overrightarrow{h}_j$
x_1 x_2 x_3 x_j

Figure 1: BiLSTM model architecture is used in the proposed approach

3.1 Dataset

We used the dataset introduced by Tahmid Hasan et al. [10]. They collected approximately 2.75 million sentence pairs, comprising of different domains. Their proposed dataset contains different sources, and they claimed that 2 million sentences out of 2.75 were new, not introduced anywhere else. Though Bengali is a resource-constrained language, introducing this corpus, Tahmid Hasan et al. [10] claimed Bengali is not a low-recourse anymore. Their proposed dataset comprising of 2,751,315 sentence pairs; the Bengali number of tokens is 35,327,967, however, the English number of tokens is slightly more that is 40,739,723

3.2 Data Preprocessing

Data preprocessing is an important task in any NLP problem, reducing computational power, and time. We performed standard preprocessing the same as other languages. Some of these tasks include data cleaning. Other tasks of data cleaning include upper case conversion to lowercase, and non-alphabetic characters are removed from words. Secondly, As humans communicate with others using different languages at the same time, we discarded foreign strings that appeared on both language pairs. Before training the model, we performed tokenization and normalization to characters and punctuations for reducing data sparsity. Neural Machine Translation White space is used to tokenize text. Two kinds of tokenization are extensively used: one is word-level tokenization, and the other one is character level. The size of the vocabulary of word-level tokenization is much larger than the size of character-level tokenization. We performed character-level tokenization. This is because we wanted to handle the out of vocabulary (OOV) words, for increasing the translation accuracy. To tackle the OOVs words we treated infrequent words as the unknown token "UNK".

3.3 Neural Machine Translation

Neural machine translation, as the name suggests, employed a neural network to conduct machine translation. Currently, this model is extensively being used. Compared to previous approaches. This is because neural translation mimics human translations in the sense that the neural network used in NMT translation is theoretically equivalent to a human neuron system, producing the same kind of functionality. We used Bidirectional Long Short-term Memory, a Recurrent Neural Network, as our NMT model. The architecture comprises an encoder and a decoder. The translation is performed, mapping the input sentence such as Bengali to a target sentence such as English with the help of the Encoder, and the decoder.

Bengali sentences are mapped to English sentences using this mapping. In the Encoder BiLSTM is used, and an attention mechanism is employed in the decoder, generating accurate output or target words. See Figure 2.

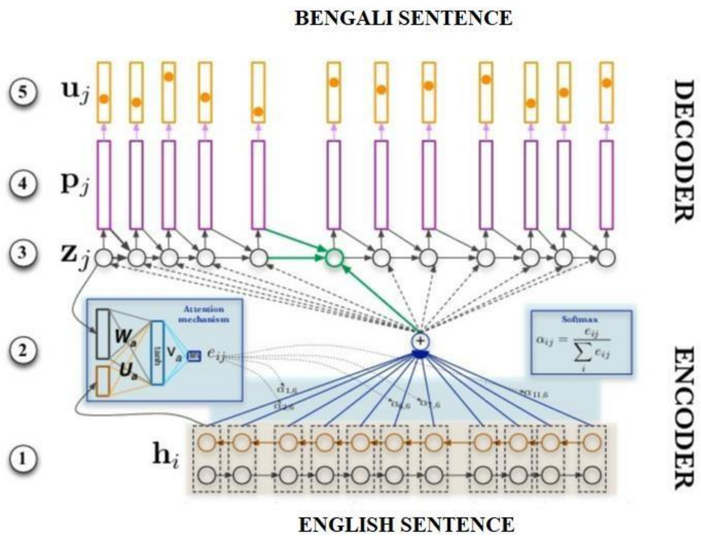


Figure 2: English to Bengali machine translation using BiLSTM with attention

3.3.1 RNN Encoder and Decoder

Encoder and Decoder are the fundamental elements of an NMT architecture. The RNN EncoderDecoder introduced by Cho et al. and Sutskever et al is widely used by the researchers. A neural network-based machine translation (NMT) system maps an input sentence or a source sentence to a target sentence. Concretely, a set of input sentences are directly modeled to a set of output sentences.

