

English to Bengali Neural Machine Translation System for the Aviation Domain

SAPTARSHI PAUL¹, BIPUL SHYAM PURKHYASTHA²

^{1, 2} Department of Computer Science, Assam University, Silchar, Assam, India

E-mail: paulsaptarshi@yahoo.co.in, bipul_sh@hotmail.com

Corresponding Author: paulsaptarshi@yahoo.co.in

Abstract: Machine translation systems for Indian languages such as Bengali and others are commonly found. Classical machine translation systems involving Bengali are available for tourism, agriculture, medical and other domains. The performances of these systems are restrained by the linguistic knowledge, that are used to develop the rules. In the recent past, notable results have been achieved by systems using neural machine translation. Well known organizations like Google and Microsoft have started using NMT models. In this paper, we explore the design and implementation of an unexplored domain in Bengali, the aviation domain. It is implemented using a neural machine translation model. In order to implement it, we have used English to Bengali parallel corpus for the aviation domain which was developed specifically for this implementation. The corpus is a unique one with large number of aviation specific OOV words and phraseologies included in it. We have used the already developed aviation preprocessing tool, E-dictionary and transliteration tool for creation of the corpus and system. Ultimately we get the output model which generates our machine translated output file in Bengali. We then apply the aviation phraseology converter and transliteration tool on the output to get a post-processed output. The two versions of the output are compared using n-gram BLEU score. The results ultimately demonstrate that NMT output with the post processing exhibits better results.

Keywords: Encoder; Decoder; Aviation; BLEU

(Received October 21st, 2020 / Accepted November 15th, 2020)

1. INTRODUCTION

Social characteristics of humans make him or her communicate with others for a variety of reasons. This may include expressing his feelings, ideas and needs. Communication is the most vital activity of humans, may it be personal or professional domain. As a result, several languages have been created by humans to communicate. Natural languages or simply languages are composed of sentences which in turn are formed of words and their guiding rules. Natural languages have found itself of interest among research communities. Researchers have been interested to find solutions in creating links among separate natural languages, so began the studies in translation among natural languages. To perform translation among different natural languages human translators and

natural language interpreters masters two or more natural languages to formulate rules and regulations between the source and target language.

Work done by human translators is time consuming and a complex task. The translator needs to have the basic understanding of both the languages. Though it is expensive and time consuming still human translation has the best translation quality. To reduce the drawbacks of human translators and to assist in human assisted translation, researchers are continuously interested in coming up with efficient machine translation systems. Machine translation or MT systems are proving to be very useful tools as they make the work of communication between humans speaking different natural languages much easier. MT systems have found their importance increased in several versatile domains, such as agriculture, tourism, in

professional use, to explore written documents in different natural languages etc. Modern MT systems are multi-lingual, user-friendly and in most case free of cost. The

comprehend translation context and the semantics of the written data. The context and content of the data to be translated is an essential aspect. To get best results, the most efficient translation system used is computer assisted one. Computer assisted translation system allows to reap the benefits of machine translation, and at the same time permits the user to perform post-processing on the results of the MT system. The obtained results can be corrected and the imperfections removed. Examples include substituting out-of-vocabulary words, phraseologies, proper nouns etc. Classical MT systems generally are restricted by the domains that they are trained with and the rules that define them. Modern SMT and NMT systems are domain dependent but their models “LEARN” from the parallel corpus. Communication through MT systems is mainly domain specific and thus the race to include more domains is on. The more is the training data from the domains, the better is the MT model translation analysis results. Technical domains such as aviation, aero-space and others are specialized domains which are made up of huge number of OOV words, phraseologies and structured English sentences. These specialized domains remain largely untranslated for Indian languages. In fact finding enough data to build data repository to train MT systems for such domains is a huge challenge. In this paper, we will discuss the design and development of such a NMT system that can handle sentences related to aviation domain for English to Bengali. Till the recent past, SMT models have been frequently used. SMT models have depended on the use of large parallel corpus. Translation units of SMT were divided and were assigned a probability score. SMT systems depends on the analyses of a sentence to translate, SMT segments the sentence written in source language into separate units. These individual units are then compared to the corpus, producing the translation with the largest probability score. The new dominant approach, Neural Machine Translation, popularly known as NMT was proposed by Kalchbrenner, N., Blunsom[1], Sutskever, I., Vinyals[2] and Cho, K., Merriënboer [3]. Recently, Google [4] and Systran [5] have adopted NMT and their translation engines are powered by NMT. NMT advocates

biggest advantage lies in the choice of multiple languages and domains. The disadvantage and main challenge to MT systems are their inability to

the use of neural networks for translation of a text written in one language to the target natural language. Compared to SMT, NMT finds its strength and flexibility in its ability to learn from previous examples, much like the way human brain works. Like SMT, NMT also requires a parallel corpus. But NMT considers the whole sentence to be translated and creates a “thought vector” that is created by adding “weights” to each constituent word of the sentence. The words in themselves are embedded. A study of NMT made by Mistry [6] states that in this way the whole sentence is a unit as a whole. NMT translate source sentences into target language ones by using the encoder-decoder model. For the encoder part, input sentences are encoded to fixed-length vector representations with embedding methods that are machine readable. In the decoder part a recurrent neural network RNN (LSTM in our case) model predicts sequence in target language word by word that corresponds to the input vector. Training is done through “Teacher Forcing” method after calculating the loss. The decoder then generates the sequence in the translated form in the target language. This paper has been organized as follows: Section 2 reviews the related existing systems on NMT and NLP tools in aviation. Section 3 discusses the RNN and LSTM models. Section 4 deals with word representation with an example related to our work. Section 5 deals in details the implementation of the aviation NMT system. Sections 6, the results obtained are discussed in details and finally, Section 7 discusses the conclusion. The Contribution of the research work discussed in the paper is as follows:

1. Creation of a pre-processing tool for developing the parallel Aviation corpus
2. Creation of Aviation OOV words post-processing tool to increase Translation Accuracy (TA).
3. Creation of Aviation phraseology converter post-processing tool to increase TA.
4. Creation of Aviation E-Dictionary to assist in creation of Aviation Corpus.
5. Creation of the unique Aviation Corpus
6. Getting an output of Open_NMT as 39.98 for Aviation Domain (first known work).
7. Application of the two developed post processing tools to increase TA to 40.58
8. Achieving an increase in TA of (+ .60%).

2. RELATED WORK

In the recent past, multiple research endeavors and work in the field of developing NMT systems have been made. For Indian languages English to the Hindi NMT system has been developed by Mishra[7]. Here use of feed-forward back-propagation neural network was explored by Shahnawaz and Mishra [7], and found effective. The developed system has been evaluated with three different methods. n-gram BLEU with a score of 0.604, METEOR 0.830, and lastly F-score with score of 0.816. A NMT system for translating English sentences into Arabic by Mishra and Marwan [8], has been successful. Here, it was assumed that the sequences are well structured. In this system, the two methods were mixed; firstly the rule based technique and secondly the feed-forward back propagation neural network. The system was evaluated by BLEU, METEOR and F-measure methods. The score achieved by n-gram BLEU was 0.6029, 0.8221 on METEOR, and 0.8386 on F-measure. A NMT system for translating standard language varieties of the same language was done by Costa-jussa et al [9]. The NMT model was implemented using a RNN and trained on Brazilian to Portuguese corpus. Evaluation results include BLEU score of 0.9. NLP systems and tools used in aviation include TUAM AVIATION [10] [11] for translation from French to English, NLP tools such as BLUE used by BOEING[11] [12] and AMRIT used by AIRBUS [11], these along with server based stand alone systems have been quite favorite in the aviation domain. Though translation systems related to aviation domain are scarce, NLP of which MT is a part has found its way into aviation applications in a big way.

