Practicum Report

Project Title:	Investigating the Correlation between Quality and Lexical Diversity in Training of Machine Translation Engines
Student ID:	19210273, 19211293
Student name:	Souradip Goswami, Aritra Dutta
Student email	souradip.goswami2@mail.dcu.ie, aritra.dutta2@mail.dcu.ie
Chosen major:	MSc in Computing (Data Analytics)
Supervisor	Dr. Dimitar Shterionov
Date of Submission	17/08/2020

DISCLAIMER

A report submitted to Dublin City University, School of Computing for Practicum, 2019/2020.

We understand that the University regards breaches of academic integrity and plagiarism as grave and serious. We have read and understood the DCU Academic Integrity and Plagiarism Policy.

We accept the penalties that may be imposed should we engage in practice or practices that breach this policy.

We have identified and included the source of all facts, ideas, opinions, viewpoints of others in the assignment references. Direct quotations, paraphrasing, discussion of ideas from books, journal articles, internet sources, module text, or any other source whatsoever are acknowledged, and the sources cited are identified in the assignment references.

We declare that this material, which we now submit for assessment, is entirely our own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

By signing this form or by submitting this material online we confirm that this assignment, or any part of it, has not been previously submitted by us or any other person for assessment on this or any other course of study. By signing this form or by submitting material for assessment online We confirm that we have read and understood DCU Academic Integrity and Plagiarism Policy (available at: http://www.dcu.ie/registry/examinations/index.shtml)

Name(s):

Souradip Goswami

Aritra Dutta

Date: 17 August 2020

Investigating the Correlation between Quality and Lexical Diversity in Training of Machine Translation Engines

Dimitar Shterionov
Assistant Professor
Tilburg University
Tilburg, Netherlands
dimitar.shterionov@adaptcentre.ie

Souradip Goswami
School of Computing
Dublin City University
Dublin, Ireland
souradip.goswami2@mail.dcu.ie

Aritra Dutta
School of Computing
Dublin City University
Dublin, Ireland
aritra.dutta2@mail.dcu.ie

Abstract*—Machine Translation (MT) is the automated process of converting a source language into a target language. Neural Machine Translation (NMT) is a relatively new MT paradigm that has shown great potential in solving complex and challenging translation tasks. Despite its superior translation performance (in terms of accuracy and fluency) to other MT paradigms, there is still a long way to go for NMT to reach human performance. Due to the differences between machine translated text and human generated (either translated or authored) text, the former is referred in literature as *Machine Translationese*. The prime objective of our work is to study two aspects of machine translationese – its quality compared to human-generated text and its lexical richness. We will investigate how different sentences in $English \leftrightarrow Bengali$, $English \leftrightarrow French$ and $English \leftrightarrow Spanish$ are translated by NMT models in different stages of its training and thereby compare lexical diversity and automatic evaluation scores. The ultimate goal is to identify a correlation between lexical diversity and evaluation metrics during the training process, trying to answer the question whether *the best model in terms of evaluation metrics is the best in terms of lexical richness and vice versa.

I. Introduction

Neural Machine Translation is a state-of-the-art corpus-based machine translation paradigm with the potential of overcoming many shortcomings of traditional statistical machine translation (SMT) systems. In spite of generating more accurate translations compared to the traditional systems, maintaining lexical richness and creating diverse outputs has to be a priority for the NMT systems. But in this regard much progress has not been made yet and NMT is quite far off from the human translated sentences. And one of the major problems that still persists is to retain the lexical richness levels typical for human translations. [Vanmassenhove et al., 2019].

Lexical richness refers to the range of different words that are used in a text. The diversity increases with the increase in the range of unique words. Berman [Berman, 2000] observed that translation results in loss of lexical richness and diversity and referred to it as 'quantitative impoverishment'. Berman studied this lexical loss from a theoretical point of view whereas Kruger used an empirical approach and came to a similar conclusion. The tendency of NMT models to over-generalize frequent words is one of the major factors contributing to the loss of richness in machine translationese as shown by the recent works of [Vanmassenhove et al., 2019] and [Toral, 2019].

As indicated by [Vanmassenhove et al., 2019], MT systems, including NMT, have difficulties in dealing with ambiguous words i.e. words having more than one meaning, as they learn the most frequent translation variant and omit the less frequent variant(s). For example, English words like 'orange', 'watch' etc. have more than one meaning and there are cases where NMT systems falter in translating these words correctly based on the context.

The objective of this paper is twofold: (i) to investigate whether there is a correlation between lexical richness and automatic evaluation scores, i.e. BLEU and TER; and (ii) to investigate to what degree NMT can handle ambiguous words in their proper context. To do so, we will conduct a set of experiments assessing lexical diversity as well as evaluation scores at different stages of training process of NMT systems. And then we will be looking into Pearson's correlation coefficient between the evaluation and lexical diversity scores and compare them with the help of facts and figures. The NMT models will then be assessed using human-crafted sentences of ambiguous words in their proper context to see whether NMT can retain the correct meaning.

II. Related Work

Neural Machine translation is a corpus-based translation system where the training is done on huge corpora of source and target sentences; it exploits the encoder-decoder principle. The work of Mikel L. Forcada [Forcada, 2017] provides a clear picture of how NMT works and its advantages over Phrase-based statistical machine translation (PB-SMT) [Callison-Burch et al., 2011]. NMT training process is iterative in nature, and unlike PB-SMT, intermediate models can be easily stored and used. One can determine the best model even if it is

not the latest iteration.[1] In our work we train NMT models to translate $English \rightarrow French$, $English \rightarrow Spanish$ and $English \rightarrow Bengali$ and vice versa.

In the field of Machine Translation, it was Berman [Berman, 2000] who first noted the deforming tendencies that are ought to happen while translating. Lexical loss is one of those vital deformations that is almost inevitable. Kruger also compared human [Kotze, 2012] translated texts to the machine translated ones and found out that the machine translated texts are much simpler. The work of Eva Vanmassenhove [Vanmassenhove et al., 2019] presented an empirical approach to quantify the loss of lexical richness in Machine Translation systems compared to Human Translation. They could identify that the more frequent words observed in the input occur even more in the output whereas the less frequent words tend to get lost in translation. As a result, the translated text gets more generalized with certain words getting completely lost. Thus, the diversity gets compromised to great extent compared to human translation. Our work is an extension to this research to assess the lexical richness at different stages of NMT training.