Several researchers used different model architecture for mapping. Kalchbrenner and Blunsom employed a convolutional neural network to encode the source sentence representation and an RNN with the standard hidden unit as the decoder[12]. Sutskever et al. [23] and Luong et al.[16], on the other hand, stacked numerous layers of an RNN with an LSTM hidden unit for both the encoder and the decoder. Cho et al. [4], Bahdanau et al. [2], and Jean et al.[11] used a different form of the RNN for both components, the gated recurrent unit (GRU), which was influenced by LSTM.

Now, let us consider $\mathbf{x} = (x_1, \cdots, x_{T_x})$ be the input sequenced vector. This sequence of vectors is converted to context vector c. The RNN hidden state can be written in the following form.

$$h_t = f(x_t, h_{t-1})$$

Here $h_t \in \mathbb{R}^n$ is the hidden state of a particular time t and f is a nonlinear function.

Combining all the hidden states, the context vector is generated using the nonlinear function, q.

$$c = q(\{h_1, \cdots, h_{T_x}\})$$

The next word, to be predicted, required a context vector and all the previous words predicted earlier time.

$$p(\mathbf{y}) = \prod_{t=1}^{T} p(y_t \mid \{y_1, \cdots, y_{t-1}\}, c)$$

where

$$y = y_1, \cdots, y_{T_N}$$

The probability of next words can be computed as a conditional probability which is modeled as

$$p(y_t \mid \{y_1, \cdots, y_{t-1}\}, c) = g(y_{t-1}, s_t, c)$$

Where g is a function of non-linear type, computing the output probability of y_t

3.3.2 Attention in Decoder

The basic components in a fundamental NMT architecture are an encoder which is responsible for computing a representation for every source sentence, and a decoder that maps the source sentence to a target word at a specific time.

We used the attention mechanism to the decoder part. In detail on how the decoder work, we may consider the conditional probability proposed by this paper [1].

$$p(y_i \mid y_1, \dots, y_{i-1}, \mathbf{x}) = g(y_{i-1}, s_i, c_i)$$

Here,$s_i$refers to the hidden state at the time i. The hidden state $s_i$can be computed by using the flowing equation.

$$s_i = f(s_{i-1}, y_{i-1}, c_i)$$

The difference between this approach and the traditional approach where a fixed context vector is being created. However, in this approach, a distinct context vector c_i is computed for every target word y_i.

The context vector has a dependency on the whole hidden state annotation and can be computed as a weighted sum of hidden state annotation h_i.

$$c_i = \sum_{j=1}^{Tx} \alpha_{ij} h_j$$

The weight of the hidden state can be computed using the flowing equation:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})}$$

Where $e_{ij} = a(s_{i-1}, h_j)$ is a scoring model that determines how much matching is achieved around the input at position j and the output at position $i$match. The importance of annotation at the hidden state h_i is directly affected by the probability of α_{ij} or associated energy e_{ij}in deciding or generating the next word. In this way, the attention mechanism is being implemented in the decoder. The decoder pays attention to some parts of the source sentence. As a result, it is not necessary to encode all the source words, reducing the fixed-length vector. Furthermore, this approach, to pay attention, spreads throughout the sequence of annotations that can be retrieved by the decoder. See Figure 4.

3.3.3 Long Short-term Memory (LSTM)

Long Short-term Memory is a recurrent neural network. The LSTM, a recurrent neural network version, is a specific type of recurrent neural network for dealing with exploding gradient difficulties in the base RNN. Unlike the Feedforward network, in the hidden layer of a recurrent neural network, however, there is a cycle. The forward and backward network can be a vanilla RNN unit, a GRU, or an LSTM unit that calculates the current hidden state given the prior hidden state. We used BiLSTM for these encoding purposes. The previous timestep's hidden layer serves as a memory element or context that can store previously processed data for future decision-making.

The state unit $S_i^{(t)}$, which possesses a linear self-loop, is the most significant component. The self-loop weight (or associated time constant) is regulated here by a forget gate unit $f_i^{(t)}$ (for time step t and cell i) which uses a sigmoid unit to adjust this weight to a value between 0 and 1. See Figure 3.