3. NEURAL NETWORK MODELS

Neural networks are basically meant to simulate the way neurons in the human brain work. AI or artificial intelligence models generally learn by example, and then makes the model to make output predictions based on the knowledge it has learned during the training process, Abiodun et al.,[13]. Compared to statistical machine translation, the NMT systems are based on RNN/ANN to translate a text written in source natural language to the target language. The underlying idea in NMT is to embed the input sentence word by word into a thought vector of fixed length that is a representation of the whole sequence. The hidden layer of neurons transforms this thought vector to a representation that is decoded by the decoder and used for both training and prediction of the output. Multiple models are available for the purpose, such as RNN, LSTM-RNN etc.

3.1 RECURRENT NEURAL NETWORK

Recurrent neural networks also known as RNN model is the most favored as it has an internal memory. It can remember the inputs Morchid, [14]. Because of the presence of internal memory, RNNs are supposed to a robust and intelligent neural network. On using RNN in encoder-decoder architecture, the inputs are considered word by word and the consecutive loops again consider them one word at a time. As a result it can “remember” the previous steps and predict better during output generation.

3.2 LSTM

LSTM stands for long short term memory. LSTM is a type of RNN architecture. LSTM are better than feed forward neural networks, as they have feedback connections. LSTM finds its uses in deep-learning domain. LSTM networks are used in encoder-decoder models of OpenNMT, etc.

4. WORD REPRESENTATION

Words and data needs to be represented in a way or format that can be input into the MT model. That is text data has to be transformed to numerical form. This can be done through many methods out of which one-hot encoding method is described with an example.

4.1 ONE HOT ENCODING

For our research we need to convert the aviation sentences/sequences written in English to numerical form. To transfer each word in the sentence in a numeric form in MT, every word is transformed into one encoding vector which can then be provided as input to the translation model. An encoding vector is nothing but a vector with 0 at every index and a 1 at a single index which corresponds to that particular word. In this way, each word has a distinct encoding vector and thus with a numerical representation we can represent each and every word in our dataset. On the way to creating these vectors, assignment of an index to each unique word in the input (source) language, and the output (target, here Bengali) language is done. On the assignment of a unique index to each unique word, creation of vocabulary for each language is done. In ideal conditions, the vocabulary for both the source and target language will contain each unique word of the language. But, in real world any languages are known to have hundreds of thousands of words that are unique in them. So, in most cases the vocabulary is trimmed to the most common words in the dataset, denoted by N. N is chosen arbitrarily, but can range from 1,000–100,000 depending on the size of the dataset. To get a better understanding of how we can use a vocabulary to create one embedding vector for every word

in our dataset, let us consider a small vocabulary set of size 12 containing the words as shown in TABLE 1

a	0
the	1
green	2
red	3
apron	4
aircraft	5
street	6
landed	7
on	8
runway	9
<SOS>	10
<EOS>	11

Table1.Example of a mini-Vocabulary

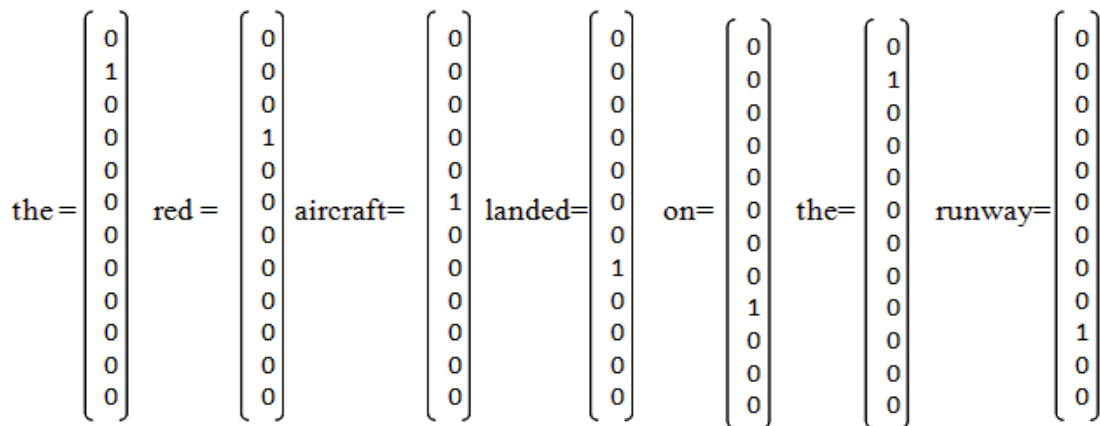


Figure.1.Representation of words as vectors

In the TABLE 1, for every word in our mini-vocabulary a unique index from 0 to 11 has been assigned. In the table start of the sentence is assigned the token <SOS> and end of the sentence is assigned the token <EOS>. The NMT model uses them to identify these crucial points in a sentence. Now, in order to convert the constituent words of the sentence “the red aircraft landed on the runway” to their respective embedding vectors, we use TABLE 1. It is done as shown below in Figure.1. In the figure we see that, each word transforms into a vector having length 12. (12 is also the size of our vocabulary) Each word also consists entirely of 0s except for a 1 at the index as was assigned in TABLE 1.

Here creation of a vocabulary for both the languages: input (source) and output (target) is done. Now this technique can be applied on each sentence in the source and target languages to transform into a format understandable by the encoder-decoder pair and that can be used for the task of machine translation.

5. IMPLEMENTATION OF THE AVIATION NMT SYSTEM

The implementation of the aviation NMT system was planned starting with understanding the need. The resource that was completely absent and needed to be created from the scratch was the English to Bengali aviation parallel corpus. Also as the aviation domain is composed of huge number out-of-vocabulary words and aviation phraseologies, the need for a parallel transliteration corpus was also felt. This corpus in the future acted as the data-repository that supported the development of the E-Dictionary [15], ATC phraseology converter [16] and OOV words converter [17]. To facilitate the pre-processing of the aviation sentences which are composed of OOV words, a preprocessing tool [18] was also created. This tool would also help in the creation of the translation corpus. The overview of the full system to be developed was constructed and it took the shape as depicted in Figure.2. After the final machine translation output was obtained and its BLEU score deducted, the output file was subjected to post-processing with the aviation phraseology converter tool [16] and OOV transliteration tool [17]. The BLEU score of this post-processed file was calculated again and a detail comparison study was conducted.