Whether an NMT systems is acceptable for use or not is judged by the community based on the evaluation scores and mainly on BLEU. The MT quality is typically measured via human evaluation, automatic evaluation and task-based evaluation as suggestive from the work of [Shterionov et al., 2018]. In our work we use BLEU and TER for automatic evaluation scores. However, in many use-cases retaining the lexical diversity is also of importance. This can be measured by YULE's I [Yule, 1944], Measurement of Translation Lexical Diversity (MTLD) [McCarthy, 2005] and Type-token ration (TTR) [Templin, 1975]. The works of [Tomás et al., 2003] provide details about the different evaluation techniques for Machine Translation. The methods include Translation Error Rate (TER), Metric for Evaluation of Translation with Explicit Ordering (METEOR) and Bilingual Evaluation Understudy (BLEU) score. We have taken references from these works and these evaluation scores form the basis of our research work where we are trying to find the correlation between different scores and thereby judge various models at different stages of their training.

NMT is a corpus dependent MT paradigm and the performance of the models are heavily dependent on the source and target sentences on which it is trained. One of the major aspect of our work emphasises on $English \leftrightarrow Bengali$ language pair. As we know, Bengali is a low resource language unlike French, English, Spanish or others. To achieve good performance of the NMT models it is essential to find enough reliable data and ensure that the data is authentic. Otherwise it may have an adverse effect on the evaluation scores resulting in poor model performance. As suggestive from the words of [Agic and Vulic, 2019a], shortage of resources impose a development and evaluation bottleneck in multilingual processing. Therefore, we have used JW300, a parallel corpus of over 300 languages with around 100 thousand parallel sentences per language pair on average to train our models.

The works of [Johnson et al., 2016] focuses on using a single NMT model to translate between multiple languages without significantly changing the basic NMT model. It presents a Zero shot translation method to translate between languages it has never been trained on. It is an interesting technique for low-resource languages like the ones we are investigating, i.e. Bengali. If the quality of our English $\leftrightarrow$ Bengali engines is not up to the mark, we aim to follow this approach to improve the performance.

The works of [Toral, 2019] tells us how Machine Translation differs from Human Translation, the characteristics of post-editese and how we can distinguish between Human and Machine translated texts. It describes vividly about lexical diversity and how the lexical density and diversity can be a distinguishing factor in judging translated texts. As our work focuses on correlating the lexical diversity scores, we have taken references from this paper to understand the know how of the lexical diversity and richness of MT systems.

The works of [Bahdanau et al., 2014] opened a new direction in the field of NMT. They extended basic encoder–decoder by letting a model (soft-)search for a set of input words, or their annotations computed by an encoder, when generating each target word. The proposed model was very promising and a stepping stone towards the progress of NMT models. Our work being completely associated with NMT, this paper was really helpful in making us understand the NMT structure.

[Koppel and Ordan, 2011] studied the reasons for differences between translated and original texts. They came to the conclusion that both interference of the source language spilling into translated texts and general impacts of translation hampers the translated texts and makes them significantly different from the original ones. They confirmed that source language is indeed a factor in determining the success of translation.

III. Datasets

A. EN-BN

English-Bengali (EN-BN) is a low-resource language pair. For our EN-BN experiments, we collected data from various sources, and pre-processed and cleaned them to compile a suitable training set that would ensure the trained MT models are of (i) sufficient quality and (ii) robust in nature. We used the recent JW300 [Agic and Vulic, 2019b] parallel corpus. JW300 contains approximately 100 million sentences and a total of 1.5 billion tokens spanning across 343 languages.

The source of the JW300 data are all publications from the www.jw.org website; the texts in this website are from various magazines. The magazines contain an immense variety of topics and the contents are organized by language and article, where articles have a unique identifier. As a result, all translations of the same article carry the same identifier value (making it parallel).

[1]Due to over-fitting, the latest iterations could perform very good based on the development set, while their performance degrades on the test set. As such one can freely choose a model of another iteration as the best one.

We also used EN-BN data from the OpenSubtitles [Lison and Tiedemann, 2016] and the GlobalVoices [Tiedemann, 2012] corpora. The former contains subtitles of movies and TV shows collected from the OpenSubtitles website[2]; the latter consists of news stories from the GlobalVoices[3] website.

To build our Bengali-English (BN-EN) and English-Bengali (EN-BN) NMT engines, we compiled a corpus of approximately 900,000 parallel sentences. To do so, first we retrieved the EN-BN part of the JW300 corpus. Our initial assessment of the aligned data revealed significant misalignment, empty lines and lines containing two or more sentences. These inconsistencies could impact the quality of the NMT system. As such we decided to pre-process the data to resolve these issues by: (i) splitting lines containing '.', '!', '?' and ' ।'[4] in the source and the target sides; (ii) removing empty lines in the source and the target sides and (iii) re-aligning the corpus using hunalign [Varga et al., 2007] and an English-Bengali dictionary[5]. From the original 366,972 (mis)aligned sentences we compiled a new parallel corpus of 342,442 sentences. Second, we retrieved the EN-BN OpenSubtitles corpus [Lison and Tiedemann, 2016]. Third, we retrieved the GlobalVoices corpus for EN-BN which contained 137,620 sentence pairs. Then we concatenated the filtered JW300, the OpenSubtitles and the GlobalVoices corpora and removed duplicate entries. Thus we created our final EN-BN corpus composed of 797,592 parallel sentences.

B. EN-ES, EN-FR

The data for EN-ES and EN-FR has been taken from the Europarl corpora [Koehn, 2005]. Approximately 1.5M randomly selected sentences had been used after duplicates and empty lines were removed from the dataset. For both EN-FR/FR-EN and EN-ES/ES-EN, the 1.5M parallel sentences shown in previous work led to decent MT performance.

The training, test and development data for each language pair have been tokenised, cleaned and split into sub-word units using Byte Pair Encoding (BPE) [Sennrich et al., 2016]. We present statistics about our data, i.e. number of parallel sentences used for training, testing and development, as well as the dictionary sizes in Table I.

Language pair	Number of parallel sents.			Voc. size	
	Train set	Test set	Dev. set	L1	L2
EN-ES	1,472,203	1,000	5,734	47,628	48,459
EN-FR	1,467,289	1,000	7,723	47,639	49,283
EN-BN	897,680	1,000	1,000	50,141	51,073

TABLE I: Data statistics: number of parallel sentences, and vocabulary sizes. L1 and L2 denote the two languages in the language pair. E.g. for EN-ES, L1 denotes EN and L2 denotes ES.