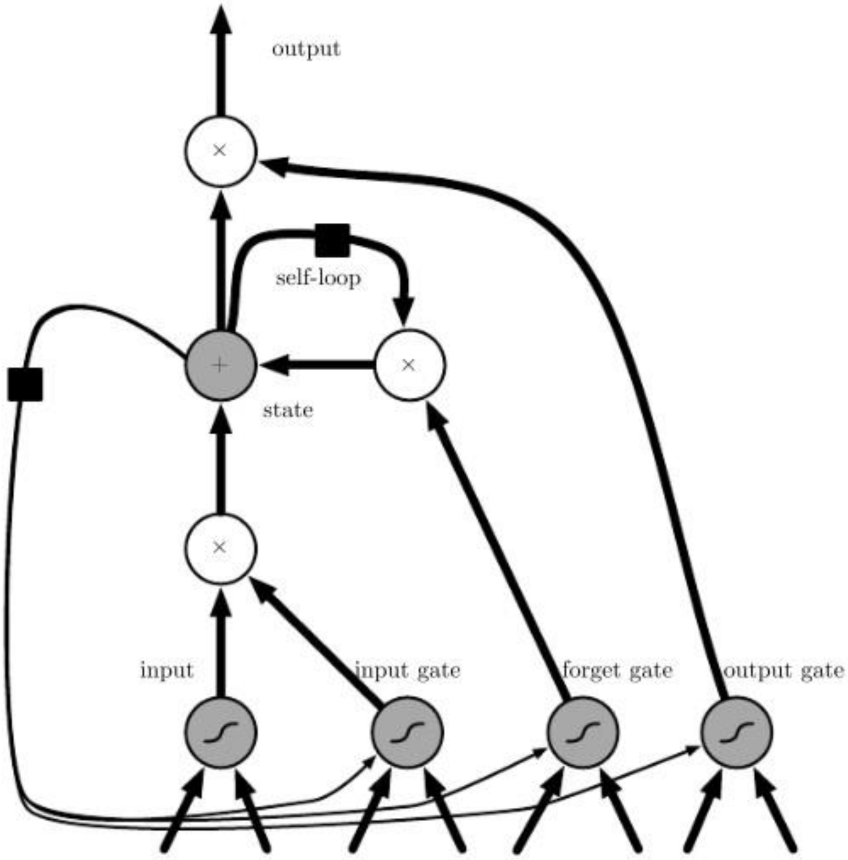


Figure 3: Block diagram of the LSTM recurrent network.

$$f_i^{(t)} = \sigma\left(b_i^f + \sum_j U_{i,j}^f x_j^{(t)} + \sum_j W_{i,j}^f h_j^{(t-1)}\right)$$

where $x^{(t)}$is the current input vector, $h^{(t)}$is the current hidden layer vector, including the outputs of all the LSTM cells, and $b^f, U^f, W^f$are the forget gates' corresponding biases, input weights, and recurrent weights.

The internal state of the LSTM cell is thus updated, albeit with a conditional self-loop weight $f_i^{(t)}$.

$$s_i^{(t)} = f_i^{(t)} s_i^{(t-1)} + g_i^{(t)} \sigma\left(b_i + \sum_j U_{i,j} x_j^{(t)} + \sum_j W_{i,j} h_j^{(t-1)}\right)$$

Where b,U, and W ande the biases, input weights, and recurrent weights into the LSTM cell, respectively. The external input gate unit $g_i^{(t)}$ is calculated in the same way as the forget gate (using a sigmoid unit to generate a gating value between 0 and 1), but with the following parameters:

$$g_i^{(t)} = \sigma\left(b_i^g + \sum_j U_{i,j}^g x_j^{(t)} + \sum_j W_{i,j}^g h_j^{(t-1)}\right)$$

The LSTM cell's output $h_i^{(t)}$ can also be turned off using the output gate $q_i^{(t)}$ Gating is also done with a sigmoid unit:

$$h_i^{(t)} = \tanh\left(s_i^{(t)}\right) q_i^{(t)}$$

$$q_i^{(t)} = \sigma\left(b_i^o + \sum_j U_{i,j}^o x_j^{(t)} + \sum_j W_{i,j}^o h_j^{(t-1)}\right)$$

It has the following settings for its biases, input weights, and recurrent weights: $b^0, U^0 \quad W^0$. As shown in Fig. 2., one of the options is to employ the cell state $s_i^{(t)}$ as an extra input (with its weight) into the three gates of the ith unit.