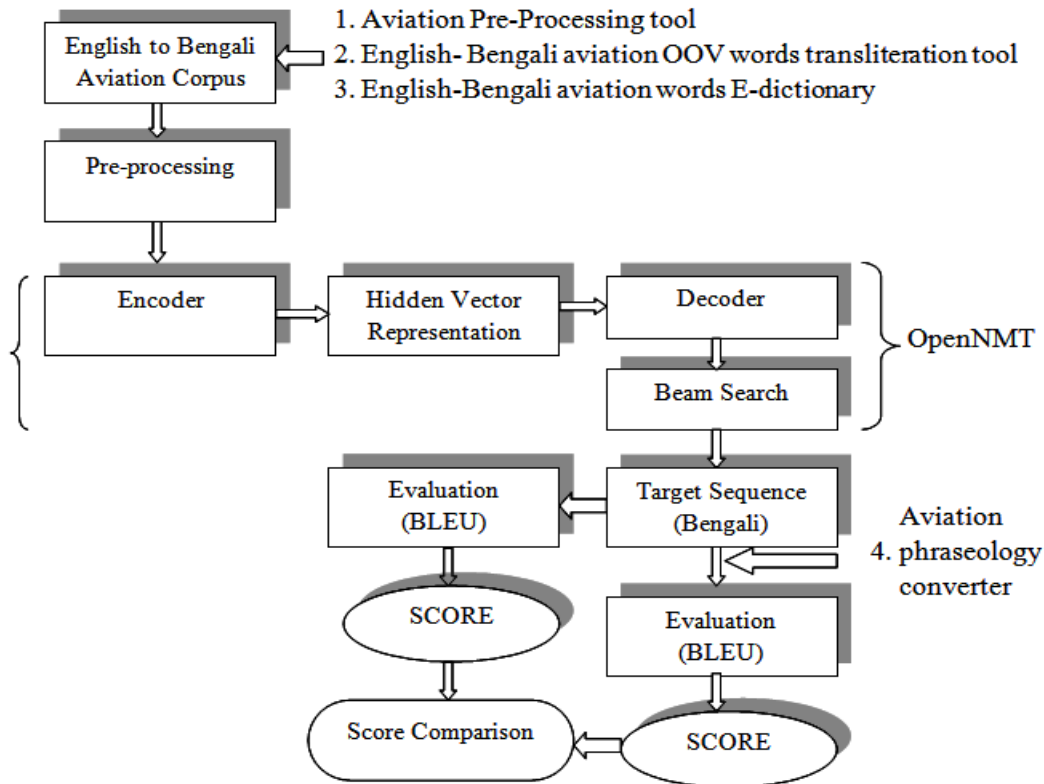


Figure.2.Overview of the English-Bengali aviation NMT system

5.1 CREATING THE ENGLISH TO BENGALIAVIATIONPARALLELCORPUS

Creation of the aviation parallel corpus was a challenge that was accomplished with the collection of data from multiple dedicated aviation sources such as AAI[19], DGCA[20],NASA-ASRS[21] and ECCAIRS [22]. The main constituent elements and features of the corpus are as given in TABLE2 and TABLE 3. The corpus creation was assisted by the E-dictionary [15] OOV transliteration tool [17] Pre-processing tool [18] that were created for this very purpose.

Domain	Language pair	Number (sentences)	Size
Aviation	English	19,959	1.5MB (in .txt)
	Bengali		3.9MB (in .txt)

Table2.Corpus size and features

Natural language	vocabulary size	ATC Phraseologies used	aviation OOV words used
English	22007	600	1200
Bengali	32265		

Table3.Vocabulary size and constituent elements of the corpus

The screenshot shows a spreadsheet with two columns: English text in column A and Bengali text in column B. The rows contain various aviation-related phrases and their translations. For example, 'it means true or accurate' is translated as 'এটার অর্থ সত্য বা নির্ভুল'. The spreadsheet is titled 'aeronautical stations are identified by name of location'.

English	Bengali
14204 it means true or accurate	এটার অর্থ সত্য বা নির্ভুল
14205 an error has been made in the message indicated	নির্দেশিত বার্তা একটি ত্রুটি হয়েছে
14206 an error has been made in the transmission	প্রেরণে একটি ত্রুটি হয়েছে
14207 did you mean the correct version	আপনি কি সঠিক সংস্করণে বোঝাতে চেয়েছিলেন
14208 the pilot is trying to ignore	বিমানচালকটি উপেক্ষা করার চেষ্টা করছে
14209 what is the readability of my transmission	আমার প্রেরণের স্পষ্টতা কেমন
14210 how do you read	আপনি কি সুত্তে পারছেন
14211 proceed with your message	বার্তাটি দিয়ে দাও
14212 pilot said i understand your message	বিমানচালক বোলো আমি আপনার বার্তা বুঝতে পেরেছি
14213 it means i will comply with your message	এর অর্থ আমি আপনার বার্তা মেনে চলবো
14214 it is not correct or not capable	এটি সঠিক নয় বা সক্ষম নয়
14215 permission not granted	অনুমতি দেওয়া হয়নি
14216 it means my transmission is ended	এটা বোঝায় আমার প্রেরণ শেষ
14217 i expect a response from you	আমি আপনার কাছ থেকে একটি প্রতিক্রিয়া আশা করি
14218 this exchange of transmission is ended	এই প্রেরণ বিনিময় শেষ
14219 no response is expected of this transmission	এই প্রেরণের কোন প্রতিক্রিয়া আশা করা হচ্ছে না
14220 indigo has the largest fleet of 210 planes	ইন্ডিগোর ২১০ টি বিমানের বৃহত্তম বিমানবহর রয়েছে
14221 to repeat all of the message this phrase is used	বার্তাটি পুরোপুরি পুনরায় পুনরাবৃত্তি করার জন্য এই বাক্যাংশ ব্যবহার করা হয়
14222 repeat the specified part of the message	বার্তাটির একটি নির্দিষ্ট অংশ পুনরায় পুনরাবৃত্তি করো
14223 last clearance is approved	শেষ অনুমোদনে অনুমোদিত
14224 a change has been made to the last clearance	শেষ অনুমোদনে একটি পরিবর্তন করা হয়েছে
14225 this new clearance supersedes the previous clearance	এই নতুন অনুমোদন পূর্বের অনুমোদন রহিত করে
14226 this phrase has been issued now	এই বাক্যাংশ এখন জারি করা হয়েছে
14227 to pass information this word is used	তথ্যটি পাস করার জন্য এই বাক্যাংশ ব্যবহার করা হয়
14228 pilot or controller may like to know something	বিমানচালক বা নিয়ন্ত্রক কিছু জানতে চেতে পারেন
14229 it has been used by the pilot	এটি বিমানচালক দ্বারা ব্যবহার করা হয়েছে

Figure.3.Example of the English to Bengali aviation parallel corpus used to train the NMT model

5.2 RUNNING THE ENGLISH-BENGLI PARALLELCORPUS IN OpenNMT

To pre-process the aviation parallel corpus and utilize it to create the NMT model, OpenNMT-py was employed. The pre-processing of the parallel corpus was done based on the requirement of the language pair. Normalization and tokenization was done for both the English and Bengali sentences while True-casing was done only for English sentences as Bengali does not have the concept of uppercase and lowercase. Figure.4. shows the layout of the pre-processing that is performed on the aviation parallel corpus. After preprocessing OpenNMT makes the parallel sentences

available to the Encoder-Decoder for the following purposes:

- The Encoder uses the preprocessed English sentences to create encodings for each word (plus the weight) and create the ultimate thought vector
- The Decoder uses the preprocessed words in the Bengali sentences and the thought vector to predict the possible output and calculates the loss.
- *Teacher forcing* is used to train the model in respective iterations.