[2]http://www.opensubtitles.org/
[3]https://globalvoices.org/
[4]The symbol is the Bengali version of full stop.
[5]https://github.com/MinhasKamal/BengaliDictionary

C. Ambiguous Word List

One of the major criteria to judge the NMT systems is not only to check for correct translations but also ensure that the translation conveys the context and the intention of the source text without tampering the meaning. As a part of our study, we looked into a number of ambiguous words in English i.e. same words with different meanings and tried to analyse how different NMT engines like Google translator and Bing Microsoft Translator translate those English words into Bengali without altering the meaning and context. Finally, we tested our EN-BN models with those ambiguous words and generated the evaluation scores for the same. Let's delve into this further with the help of an example. Consider the simple example of the word "orange". In English "orange" can be used to describe a *fruit* or a *color*. The context is usually enough to judge whether a particular sentence signifies orange as a fruit or a colour. We created a list of around ninety such words and compiled sentences with them in such a way that both meanings are expressed and the context is clearly stated. Then we converted these sentences into Bengali using Google translate and Bing Microsoft Translator to check whether these translators are able to pick the context and convert the ambiguous words properly. Finally, we tested our EN-BN NMT models with those ambiguous words to generate evaluation scores.

IV. Empirical Evaluation

A. NMT models

For our experiment we used OpenNMT[6] [Klein et al., 2017] – a framework for neural sequence learning and neural machine translation. OpenNMT was started by the Harvard NLP group and since its inception it has been used in various research and industry applications. It is distributed for two popular deeplearning frameworks: OpenNMT-py[7] for pytorch[8] and OpenNMT-tf[9] for TensorFlow[10].

In this work we used the OpenNMT-py implementation. We trained two types of models – RNN using LSTM units (abbreviated in this paper as LSTM) and Transformer [Vaswani et al., 2017a] (abbreviated as TRANS). During training intermediate LSTM models are saved every 5000 steps and intermediate TRANS models are saved every 500 steps. These intermediate models are used in our evaluation.

B. Translation Evaluation Shell

A major part of our work includes collecting the evaluation metrics for the translations to check the quality of the same. For this purpose, we developed a shell script to calculate the BLEU-TER and lexical diversity scores for each model generated by running the data corpus on a test data set. The

[6]https://opennmt.net/
[7]https://opennmt.net/OpenNMT-py/
[8]https://pytorch.org/
[9]https://opennmt.net/OpenNMT-tf/
[10]https://www.tensorflow.org/

script essentially takes three arguments as input- (i) the model directory containing the models which were retrieved upon training the word corpus with LSTM and TRANS engines, (ii) a source test file i.e. the text that is to be translated, and (iii) the target test file i.e. text set that is the corresponding translation of the source text. The idea is to loop through the entire model directory for a specific word-pair and get the evaluation scores for each model translations on the test set, and generate the result in a file. There are two python scripts inside the shell- one for calculating the BLEU and TER scores whereas the other calculates the lexical diversity scores namely TTR, MTLD and YULE's I. At each iteration, the script calls a translate shell and executes the same which requires two arguments, the file to be translated and the model name which is being used to translate the text. After the translation is done, the reference file (target file), the translated file and the target language is passed into the BLEU TER python script to get the respective scores via comparative studies. Similarly, the translated file is passed as input to the lexical diversity python script to get the TTR, MTLD and YULE's scores. For each model the scores are appended to two different files and the same is used to get the scores for each model. Once the models are trained, we predicted the translations on a test set and calculated the BLEU-TER and Lexical diversity scores. With the help of the translation evaluation automated shell script, we calculated different metrics like BLEU, TER, MLTD, and YULE's I for the test set (a word corpus of different languages). The script works by taking three arguments- the model directory, the source translation text file and the target translation text file. The script generates two output files. One file contains the automatic evaluation scores namely BLEU and TER. The other file contains the lexical diversity scores i.e. the scores for TTR, MTLD and YULE's I.

V. Experiments

In this paper we have primarily worked with three language pairs, English-Spanish (EN-ES), English-French (EN-FR) and English-Bengali (EN-BN). The neural translation engines for EN-ES and EN-FR were trained beforehand using OpenNMT-py with a training data of approximately 1.5 million. The EN-BN data has been trained with approximately 900,000 parallel data corpus. To train the mentioned language pairs we have built two types of machine translation systems. We have trained attentional RNN [Bahdanau et al., 2015] and Transformer [Vaswani et al., 2017b] MT systems. Post training we draw our conclusions on quality and lexical diversity of the translations with the unseen (test) data.

A. MT Systems

The idea was to train a standard MT system (forward system) for all language pairs. RNN and Transformer systems have been used from OpenNMT-py. Both the systems were trained for a total of 200k steps, where it saves an intermediate model at every 5000 steps for RNN and 500 steps for Transformer. The perplexity of each model was recorded at every step on the development set and the best model was chosen which had the lowest perplexity. Let us discuss in detail the parameter setting for both RNN and Transformer

- RNN: The number of units is set to 512 and the type of the RNN is bidirectional LSTM (thus in the remainder of this paper we refer to all RNN models with LSTM), with 4 layers of encoder and decoder. MLP is the attention type, dropout is set to 0.2; Learning optimizer is Adam [Kingma and Ba, 2015] and learning rate of 0.0001; the batch size is 128.
- Transformer: The number of layers for both encoder and decoder is 6, with size of 512. Other parameters are transformer ff: 2048, number of heads: 8, dropout: 0.1, batch size: 4096, batch type: tokens, learning optimizer Adam with beta 2 equal to 0.998, learning rate of 2.

The training for all the mentioned neural systems is distributed over 4 nVidia 1080Ti GPUs and learning rate decay is enabled for all. The RNN system settings used for our experiments are optimal according to [Britz et al., 2017]. The setting for transformer is the same which has been suggested by OpenNMT community that lead to quality at par with the original transformer work [Vaswani et al., 2017b].

1) Lexical Diversity and Quality Metrics

To test the effectiveness of the machine translation engines, we have collected various evaluation metrics and looked into the same. The lexical diversity scores and the automatic evaluation scores have been used to judge the NMT model performances.