3.3.4 Sequence Annotation in Bi-directional LSTM

BiLSTM refers to a Bi-directional LSTM consisting of forward and backward LSTM. The forward LSTM reads the input sequences from left to right, denoting $\vec{h}_j^\top$, in contrast, the backward LSTM reads the input in reverse order, denoting $\overleftarrow{h}_j^\top$, making an annotation pair for input sequences.

$$h_j = \left[\vec{h}_j^\top ; \overleftarrow{h}_j^\top\right]^\top$$

As a result, the annotation h_j contains information both the previous and the next words (see figure 4.)

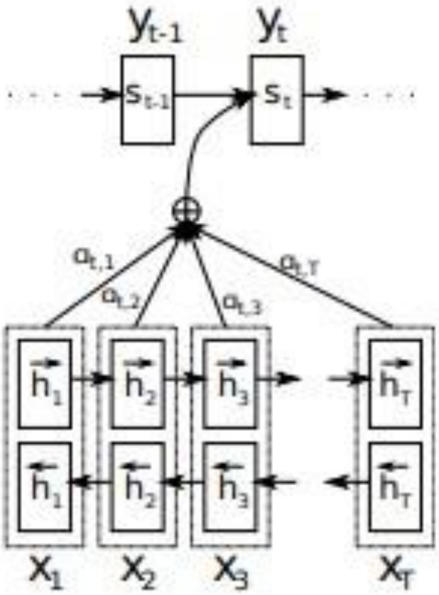


Figure 4: The target word $y_t$is generated given the input words sequences $\mathbf{x} = (x_1, \cdots, x_{Tx})$ and the previously hidden layer's, s_{t-1},information.

CHAPTER 4 Result and Discussion

4.1 Experimental Set up

The entire project was completed using the Python programming language, having distribution's edition 3.7., and trained on Google Colaboratory. There are several opensource tools built based on Python such as Keras, NLTK, NumPy, Windows 10, Python 3.8, Sklearn, Pandas, etc were used in our project for better model building. Furthermore, a deep learning framework, TensorFlow were used to build BiLSTM Recurrent neural Network. Pandas are used for data manipulation, and visualization, Matplotlib was used. Machine learning tools, Scikit-learn, Keras were used along with TensorFlow. Keras also combines the benefits of Theano and Tensor Flow to create a neural network model. The configuration we used is Local PC Google Colab with 35GB TPU, i7 16GB RAM, 1 TB HDD

4.2 Result

In this section, we presented the result we have gotten in our experiment during training and testing time. For the MT, Different automatic evaluation metrics such as BLEU. We, therefore, addressed them in our result section. After collecting data, we preprocessed the dataset.

Training the model is an important step, as there are many parameters involved in the model architecture. Without effective optimization, there may have much fluctuation of the test results, becoming difficult to choose the best model. Model parameters include a learning rate of 0.01 We used the ReLu activation function in all the hidden layers, softmax is used in the output layer. After training up to 30 epochs, we have obtained an accuracy score of 96.49. We have also shown the cross-entropy along with accuracy in Figure 5.
concerning the number of epochs.

Bilingual Evaluation Understudy in short BLEU is a metric for evaluating the MT models. This metric makes a correlation between human translation and machine translation. The idea behind this metric is that the more the score the better the MT. This is one of the most popular metrics used in the MT evaluation metrics. The closer the MT to the human translation the better the translation quality. We have obtained a test accuracy on the test set is 28.79.