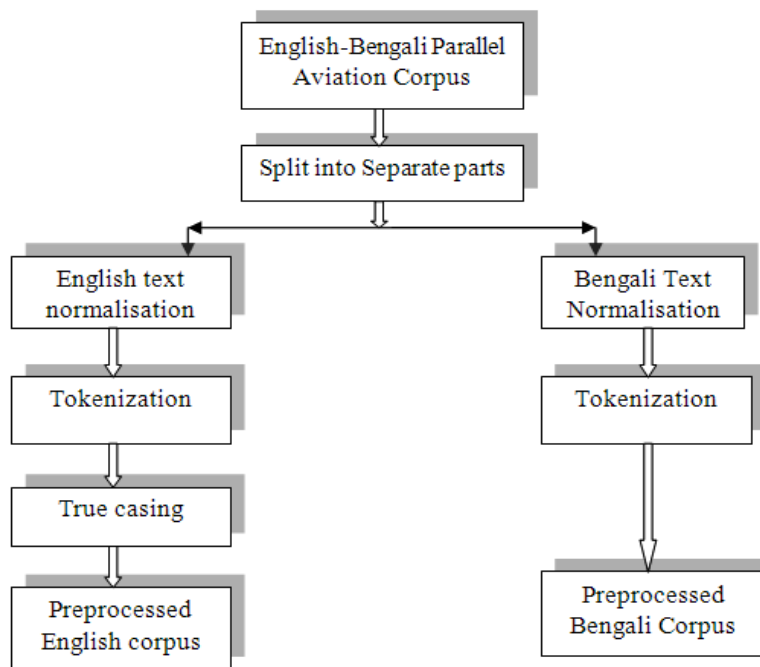


Figure.4.Pre-processing steps on the parallel corpus.

5.3 ENCODER-DECODER OPERATIONS

The encoder-decoder architecture for the aviation machine translation system in training mode is given in Figure.5. In the training mode the encoder takes in the English sentence word by word, adds weight to it leading to the creation of a *thought vector*. Figure.6.

depicts the creation of the *thought vector* for the English sentence “the aircraft has landed”. Here weights are added to the word encodings and the ultimate *thought vector* is created. The details of working of the encoder are described in the next section.

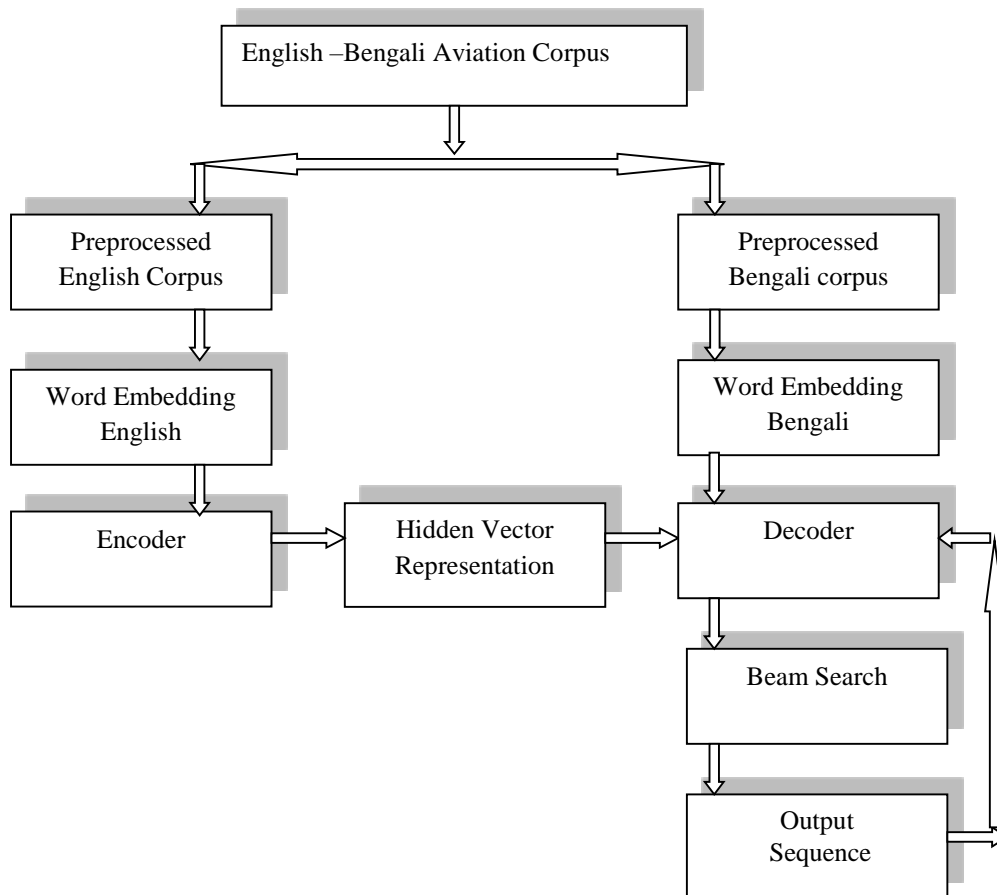


Figure.5.Training architecture of the NMT model

5.3.1. WORKING OF THE ENCODER

The encoder used is an LSTM-RNN responsible for creating the thought vector. It is responsible for adding weights (at each time step) and converting the input sequence word by word to the hidden vector. In our model the encoder reads the input source sequence and summarizes the information in something called as the internal state vectors i.e. hidden state and cell state vectors as it is LSTM. At each time step “t”, the model updates a hidden vector “h”, by using information from the source word that has been input to the model during that particular time step. The hidden vector stores information about the input sentence (word by word). So we can represent encoding of the sentence “The aircraft has landed” as given in Figure.6.

In Figure.6. Individual words in the input source sentence are fed into the encoder in a number of consecutive time steps, where

t= time step
 h = hidden state / vector
 c= cell state / vector
 W= Weight
 E= superscript E = hidden state of the encoder
 D= superscript D = hidden state of the decoder

In Figure.6. “W^E” represent weight matrices, which are used to achieve accurate translations. The final encoder hidden state, i.e. (h^E_{t=4}) ultimately becomes the “thought vector” and is marked with superscript “D” meaning decoder and time step is made “t=0”. In this way the final hidden vector of the encoder is placed as the initial hidden vector of the decoder. Thus we pass the encoded meaning of the source sentence to the decoder. Now it is to be translated to a sentence in the output or target language. But unlike the encoder, the decoder is needed to output variable length translated sentence. As a result, the decoder has to output a prediction word at each time step till we have an output of a complete sentence.