Language diversity score Lexical diversity (LD) refers to the range of different words that are used in a text. The more the range, the higher the diversity [Mccarthy and Jarvis, 2010]. LD can be assessed in different ways and each of them have their specific drawbacks. Therefore, we evaluated LD by using three different widely used metrics: type/token ratio (TTR), Yule's I (Yule, 1944), and the measure of textual lexical diversity (MTLD). One of the most commonly used lexical richness metric is TTR i.e. type token ratio. TTR is defined by the ratio of number of different words in a text to the total number of words . A high/low TTR indicates a high/low degree of lexical diversity. While TTR is one of the most widely used metrics, it does have some drawbacks linked to the assumption of a linear relation between the types and the tokens. Because of which, TTR is only valid when comparing texts of a similar size as it decreases when texts become longer due to repetitions of words [Brezina, 2018].

Yule's characteristic constant, or Yule's K, is a probability model of the changes that take place in the lexical frequency spectrum of a text as the text becomes longer. Yule's K and its reverse Yule's I are better than TTR with respect to the text length. [Oakes and Ji, 2012]. Another metric used to study lexical richness and diversity is MTLD. The difference in these two methods is that MTLD is evaluated sequentially as the mean length of sequential word strings in a text that maintains a given TTR value. Studies have shown that MTLD is most robust with respect to text length.

Recent improvements in MT systems have been largely contributed by BLEU (bilingual evaluation understudy), it is considered as a cheap, reliable evaluation measure for developers to compare their systems [Papineni et al., 2002]. BLEU measures the number of ngrams of varying length of the output of the system that occurs within the set of references and calculates a translation score for the same. It requires a considerable amount of references and sentences in order to correlate human judgments and is relatively counter-intuitive.

Translation error rate or translation edit rate (TER) [Snover et al., 2006] measures the amount of post-editing required after machine translations. It measures the number of edits or actions required by a human translator in order to edit or change an output segment line into one of the references. It is language independent, corresponds with post-editing efforts and is easy to use. Most importantly it focuses on avoiding the knowledge-intensiveness and labor-intensiveness of human judgements on machine translation outputs.

2) EN-ES and EN-FR

For the language pairs English-Spanish and English-French, the neural translation engines were already trained with parallel text from Europarl Corpora. The next part was to calculate the evaluation metrics, as discussed above with the trained models. A total of 1000 records were selected which were Byte Pair Encoded (BPE). BPE segments the parallel corpus to obtain a frequent sequence of combined characters. With the help of our translation evaluation shell that we developed, we passed the model directory containing the models for EN-ES and EN-FR translations one by one along with the BPE source and target files to check how the translation performs on unseen data and recorded the BLEU, TER, TTR, MTLD, YULE score for further analysis.

3) EN-BN

The final EN-BN corpus of 797,592 parallel sentences were used for training our translation engines. MT systems, RNN and Transformer as discussed, were used for training the models from the OpenNMT-py ecosystem. Once the models were trained, we repeated the experiments that were performed for EN-ES and EN-FR. The test set for EN-BN evaluation consisted of two different dataset, both comprising of 1000 parallel texts. One data set was BPE-d and the other test set was used from the SUPara benchmark [Mumin et al., 2018]. Similarly, EN-BN model directory along with the test sets were passed as arguments into translation evaluation shell in order to calculate the evaluation scores.

B. Results and Discussion

Our objective is to assess whether the MT performance changes in the same way for both evaluation metrics and lexical diversity metrics in different stages of training the engines. To do so, we first investigated the linear correlation between BLEU and TER on one hand, and TTR, Yule's I, MTLD on the other hand. We then compared the individual stages of model training and tried finding out the stages of training where these scores were highest (lowest for TTR).

1) Linear correlation between evaluation and lexical diversity metrics

We use Pearson's correlation coefficient *r* to study the linear correlation between two types of metrics i.e. automatic evaluation and lexical diversity. Pearson Correlation coefficient, also known as Pearson's r, is the statistical measure of the linear correlation between two variables X and Y (evaluation metrics and lexical diversity in our case). The values lie in between -1 and +1 signifying positive and negative collinearity respectively. A value close to 0 signifies no/minimal correlation. Pearson's r is the co-variance of two variable divided by their standard deviation. The prime intention of carrying out a Pearson's r test is to check how the automatic evaluation and the lexical diversity evolve at different stages of model training and determine whether the iterative training of NMT models improves both quality, i.e. adequacy and fluency, and lexical diversity in the same way. In addition, linear correlation may be used when it comes to quantifying the degree of separation between systems [Graham et al., 2015].

For our experiment, we are looking at the correlation between evaluation metrics (quality) and lexical diversity metrics. Table II shows Pearson *r* between automatic evaluation metrics and lexical diversity for LSTM models whereas Table III shows the scores for Transformer models, calculated over all training iterations. While higher values in all metrics indicate better performance, it is the other way round for TER – the lower, the better. For ease of readability, we have ignored the negative sign in the Pearson's *r* value for TER. The Pearson *r* between BLEU and TTR, BLEU and MTLD , BLEU and YULE's I is suggestive of the fact that there exists a strong positive correlation between BLEU and lexical diversity. The Pearson *r* between TER and TTR, TER and YULE's I, TER and MTLD also shows strong positive correlation in most of the cases. The lesser the TER for a model, the better it is. And with increase in TTR, YULE's I and MTLD, the TER decreases i.e. the betterment of lexical diversity scores signify a better TER score.

NMT engine	Evaluation metric	Lexical diversity metric		
		TTR	Yule's I	MTLD
EN-FR LSTM	BLEU	0.9909	0.9837	0.8379
	TER*	0.9956	0.9640	0.8881
EN-ES-LSTM	BLEU	0.9938	0.9662	0.9308
	TER*	0.9953	0.9456	0.9422
EN-BN-LSTM	BLEU	0.9947	0.9588	0.9800
	TER*	0.9927	0.9054	0.9758
FR-EN-LSTM	BLEU	0.9963	0.9639	0.9746
	TER*	0.9946	0.9410	0.9916
ES-EN-LSTM	BLEU	0.9941	0.9519	0.9942
	TER*	0.9982	0.9468	0.9970
BN-EN-LSTM	BLEU	0.9525	0.8745	0.7557
	TER*	0.9475	0.8655	0.7094

TABLE II: Pearson's coefficient *r* for evaluation and lexical diversity metrics for the LSTM engines. For ease of comparison, we have ignored the − sign in front of the values for TER, thus we are using the TER* label. Since the lower the TER the better, which is contrary to the other metrics, we can safely ignore the negative sign.