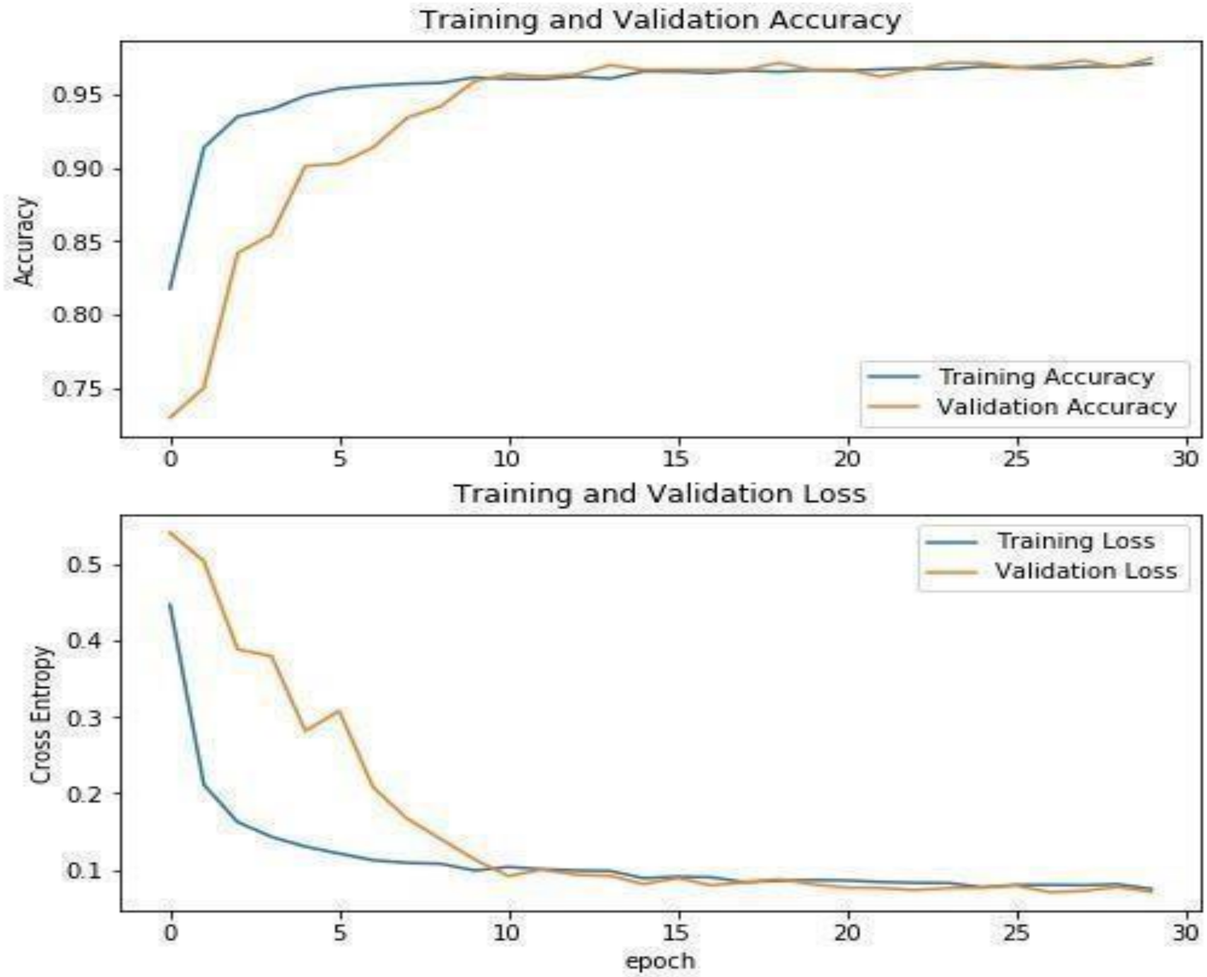


Figure 5: The accuracy of training and testing is shown in the upper part of this figure. And the cross-entropy is shown in the bottom part. All the concerning the number of epochs.