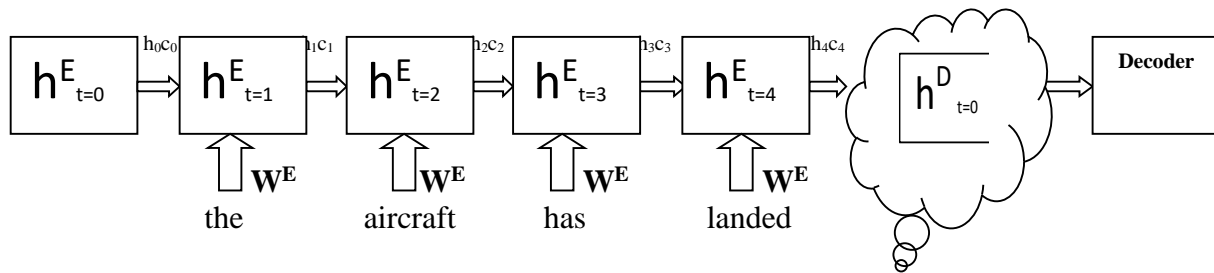


Figure.6. Creating the thought vector of the sentence “the aircraft has landed”

Ultimately, the encoder summary is as follows: the encoder reads the input sequence i.e. source sentence word by word and preserve the internal states of the LSTM network generated after the last time step “h₄, c₄”. “h₄c₄” (Figure6)

are known as the “encoding” of the input sentence/sequence. They summarize or encode the entire input sentence in a vector form. The decoder starts generating the output once it has read the entire sequence.

5.3.2. WORKING OF THE DECODER

While the encoder has only one role to play in both the phases of training and inference, the decoder has two different significant roles to play.

1. Training mode
2. Inference mode

In the training mode the decoder takes in the aviation English sentence *thought vectors* and then uses it to train the model through predicting the corresponding Bengali

words in each loop. Then on calculating the loss it uses *Teacher forcing* in the following iterations to perfect itself. In the translation mode the decoder simply predicts the output sequence word by word depending on the training it received. Figure.7. depicts the architecture of encoder-decoder in Testing / Translation mode where <SOS> and <EOS> are the unique index for start of sentence and end of sentence respectively. These two indexes are used to mark the start and end of the sentences. These English and Bengali sentences are composed of the combination of words that have formed the vocabulary of our aviation parallel corpus. They number in 22007 and 32265 respectively for English and Bengali.

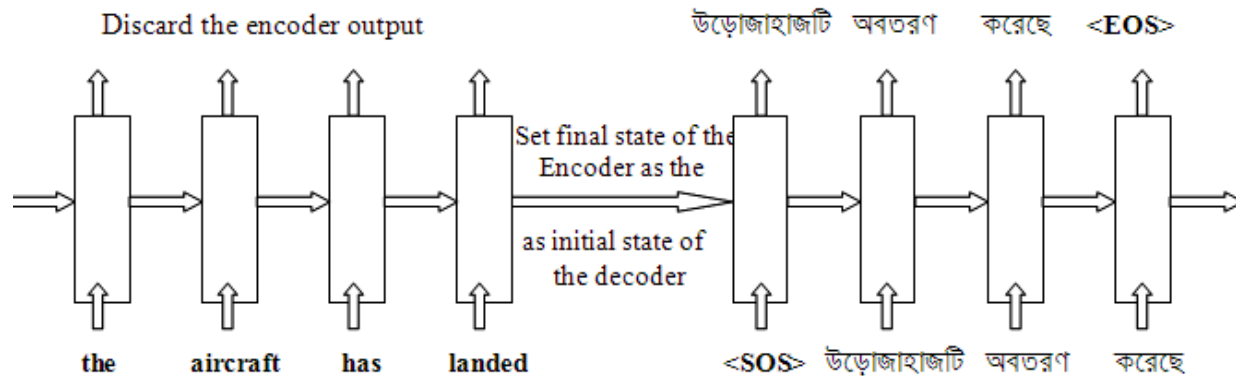


Figure.7. Translation model for the aviation MT system

6. RESULTS AND EVALUATION

The English to Bengali aviation machine translation system has been implemented by training and validating the model with the English-Bengali aviation corpus. The corpus has been broken up for training and validation. Both the training and validation files are in turn grouped into source and target files, the details of which are given in TABLE 4. For creating the aviation model, on running the corpus, the OpenNMT calculation for the “src” (Source) and “trt” (Target) vocabulary and their embedding for the RNN encoder and decoder are computed as follows:

```
(jefson08)
jefson08@DonCorleone:~/Codes/OpenNM
T-py$ onmt_train -data data/demo -
save_model_demo-model -gpu_ranks 0
[2020-08-14 23:03:54,256 INFO]
*src vocab size = 22007
[2020-08-14 23:03:54,256 INFO]
*tgt vocab size = 32265
[2020-08-14 23:03:54,256 INFO]
Building model...
[2020-08-14 23:03:57,165 INFO]
NMTModel(
  (encoder): RNNEncoder(
    (embeddings): Embeddings(
      (make_embedding): Sequential(
```

```
(emb_luts): Elementwise(
  (0): Embedding(22007,
500, padding_idx=1)
  )
  )
  (rnn): LSTM(500, 500,
num_layers=2, dropout=0.3)
  )
  (decoder): InputFeedRNNDecoder(
    (embeddings): Embeddings(
      (make_embedding): Sequential(
        (emb_luts): Elementwise(
          (0): Embedding(32265,
500, padding_idx=1)
          )
        )
        (dropout): Dropout(p=0.3,
inplace=False)
        (rnn): StackedLSTM(
          (dropout): Dropout(p=0.3,
inplace=False)
          (layers): ModuleList(
            (0): LSTMCell(1000, 500)
            (1): LSTMCell(500, 500)
          )
        )
      )
    )
  )
)
```

Corpus Details	Files	Parallel Sentences	Size
Main Corpus	English.txt	19959	1.5 MB
	Bengali.txt		3.9 MB
Train Corpus	src-train.txt	16239	1.2 MB
	tgt-train.txt		3.1 MB
Validate/Test Corpus	src-aviation_only_translation.txt	3720	311.6 KB
	tgt-aviation_only_translation.txt		780.7 KB

Table4.Number of sentences and file structure used for training and validation

The total number of iterations performed by the OpenNMT is 100000. The aviation NMT model creates the output file named “*pred_aviation_only_translation.txt*” This file is subjected to both manual evaluation process and BLEU machine evaluation system. For manual evaluation only the machine translated file is considered while for the BLEU evaluation both the machine translated file and the post-processed machine translated file is considered.

6.1 MANUAL EVALUATION METHOD

In this method the TA or translation accuracy of the machine translation system is evaluated on the adequacy and fluency of the input and output texts. The adequacy checks are made in order to find if the meaning of the sentences in the source and target text is same or not. At the same time fluency checks are made to determine if the translated sentences are grammatically correct or not. Often these manual evaluation methods are made with professional evaluators. Manual evaluation method is time consuming and in certain cases, a costly procedure. But it remains the most accurate evaluation technique of machine translation outputs. The main requirement of a manual evaluator is that the evaluator should have proper knowledge of both the source and target language. To evaluate the translation accuracy of the

English to Bengali aviation machine translation system, human evaluation technique has been used along with BLEU. The analysis of the output in Bengali compared to the input in English has been done by a linguist. The translation has been evaluated on the basis of percentage of translation accuracy of the English and translated Bengali sentences. On dividing the percentage levels of the translation accuracy into different levels, their meaning can be put as follows:

Levels	Meaning
perfect	easy to understand
fair	need minor correction
acceptable	broken but understandable
unacceptable	not understandable

Table5. Levels of manual evaluation technique

With the levels of evaluation fixed, the analysis found the approximate levels of the TA (in percentage) as given in TABLE 5 and Figure.8.