NMT engine	Eval. metric	Lex. div. metric		
		TTR	Yule's I	MTLD
EN-FR-TRANS	BLEU	0.9688	0.9759	0.8362
	TER*	0.8341	0.7686	0.8131
EN-ES-TRANS	BLEU	0.9859	0.9796	0.7486
	TER*	0.9977	0.9553	0.8266
EN-BN-TRANS	BLEU	0.9811	0.9328	0.9577
	TER*	0.9567	0.9567	0.9500
FR-EN-TRANS	BLEU	0.9671	0.9578	0.9307
	TER*	0.9970	0.9216	0.9876
ES-EN-TRANS	BLEU	0.9860	0.9582	0.9629
	TER*	0.9961	0.9086	0.9920
BN-EN-TRANS	BLEU	0.9377	0.8876	0.4889
	TER*	0.9514	0.8837	0.6170

TABLE III: Pearson's coefficient r for evaluation and lexical diversity metrics for the LSTM engines. For ease of comparison, we have ignored the $-$ sign in front of the values for TER, thus we are using the TER* label. Since the lower the TER the better, which is contrary to the other metrics, we can safely ignore the negative sign.

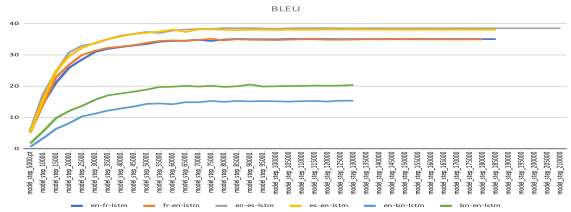
Table II and Table III show strong correlation between quality evaluation and lexical diversity metrics. Further, we notice that BLEU and TER correlate more strongly to TTR than Yule's I and MTLD. For $English \rightarrow French$ and $English \rightarrow Spanish$ LSTM models, it is evident that the correlation between quality evaluation metrics (BLEU and TER) and MTLD is less strong; however, this is not valid for the EN-BN and BN-EN LSTM systems, where the least strong correlation is between the quality metrics and YULE's I. For the TRANS models, overall the correlation between quality and lexical diversity metrics is in the order TTR, Yule's I, MTLD from stronger to weaker.

From our results and their summary in Table II and Table III it is difficult to pinpoint only one quality metric that correlates stronger/weaker to the lexical diversity metrics. For LSTM models, TTR correlates stronger with BLEU in 3 (out of 6) cases and with TER in the other 3 cases. For TRANS models, TTR correlates stronger with BLEU in 4 cases and TER in 2. For both LSTM and TRANS, BLEU correlates better than TER to Yule's I – 6 out of 6 and 5 out of 6 cases, respectively; for MTLD, however, the results show an opposite trend – TER correlates better than BLEU for 4 out of the 6 cases for both LSTM and TRANS models.

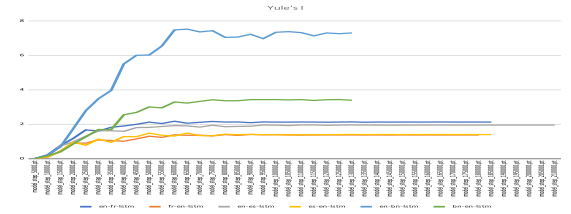
From these results and their analysis we can draw one main conclusion: in order to exploit both quality evaluation metrics and lexical diversity metrics, it is advisable to take into account either BLEU and Yule's I or TER and MTLD – the metrics with the least correlation overall (or a combination of the four).

2) Slope analysis

The graphs in Figure 1 and Figure 2 reflect the trend in the evolution of the automatic evaluation and lexical diversity scores. These are drawn from the value sequences for which we conducted our slope analysis. They illustrate that the scores improve until a certain stage of training the models. Then there comes a stage of training after which the scores do not improve further i.e. a threshold is reached.

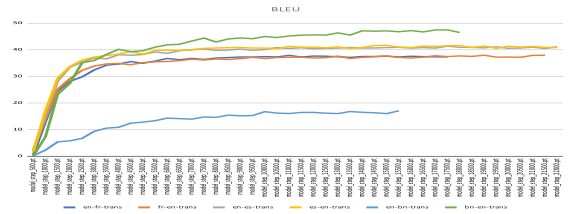


(a) BLEU

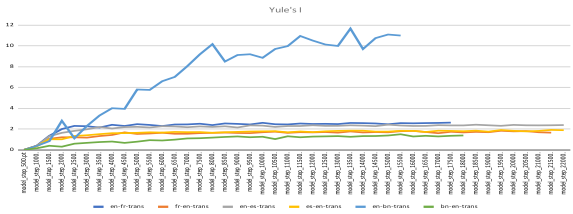


(b) Yule's I

Fig. 1: BLEU and Yule's I score evolution for LSTM models at different stages of training



(a) BLEU



(b) Yule's I

Fig. 2: BLEU and Yule's I score evolution for TRANS models at different stages of training

The above tendencies are clearly confirmed by the slope analysis summarised in Tables IV, V, VI and VII.

To better understand when the highest performance is reached in terms of quality and lexical diversity metrics, we performed a slope analysis on the sequences of values for each metric and the intermediate models. The goal of this analysis is to: (i) identify at what stage of the training process highest values are achieved, i.e. no more significant improvement is possible; (ii) after the point of highest performance whether there is a degradation or there is scope for continuous improvement if the model is allowed to train for

longer time. With respect to (ii) we must note that although we use an early stopping mechanism based on no improvement in 5 consecutive intermediate models, the criteria we use is perplexity and so the quality evaluation and lexical diversity values could potentially continue rising beyond the stopping point of the model training.

For each model and each quality evaluation or lexical diversity metric we employ the following approach:

1) identify the *best* value for each of the sequences – the maximum for BLEU, MTLD, Yule's I, TTR or the minimum value for TER – as well as the point of the training process in which this best value occurs, i.e. the *best point*. In case there are multiple occurrences of the best value, we consider the first one.
2) compute the largest subset of values surrounding the best value with the smallest slope, i.e. identify the widest interval around the best value with the least change – we refer to this interval as *plateau*. To compute the plateau, we are looking for the slope value closest to 0.0.
3) calculate the slope for the sequence of values before the left and right endpoints of the *plateau*. This would indicate whether indeed the plateau is a saturation point prior to and beyond which there are significant changes.
4) for ease of comparison we present the best point, the left and the right endpoints of the plateau interval in percentages of the overall training process. For example, for an NMT system with 20 intermediate models, if the best point (let's say for BLEU) is at intermediate model no. 10 having the left endpoint of the plateau at model no. 8 and the right endpoint at model no. 20, we will represent these as *best at 50% of training, left at 40% and right at 100%*.