4.3 Result Comparision

There are several approaches currently available, however, neural machine translation outperformed other traditional techniques. We, therefore mainly compared our approach to other NMT-based techniques. From Table 1., we see that Mumin et al proposed two approaches which are SMT and NMT based, securing results of 17.43 and 22.68 [17]. Yet, the NMT based outperformed the SMT. Furthermore, Hasan et al. also proposed an NMTbased approach which is also outperformed the SMT [9]. In addition to this, Tahmid Hasan et al. used a transformer-based model in their experiment, and show that their approach outperformed the earlier mentioned results, claiming a BLEU score of 22.02 for En→Bn [10]. We used a BiLSTM encoder and decoder, having an attention mechanism in the decoder. We obtained a BLEU score of 23.05 for En→Bn. We, however, do not perform Bn→En translation.

In order to implement the proposed approach in a real-world scenario, we showed that our approach can be used interchangeably with other automatic translators for English to Bengali translation. See Table 2. The result we obtained, outperformed other giant works such as Google, Bing, etc. Then we used the RisingNews test dataset introduce by [10], this dataset contains 200 articles, having 600 validation and 1000 test pairs. From Table 3. it is apparent that the accuracy we obtained outperformed the baseline

Table 1. The BLEU score of SMT and NMT-based model are shown

Method	**En→Bn**
SMT [17]	15.27
NMT [18]	16.26
OpenNMT [10]	22.02
Our approach	23.05

Table 2. The comparison is shown, having 23.92 SacreBLEU scores of our approach with the automatic translator such as Google, Bing.

Translator	**SuPara (En→Bn)**
Google	11.1
Bing	10.7
OpenNMT [10]	22.0
Proposed	23.92

Table 3. The BLEU score of the RisingNews test set is shown with other approaches.

Metric	**En→Bn**
BLEU	27.73
SacreBLEU	27.7
BLEU (Proposed)	28.75

4.4 Discussion

Language translation especially automatic translation is not an easy task to be performed using a machine that has no intelligence. However, with the help of Artificial Intelligence, and the dominant advancement in the neural network, the artificial agent's intelligence competes with the human being. Neural Machine Translation, having such intelligence can be used in real-world problems like language translation. The NMT approach we proposed, outperformed all the mentioned results, indicating that we may implement this system to translate from English to Bengali. However, We do not perform translation from Bengali to

English.

Chapter 5 Conclusion and Future Work

Automatic language translation is a crying need demand for various reasons such as business travel, news understanding, in this modern age when people mostly depend on technology. Bengali is the seventh most widely spoken language in the world, however, the works done on Bengali machine translation are not up to the mark compared to the resources-rich language such as English. Finding this research gap we were motivated to explore Bengali to English Machine Translation. We used BiLSTM in the encoder and decoder. An attention mechanism was also introduced in the decoder. We found this approach is effective for accurate machine translation. As in the BiLSTM, the sentences are taken from the left as well as from the right, the mapping becomes more accurate. Furthermore, all the information in the input text is not necessary, some texts are more crucial in which we need to pay attention to more. That’s why we introduce an attention mechanism in the decoder.
Yet there are some challenges including the corpus size as Bengali is a resource-constrained language.

In the future, we want to introduce a new dataset containing millions of sentence pairs to make Bengali a resource-rich language. Furthermore, we also investigate transformer models such as BERT for NMT.

Reference:

[1] Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.* (September 2014). Retrieved November 14, 2021, from https://arxiv.org/abs/1409.0473v7

[2] Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.* (September 2014). Retrieved November 15, 2021, from https://arxiv.org/abs/1409.0473v7

[3] Tamali Banerjee, Anoop Kunchukuttan, and Pushpak Bhattacharya. 2018. Multilingual Indian Language Translation System at WAT 2018: Many-to-one Phrase-based SMT. (2018). Retrieved November 24, 2021, from https://github.com/anoopkunchukuttan/

[4] Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. (September 2014), 103–111. DOI:https://doi.org/10.3115/v1/w14-4012

[5] Sandipan Dandapat and W. Lewis. 2018. Training deployable general domain MT for a low resource language pair: English–Bangla. *undefined* (2018).