Levels	Translated aviation sentences in Bengali
perfect	42 Percent
fair	36 percent
acceptable	14 percent
unacceptable	8 percent

Table6. Levels of TA percentage of the translated aviation sentences in target language

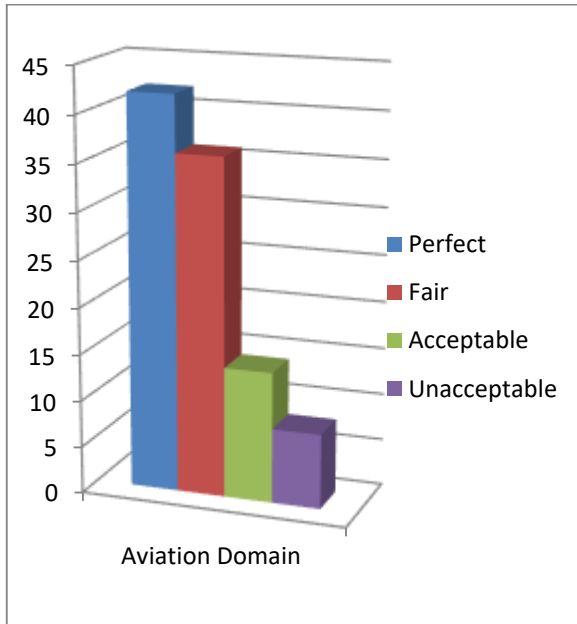


Figure.8.Graphical representation of the manual evaluation results

6.2 BLEU

Being an inexpensive, language independent and efficient evaluation technique Bilingual Evaluation Understudy or BLEU is among the most popular evaluation technique for SMT and NMT system outputs. The command to find the BLEU score of the aviation English to Bengali MT system is as shown in Figure.9. In the figure, “tgt-aviation_only_translation.txt” is the human translated file while “pred_aviation_only_translation.txt” is the machine translated file. On comparing them the BLEU score is found to be 39.97 for the aviation NMT system. TABLE 6 gives the details of the files used and BLEU score.

```

Applications: Terminal
Aug29 11:36 AM 30.5°C
jefson08@DonCorLeone: ~/Codes/OpenNMT-py
(jefson08) jefson08@DonCorLeone:~/Codes/OpenNMT-py$ perl tools/multi-bleu.perl data/tgt-aviation_only_translation.txt < pred_aviation_only_translation.txt
BLEU = 39.97, 49.9/43.1/40.6/38.7 (BP=0.932, ratio=0.934, hyp_len=22273, ref_len=23848)
(jefson08) jefson08@DonCorLeone:~/Codes/OpenNMT-py$

```

Figure.9.Command to evaluate the BLEU score

Aviation Corpus	Files	BLEU score
Main Corpus	English.txt	39.97
	Bengali.txt	
Train Corpus	src-train.txt	
	tgt-train.txt	
Validate/Test Corpus	src-aviation_only_translation.txt	
	tgt-aviation_only_translation.txt	
Machine Translated output file	pred_aviation_only_translation.txt	

Table7.BLEU score of the aviation English to Bengali MT system

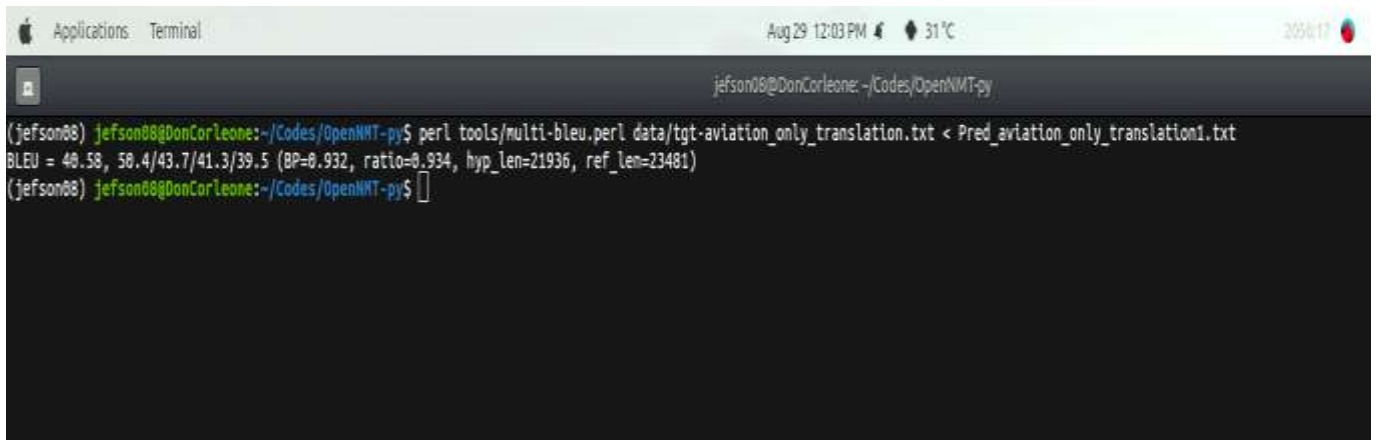
6.2.1 APPLICATION OF THE POST PROCESSING TOOL AND RESULT COMPARISON

In the aviation machine translation output file “*pred_aviation_only_translation.txt*” (consider as **output-file1**), we observed that some aviation OOV words, aviation phraseologies, proper nouns and technical terms remain untransliterated. This results in reducing the accuracy of the translation, its BLEU score and manual score. So the OOV words transliteration tool [17] and aviation phraseology converter [16] are both used as post-processing to increase the translation accuracy count.

Though time consuming this is done manually as manual transliteration provides accurate results.

So, after using the transliteration tools on the output of the English to Bengali aviation machine translation system we again calculate the BLEU score as given in Figure.10.

This time the name of the updated file is “*pred_aviation_only_translation.txt1*” (Consider as **output-file2**) which is compared with the target validation file (human translated) “*tgt-aviation_only_translation.txt*”. On computing the BLEU score it was found to be 40.58.



```

Applications: Terminal
Aug 29 12:03 PM 31°C
jefson08@DonCorleone: ~/Codes/OpenNMT-py
(jefson08) jefson08@DonCorleone:~/Codes/OpenNMT-py$ perl tools/multi-bleu.perl data/tgt-aviation_only_translation.txt < Pred_aviation_only_translation1.txt
BLEU = 40.58, 58.4/43.7/41.3/39.5 (BP=0.932, ratio=0.934, hyp_len=21936, ref_len=23481)
(jefson08) jefson08@DonCorleone:~/Codes/OpenNMT-py$

```

Figure.10.Command to evaluate the BLEU score of the post-processed file.