We show the best points as well as the plateau in Table IV and Table V for LSTM and TRANS models accordingly. We also show the slope for the sequences before, at and after the plateau in Table VI and Table VII for LSTM and TRANS models. With '–' we indicate that there is no interval for which the slope can be computed as the plateau extends until 100% of the training progress.

The following conclusions can be drawn from our results:

- From the results in Table IV, comparing the left endpoints of the plateaus, we observe that in more cases the plateaus for BLEU (4 out of 6 for TTR, 3 out of 6 for Yule's I and 4 out of 6 for MTLD) and TER (4 out of 6 for TTR, 3 out of 6 for Yule's I and 5 out of 6 for MTLD) is reached at a later stage of the training – the earliest being 37% for BLEU and 40% for TER; for TTR, Yule's I and MTLD the earliest is at 27%, 27% and 16% respectively. If we compare the best points for each metric and model, we notice that more often quality evaluation scores reach their peak at a later stage than the lexical diversity scores for which the peak is achieved much quicker. Analysing the length of the plateau, i.e. the training duration during which values changed the least, shows that BLEU and TER are more unstable than the lexical diversity metrics, indicated by the smaller intervals. These observations indicate that *for LSTM models the optimal lexical diversity scores are reached earlier in the training process than the optimal quality evaluation scores and their values do not change significantly during the remaining of the process*. Despite the above statement, we ought to note that lexical diversity metrics are often scored with very small values, meaning that small fluctuations in the score may indicate a large difference in the lexical diversity of the text. That is why for future work we aim to conduct a thorough human evaluation.
  For TRANS models (see Table V), however, the results are quite different. Firstly, we notice that the best point for the lexical diversity metrics is in general reached at a later stage than for BLEU or TER. The stages at which the best BLEU and TER scores are reached is also later than those found in LSTM models. For some cases (e.g. EN-BN – BLEU, TER, MTLD) the peak is reached at 100% training progress, i.e. at its end. Next, the length of the plateau is significantly low and after that the score usually either remains constant or degrades (the slope is negative, equal to zero or '–'). These results indicate that *for TRANS models, more training is necessary to reach an optimal lexical diversity as well as quality evaluation points in comparison to LSTM.*
- From Table VI we notice that while BLEU scores tend to slightly degrade (negative slope) after the plateau has been reached, TTR and MTLD scores remain mainly stable, i.e. no significant fluctuations are seen. But for Yule's I, 14 out of the 6 models have a negative slope, hinting towards a decline.
  However, in the case of TRANS (see Table V), the results are different and shows more positive *after slopes*. From this we can state that *while LSTM models would reach a stable best point after which their performance may start degrading (in terms of both quality evaluation as well as lexical diversity), TRANS system may potentially benefit from more extensive training with a more relaxed early stopping criteria*.
  It is also obvious from the summarised results that the EN-BN and BN-EN models behave less systematically and the corresponding results are not conclusive. This might be due to lower amounts of data, which may not be sufficient enough to achieve a stable model. These models can be improved further by training with additional parallel data.

3) Ambiguous words

An efficient way to judge the proficiency of NMT systems is to check its mettle in dealing with ambiguous words (a word whose meaning varies depending on context). Therefore, we created a corpus of English sentences containing 92 ambiguous words in such a way that the context is absolutely clear. These sentences were passed through Google Translate[11] and Bing

[11]https://translate.google.com/

Language pair	BLEU				TER				TTR				YULE'S I				MTLD			
	Best	Left	Right	Len.	Best	Left	Right	Len.	Best	Left	Right	Len.	Best	Left	Right	Len.	Best	Left	Right	Len.
EN-FR	45	43	56	13	78	67	78	11	32	27	100	73	32	27	40	13	27	16	70	54
EN-ES	57	38	73	35	54	52	100	48	54	52	88	36	54	52	66	14	52	16	85	69
EN-BN	96	69	96	27	46	42	80	38	50	42	100	58	50	42	76	34	50	38	73	35
FR-EN	41	38	72	34	61	44	61	17	44	41	100	59	50	47	66	19	52	47	91	44
ES-EN	40	37	75	38	43	40	48	8	35	27	59	32	35	27	72	45	27	24	70	46
BN-EN	69	65	92	27	100	80	100	20	80	69	92	23	96	76	100	24	46	34	100	66

TABLE IV: Slope analysis table for the LSTM models showing the best point, the plateau indicated by its left and right endpoints as well as its length (abbreviated as "Len."). Values are in percentage of the training progress of each model.

Language pair	BLEU				TER				TTR				YULE'S I				MTLD			
	Best	Left	Right	Len.	Best	Left	Right	Len.	Best	Left	Right	Len.	Best	Left	Right	Len.	Best	Left	Right	Len.
EN-FR	62	54	100	46	80	65	80	15	80	65	91	26	100	77	100	23	80	45	85	40
EN-ES	79	59	79	20	100	88	100	12	93	84	100	16	68	63	75	12	75	65	100	35
EN-BN	100	70	100	30	100	80	100	20	96	83	96	13	87	70	93	23	100	83	100	17
FR-EN	88	76	100	24	88	86	93	7	90	53	90	37	90	55	90	35	93	58	97	39
ES-EN	68	59	100	41	79	77	84	7	88	59	95	36	97	79	97	18	97	56	97	41
BN-EN	97	77	97	20	77	75	88	13	86	75	100	25	86	80	91	11	86	77	97	20

TABLE V: Slope analysis table for the TRANS models showing the best point, the plateau indicated by its left and right endpoints as well as its length (abbreviated as "Len."). Values are in percentage of the training progress of each model.