[6] Sajib Dasgupta, Abu Wasif, and Sharmin Azam Cse. An optimal way of machine translation from English to Bengali. Retrieved November 24, 2021, from https://www.researchgate.net/publication/228734022

[7] Judith Francisca and Mamun Mia. Adapting rule based machine translation from english to bangla.

[8] Zhiyu Guo and Minh Le Nguyen. 2020. Document-Level Neural Machine Translation Using BERT as Context Encoder. 101–107. Retrieved November 5, 2021, from https://aclanthology.org/2020.aacl-srw.15

[9] Md Arid Hasan, Firoj Alam, Shammur Absar Chowdhury, and Naira Khan. 2019. Neural machine translation for the bangla-english language pair. *2019 22nd Int. Conf. Comput. Inf. Technol. ICCIT 2019* (December 2019). DOI:https://doi.org/10.1109/ICCIT48885.2019.9038381

[10] Tahmid Hasan, Abhik Bhattacharjee, Kazi Samin, Masum Hasan, Madhusudan Basak, M Sohel Rahman, and Rifat Shahriyar. Not Low-Resource Anymore: Aligner Ensembling, Batch Filtering, and New Datasets for Bengali-English Machine Translation. Retrieved November 5, 2021, from https://github.com/facebookresearch/

[11] Sébastien Jean, Orhan Firat, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2015. Montreal Neural Machine Translation Systems for WMT'15. (December 2015), 134–140. DOI:https://doi.org/10.18653/V1/W15-3014

[12] Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent Continuous Translation Models. 1700–1709. Retrieved November 15, 2021, from https://aclanthology.org/D13-1176

[13] Philipp Koehn and Rebecca Knowles. 2017. Six Challenges for Neural Machine Translation. (July 2017), 28–39. DOI:https://doi.org/10.18653/V1/W17-3204

[14] Oleksii Kuchaiev, Boris Ginsburg, Igor Gitman, Vitaly Lavrukhin, Jason Li, Huyen Nguyen, Carl Case, and Paulius Micikevicius. 2018. Mixed-Precision Training for NLP and Speech Recognition with OpenSeq2Seq. (May 2018). Retrieved November 15, 2021, from https://arxiv.org/abs/1805.10387v2