The BLEU scores of the output files of aviation MT system, before and after application of the post-processing tools are as follows:

Domain	BLEU SCORE	
	Before using Transliteration Tools	After Using Transliteration Tools
Aviation	39.97	40.58

Table8.BLEU score comparison

Now, we compare the same files at “<https://www.letsmt.eu/Bleu.aspx>” to get a more detail BLEU result. Here we can enter the human

translated validation file and machine translated files (output-file1 and output-file2) and gets not only the BLEU score but also the individual and cumulative n-gram results. There are two types of BLEU score: Individual and cumulative. An individual n-gram score means that it is the result of just matching grams of one specific single order, such as single words or 1-gram, word pairs or 2 gram also known as bigram. The attached weights are specified as a unique one where the gram orders are referred to by each index. Whereas, cumulative scores mean the calculation of individual n-gram scores, all orders from 1 to n along with weighting them by calculating the weighted GM. Both the output files are evaluated and compared with each other using BLEU score metrics. While Figure.11. gives us the detail evaluation result of output-file1 Figure.12 gives us the detail evaluation result of output-file2.

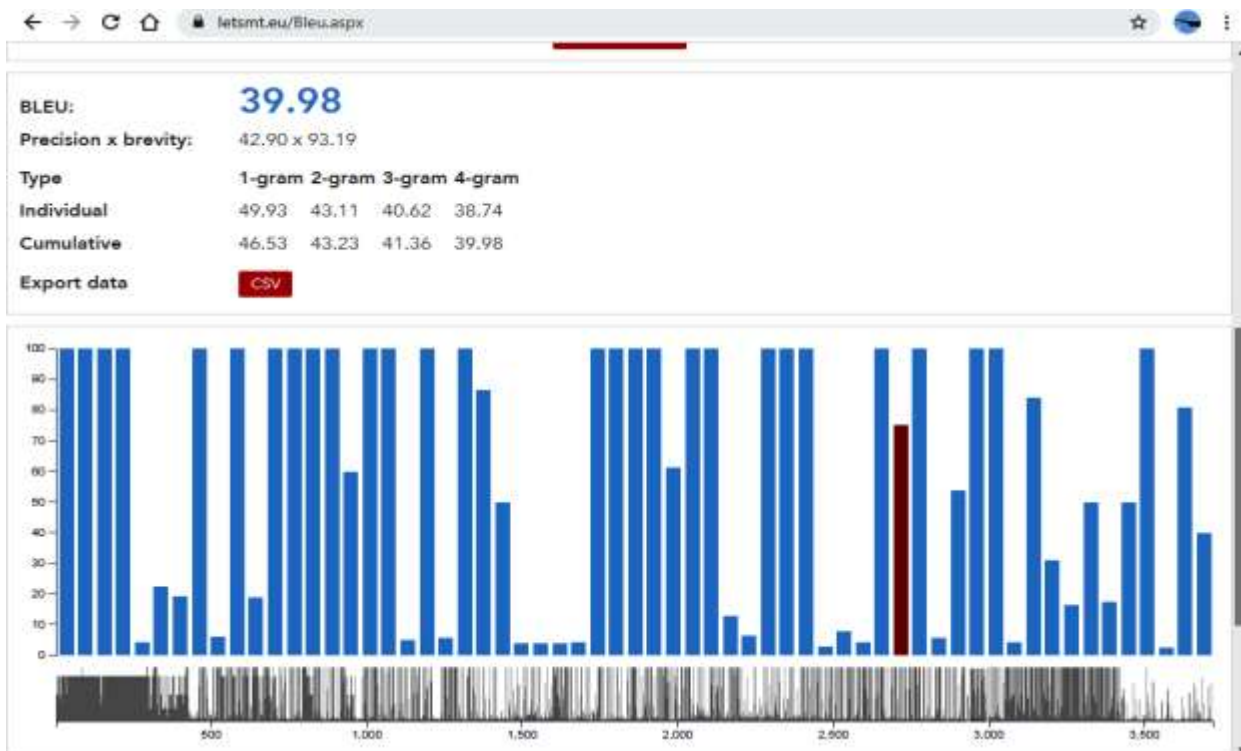


Figure.11.Detail BLEU evaluation for output-file1



Figure.12.Detail BLEU evaluation for output-file2.

As noted from Figure.11. and Figure.12. the BLEU score obtained from output-file2 is better than that obtained from output-file1. Again from Figure13. we observe that the for cumulative n-gram score of 1-gram, 2-gram, 3-gram and 4-gram output-file2 scores better then output-file1.As all

NMT systems are dependent on size of parallel corpus and its quality, increasing the size of the corpus and including more aviation terminologies will help improve the translation quality and in turn the translation accuracy (TA).

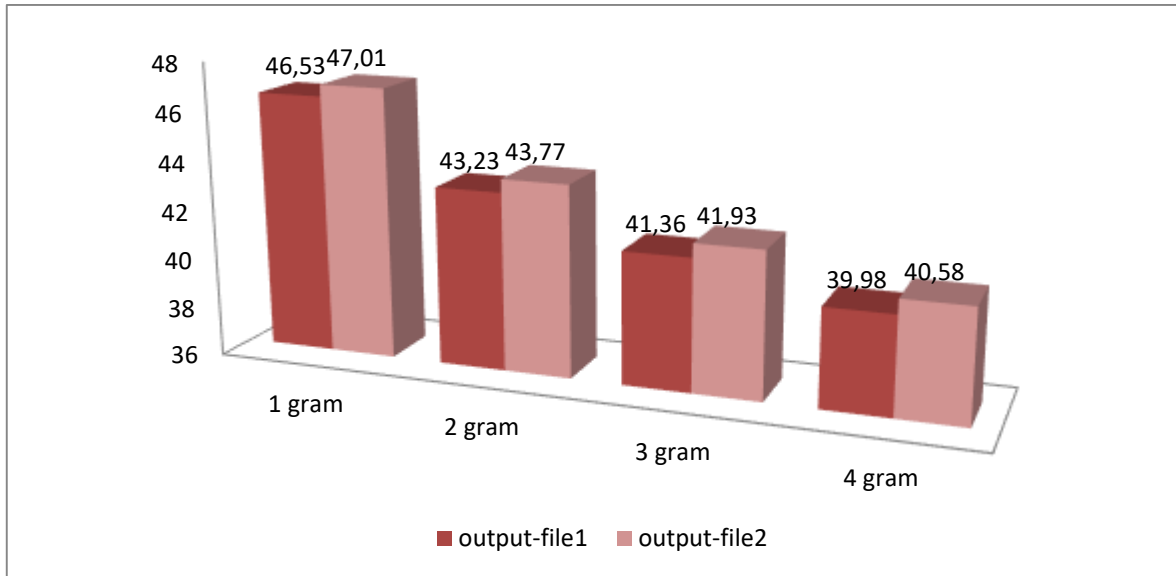


Figure.13. Comparison of cumulative n-gram scores for output-file1 and output-file2