Language pair	BLEU			TER			TTR			YULE'S I			MTLD		
	plateau	before	after	plateau	before	after	plateau	before	after	plateau	before	after	plateau	before	after
EN-FR	-0.0015	1.6047	-0.0004	0.0000	-0.7883	0.0001	0.0000	0.0143	–	-0.0002	0.2585	-0.0002	0.0000	0.0886	-0.0001
EN-ES	0.0000	1.6770	-0.0007	0.0002	-1.1029	–	0.0000	0.0045	0.0000	0.0000	0.0759	0.0001	0.0000	0.0627	0.0000
EN-BN	-0.0027	0.7819	-0.0292	0.0079	-3.7861	-0.0343	0.0001	0.0246	–	-0.0047	0.7650	-0.0134	0.0006	0.3487	0.0043
FR-EN	0.0008	1.9288	-0.0001	0.0000	-1.9704	0.0014	0.0000	0.0078	–	-0.0002	0.0824	-0.0001	0.0000	0.0468	-0.0003
ES-EN	0.0002	2.1450	-0.0001	0.0000	-2.3156	-0.0004	0.0000	0.0145	0.0000	0.0001	0.1647	-0.0002	0.0000	0.1226	0.0002
BN-EN	0.0003	1.0473	0.1207	0.0000	-1.2877	–	0.0000	0.0104	-0.0003	-0.0005	0.2045	–	-0.0001	0.1594	–

TABLE VI: Slope values for the sequences plateau, before and after for the LSTM models

Language pair	BLEU			TER			TTR			YULE'S I			MTLD		
	plateau	before	after	plateau	before	after	plateau	before	after	plateau	before	after	plateau	before	after
EN-FR	-0.0004	1.3848	–	0.0000	-1.3885	0.0345	0.0000	0.0033	0.0008	0.0024	0.0544	–	0.0000	0.0369	0.0010
EN-ES	0.0003	0.8014	-0.0339	0.0000	-0.4127	–	0.0000	0.0016	–	0.0000	0.0499	0.0028	0.0000	-0.0163	–
EN-BN	0.0039	0.7066	–	-0.0257	-1.7717	–	0.0003	0.0089	-0.0088	0.0021	0.5332	0.1232	0.0006	0.1176	–
FR-EN	0.0020	0.4873	–	0.0000	-0.4647	-0.1000	0.0000	0.0031	-0.0014	0.0005	0.0484	-0.0451	0.0000	0.0156	-0.0249
ES-EN	-0.0001	0.7831	–	0.0000	-0.5969	0.0238	0.0000	0.0026	0.0007	0.0002	0.0286	-0.0262	0.0000	0.0171	-0.0500
BN-EN	0.0176	1.0882	-0.9681	-0.01	-1.7616	-0.2300	0.0001	0.0043	–	-0.0013	0.0422	0.0157	-0.0004	0.0344	0.0773

TABLE VII: Slope values for the sequences plateau, before and after for the TRANS models

Microsoft Translator[12] to convert them into Bengali and judge how these state-of-the-art systems deal with the ambiguous words. The Google translator could not correctly deal with 28 of those 92 ambiguous words whereas the Bing Microsoft Translator was unable to translate 22 of these properly. The results clearly suggested that even the most advanced NMT systems face difficulty in dealing with these words and they are in fact quite far off from the human translated texts.

Next, we tested our NMT systems with these ambiguous words to judge how far off our NMT models are from the advanced NMT engines like Google translator and Bing Microsoft Translator but also to see whether using models that have reached highest quality evaluation scores or highest lexical diversity scores is better. We translated the set of sentences and compared the translations to our human translations.[13] We only tested models with highest quality and lexical diversity scores obtained in our experiment. Table VIII shows the quality metrics; Table IX shows the lexical diversity metrics for the sentences containing those ambiguous words.

From Table VIII, we noticed that the translation quality of our NMT models is quite low. The primary reason for this low performance is the unavailability of enough parallel data required to train good NMT models. Having said that, the quality judged by the evaluation metrics on our test set is high enough because of the closed-domain data even in case of general sentences. In the future, we aim to exploit Zero Shot Translation and use Transliteration to overcome the data sparsity issue for this language pair.

Although the quality and lexical diversity scores for the translations of the ambiguous words were below average,

[12]https://www.bing.com/translator

[13]Human translations were generated by native Bengali speakers proficient in English.

NMT engine	Best Models	BLEU	TER
EN-BN-LSTM	model step 60000	10.61	70.30
	model step 65000	11.01	72.90
	model step 90000	11.78	73.00
	model step 125000	11.46	72.70
	model step 130000	11.52	72.80
EN-BN-TRANS	model step 13500	8.26	74.00
	model step 15000	5.70	75.80
	model step 15500	6.06	76.60
BN-EN-LSTM	model step 70000	13.27	63.10
	model step 80000	12.59	64.20
	model step 85000	13.39	62.90
	model step 110000	13.03	63.00
BN-EN-TRANS	model step 14000	15.80	61.6
	model step 15500	16.85	59.8
	model step 17000	15.95	60.0
	model step 17500	16.43	60.6

TABLE VIII: Quality evaluation scores for translations of sentences with ambiguous words for the EN-BN and BN-EN models with the best scores.

NMT engine	Best Models	MTLD	TTR	Yule's I
EN-BN-LSTM	model step 60000	15.4367	0.3984	12.7772
	model step 65000	15.1564	0.4090	13.5705
	model step 90000	15.2218	0.4137	14.0049
	model step 125000	15.3831	0.4133	14.2187
	model step 130000	15.4850	0.4134	13.9966
EN-BN-TRANS	model step 13500	15.5612	0.5059	23.1949
	model step 15000	15.7306	0.5074	22.6752
	model step 15500	15.8851	0.5031	23.6544
BN-EN-LSTM	model step 70000	13.1461	0.3320	5.7997
	model step 80000	13.2417	0.3402	6.3841
	model step 85000	13.2618	0.3343	5.8713
	model step 110000	13.2561	0.3378	6.1473
BN-EN-TRANS	model step 14000	13.6408	0.3908	8.6587
	model step 15500	13.5404	0.3816	8.4468
	model step 17000	13.3341	0.3742	8.3568
	model step 17500	13.3861	0.3860	8.8490

TABLE IX: Lexical diversity scores for translations of sentences with ambiguous words for the EN-BN and BN-EN models with the best scores.

it was interesting to see that few translations were correct and the translated text could differentiate the context of the ambiguous words as intended. The quality of those sentences that were translated correctly matched with advanced engines like Google translator and Bing Microsoft Translator. This allows us to make a conclusion that if the MT engines are trained with more data then the ambiguous translations ought to give better results.