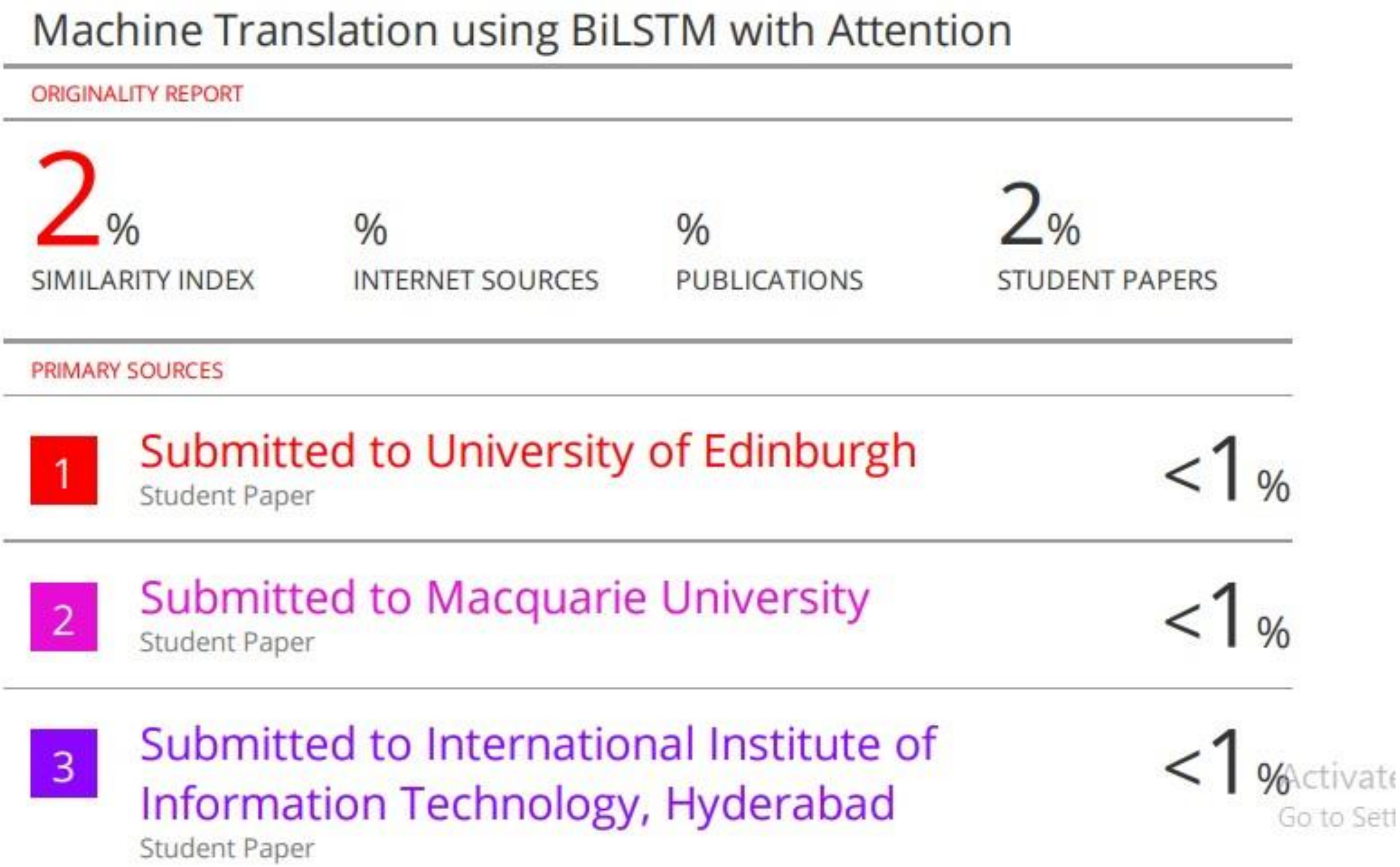
[15] Nelson F Liu, Jonathan May, Michael Pust, Kevin Knight, and Paul G Allen. 2018. Augmenting Statistical Machine Translation with Subword Translation of Out-ofVocabulary Words. (August 2018). Retrieved November 24, 2021, from https://arxiv.org/abs/1808.05700v1

[16] Minh Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. *Conf. Proc. - EMNLP 2015 Conf. Empir. Methods Nat. Lang. Process.* (August 2015), 1412–1421. DOI:https://doi.org/10.18653/v1/d15-1166

[17] M A A Mumin, Md Hanif Seddiqui, Muhammed Zafar Iqbal, and Mohammed Jahirul Islam. 2019. shutorjoma: An english↔ bangla statistical machine translation system. *J. Comput. Sci. (Science Publ.* (2019).

[18] Mohammad Abdullah Al Mumin, Md Hanif Seddiqui, Muhammed Zafar Iqbal, Mohammed Jahirul Islam, Mohammad Abdullah, and Al Mumin. 2019. Neural Machine

Translation for Low-resource English-Bangla. *J. Comput. Sci.* 15, 11 (November 2019), 1627–1637. DOI:https://doi.org/10.3844/JCSSP.2019.1627.1637



Machine Translation using BiLSTM with Attention

ORIGINALITY REPORT

2% SIMILARITY INDEX

% INTERNET SOURCES

% PUBLICATIONS

2% STUDENT PAPERS

PRIMARY SOURCES

1 Submitted to University of Edinburgh
Student Paper
<1%

2 Submitted to Macquarie University
Student Paper
<1%

3 Submitted to International Institute of Information Technology, Hyderabad
Student Paper
<1%

[19] Sudip Naskar and Sivaji Bandyopadhyay. Use of Machine Translation in India: Current Status. Retrieved November 24, 2021, from http://www.uohyd.ernet.in/

[20] Goutam Kumar Saha. 2005. The E2B machine translation. *Ubiquity* 2005, August (August 2005), 1–1. DOI:https://doi.org/10.1145/1088431.1088432

Thank You!

Best of Luck