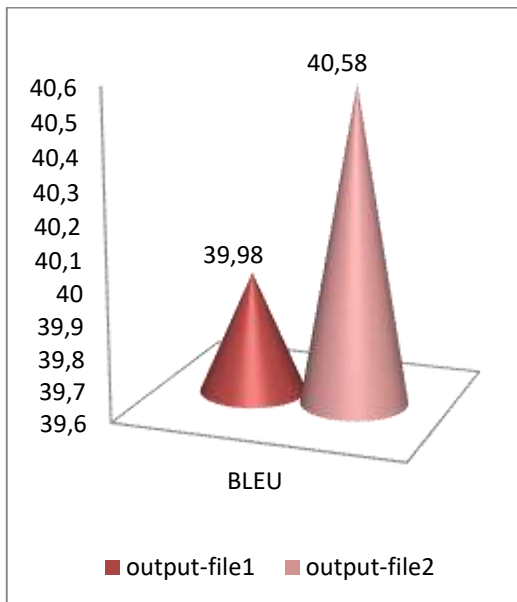


Figure.14. BLEU score comparison between output-file1 and output-file2

in the Figure.13. and Figure.14. we see the detail score comparison of the two output files. While Figure.13 shows us the comparative study of the cumulative n-gram for both the output files, the Figure.14. shows us the BLEU score comparison of the output files. From the results we conclude that the output-file2 has better translation accuracy than

output-file1. That is, the post-processed output file exhibits better results in translation accuracy (TA). We can summarize the strengths of the aviation NMT system as following:

- The model or translation system as a whole is based on neural networks that are capable of giving predictions based on the “teachings” and “trainings” done on the basis of the aviation parallel corpus.
- Along with the use of individual words, we have assigned 1200 OOV words unique to aviation and 600 aviation phraseologies for training. This allows the aviation NMT system to handle aviation words and sentences better than any existing system.
- It is the first known translation system involving Bengali for the aviation domain.
- The training datasets (parallel corpus) are available together and individually for both translation and transliteration applications.

7. CONCLUSION

In this paper, we have presented a system based on neural translation system for the aviation domain. This system was implemented using a LSTM-RNN, and the steps have been described in detail. The system was trained using a unique English-Bengali parallel corpus containing about 19959 aviation sentences which in turn has made use of 1200 aviation OOV words and 600 aviation phraseologies. The model was evaluated using both manual evaluation method and BLEU score. The output is post processed and the BLEU score of the post-processed file is also calculated and compared to the first version. BLEU score of 39.98 and 40.58 are scored for the two files. Considering that it is the first time such an attempt has been made for a technical domain such as aviation in an Indian language, the results for both the manual score and BLEU score is considered to be satisfactory. It was observed that for the English to Bengali aviation machine translation system the BLEU score depend on the quality and size of the parallel corpus and also the proper normalization of the aviation OOV words and phraseologies. Higher BLEU score denotes better translation. While the pre-processing tool and E-dictionary helped in the development of the aviation corpus, the application of the transliteration tools on the output of the MT system have helped improve the translation accuracy of the aviation machine translation system as a whole. Although these results have been found to be encouraging, the translation accuracy of the NMT system can be improved. By increasing the size of the English to Bengali aviation parallel corpus and including more aviation related OOV words and phraseologies the prediction results can be made more accurate i.e. the TA can be increased.

REFERENCES

- [1] Kalchbrenner, N., Blunsom, P., 2013. Recurrent continuous translation models. In: Yarowsky, D., Baldwin, T., Korhonen, A., Livescu, K., Bethard, S. (Eds.), Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Seattle, pp.1700–1709.
- [2] Sutskever, I., Vinyals, O., Le, Q.V., 2014. Sequence to sequence learning with neural networks. In: Proceedings of the 27th International Conference on Neural Information Processing Systems, 2, 3104–3112.
- [3] Cho, K., Merriënboer, B.v., Bahdanau, D., Bengio, Y., 2014. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. arXiv preprint arXiv:1409.1259.
- [4] Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., et al., 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. arXiv:1609.08144.
- [5] Crego, J., Kim, J., Klein, G., Rebollo, A., Yang, K., Akhanov, J. S., et al., 2016. SYSTRAN's Pure Neural Machine Translation Systems. CoRR abs/1610.05540.
- [6] Mistry, J.G., Verma, A., Bhattacharyya, P., 2017. Literature Survey: Study of Neural Machine Translation. Resource Centre for Indian Language Technology Solutions (CFILT).
- [7] Shah Nawaz, Mishra, R.B., 2012. A neural network based approach for English to Hindi machine translation. Int. J. Computer Applications, pp 50–56.
- [8] Mishra, R., Marwan, A., 2014. ANN and rule based method for English to Arabic machine translation. Int. Arab J. Information Technol., 396–405
- [9] Costa-jussa, M.R., Zampieri, M., Pal, S., 2018. A Neural Approach to Language Variety Translation. VarDial@COLING.
- [10] Pierre Isabelle and Laurent Bourbeau, "TAUM-AVIATION: its Technical Features and some experimental Results", Computational Linguistics, Volume 11, Number 1, January-March 1985. Pages: 18-27.
- [11] Saptarshipaul, Bipulshyampurkhyastha, "NLP tools used in civil aviation: A Survey", International journal of advanced research in computer science, volume 9, Number 2, March-April-2018
- [12] Peter Clark and Phil Harrison "Boeing's NLP System and the Challenges of Semantic Representation" WA 98124, ACL , STEP '08 Proceedings of the 2008 Conference on Semantics in Text Processing, Pages 263-276, Venice, Italy, September 22 - 24, 2008
- [13] Abiodun, O.I., Jantan, A., Omolara, A.E., Dada, K.V., Mohamed, N.A., Arshad, H., 2018. State-of-the-art in artificial neural network applications: a survey. Heliyon.
- [14] Morchid, M., 2017. Internal Memory Gate for Recurrent Neural Networks with Application to Spoken Language Understanding. INTERSPEECH.
- [15] Saptarshipaul, "Bilingual (English to Bengali) Technical E- Dictionary for Aviation OOV Words", International journal of Engineering and advanced technology, volume 9, issue 2, December 2019
- [16] Saptarshi Paul, BipulshyamPurkhyastha, "An NLP tool for decoding the ATC Phraseology from English to Bengali", proceedings of the International conference on smart electronics and communication (ISOSEC2020), 10-12 September 2020 (IEEE XPLORE complaint, ISBN: 987-1-7281-5461-93)
- [17] Saptarshi Paul, BipulshyamPurkhyastha. English to Bengali Transliteration Tool for OOV words common in Indian Civil Aviation. Journal of Advanced Database Management & Systems.2019; 6(1): 23–32p.
- [18] Saptarshi Paul, BipulshyamPurkhyastha, "Preprocessing Aviation Slangs and OOV Words for Machine Translation", proceedings of International conference in recent trends in Electronics and Computer Science, (ICRTECS-2019) organized by NIT-Silchar , March 18 to 19, 2019.
- [19] <https://www.aai.aero/hi/system/files/resources/>
- [20] http://dgca.nic.in/accident/reports/contents_acc_rep.html
- [21] <https://asrs.arc.nasa.gov/>
- [22] <https://ec.europa.eu/jrc/en/scientific-tool/eccairs-european-central-repository-aviation-accident-and-incident-reports>