VI. Conclusions and Future Work

The work presented in the paper focused on the correlation between automatic evaluation and lexical diversity metrics. We used three language pairs and trained NMT models in six possible directions. We assessed both LSTM and Transformer architectures in this paper. Our experiments traced the scores for BLEU, TER and for TTR, Yule's I and MTLD at intermediate training stages, i.e. for models saved at each 5000 training steps for LSTM and 500 for Transformer. We focused on (i) assessing the linear correlation between automatic evaluation and lexical diversity metrics using Pearson's correlation coefficient *r*, (ii) analysis of the slope of the sequences of metric values and (iii) analysis of translations of ambiguous words. The presented results showed that there exists strong correlation between the automatic evaluation metrics and TTR, but not so strong between BLEU and Yule's I and TER and MTLD. This hints of exploiting benefits of both metrics to device a more informative metric. Further work on more language pairs will provide support to our findings. We noticed that scores surged and improved up to a certain stage of training and then they often degraded or remained unchanged. For Transformer it seems that more training would be beneficial as indicated by the slope analysis.

A general understanding in terms of machine translation is that the quality of translations of an NMT system is directly proportional to the amount of training data. Although, ample amount of parallel open data is available for some language pairs, that is not the case for all, such as low-resource languages like Bengali or Hindi. Our analysis of the ambiguous words showed that NMT models that we trained with freely available data are not generalised. We also showed that context is important and in some cases, given the appropriate context, even our restricted NMT systems can deal correctly with such ambiguous words. However, we also noticed that even powerful MT giants such as Google translate and Bing Microsoft Translator can falter if the contexts are vague, indicating that machine translationese has still a long way to go to reach human quality.

To mitigate the data sparsity issue and improve the translation systems for Bengali, in the future we will work towards new data generation using TRANSLIT [Benites et al., 2020] to generate synthetic data and complement other (synthetic or authentic) data. We will also look into Zero shot translation (ZST) [Johnson et al., 2016] which has been found to be quite successful in dealing with low resource languages.

During this work we were faced with a cumbersome process of training, evaluating and analysing the results. Keeping this in mind a part of our future work is to develop a user-friendly interface wherein one can view the automatic evaluation scores and lexical diversity scores for each intermediate model and choose one of their liking as the default translation model. The user will also be able to specify their preference prior to training – higher BLEU/TER or higher lexical diversity. We have already developed a beta version of the interface and the same will be incorporated soon.

References

[Agic and Vulic, 2019a] Agic, Z. and Vulic, I. (2019a). JW300: A wide-coverage parallel corpus for low-resource languages. In Korhonen, A., Traum, D. R., and Màrquez, L., editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3204–3210. Association for Computational Linguistics.

[Agic and Vulic, 2019b] Agic, Z. and Vulic, I. (2019b). JW300: A wide-coverage parallel corpus for low-resource languages. In Korhonen, A.,

Traum, D. R., and Màrquez, L., editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3204–3210. Association for Computational Linguistics.

[Bahdanau et al., 2014] Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

[Bahdanau et al., 2015] Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In Bengio, Y. and LeCun, Y., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

[Benites et al., 2020] Benites, F., Duivesteijn, G. F., von Däniken, P., and Cieliebak, M. (2020). TRANSLIT: A large-scale name transliteration resource. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 3265–3271. European Language Resources Association.

[Berman, 2000] Berman, A. (2000). Translation and the trials of the foreign.

[Brezina, 2018] Brezina, V. (2018). *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge University Press.

[Britz et al., 2017] Britz, D., Goldie, A., Luong, M., and Le, Q. V. (2017). Massive exploration of neural machine translation architectures. *CoRR*, abs/1703.03906.

[Callison-Burch et al., 2011] Callison-Burch, C., Koehn, P., Monz, C., and Zaidan, O. (2011). Findings of the 2011 workshop on statistical machine translation. In Callison-Burch, C., Koehn, P., Monz, C., and Zaidan, O., editors, *Proceedings of the Sixth Workshop on Statistical Machine Translation, WMT@EMNLP 2011, Edinburgh, Scotland, UK, July 30-31, 2011*, pages 22–64. Association for Computational Linguistics.

[Forcada, 2017] Forcada, M. L. (2017). Making sense of neural machine translation. *Translation Spaces*, 6(2):291–309.

[Graham et al., 2015] Graham, Y., Baldwin, T., and Mathur, N. (2015). Accurate evaluation of segment-level machine translation metrics. In Mihalcea, R., Chai, J. Y., and Sarkar, A., editors, *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, pages 1183–1191. The Association for Computational Linguistics.

[Johnson et al., 2016] Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F. B., Wattenberg, M., Corrado, G., Hughes, M., and Dean, J. (2016). Google's multilingual neural machine translation system: Enabling zero-shot translation. *CoRR*, abs/1611.04558.

[Kingma and Ba, 2015] Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In Bengio, Y. and LeCun, Y., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

[Klein et al., 2017] Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. (2017). OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

[Koehn, 2005] Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

[Koppel and Ordan, 2011] Koppel, M. and Ordan, N. (2011). Translationese and its dialects. In Lin, D., Matsumoto, Y., and Mihalcea, R., editors, *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA*, pages 1318–1326. The Association for Computer Linguistics.

[Kotze, 2012] Kotze, H. (2012). A corpus-based study of the mediation effect in translated and edited language. *Target*, 24:355–388.

[Lison and Tiedemann, 2016] Lison, P. and Tiedemann, J. (2016). OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

[Mccarthy and Jarvis, 2010] Mccarthy, P. and Jarvis, S. (2010). Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42:381–92.

[McCarthy, 2005] McCarthy, P. M. (2005). An Assessment of the Range and Usefulness of Lexical Diversity Measures and the Potential of the Measure of Textual, Lexical Diversity (MTLD). In *PhD Thesis, Dissertation Abstracts International, Volume 66:12*. University of Memphis, Memphis, Tennessee, USA.

[Mumin et al., 2018] Mumin, M. A. A., Seddiqui, M. H., Iqbal, M. Z., and Islam, M. J. (2018). SUPara-Benchmark: A Benchmark Dataset for English-Bangla Machine Translation.

[Oakes and Ji, 2012] Oakes, M. and Ji, M. (2012). *Quantitative Methods in Corpus-Based Translation Studies: A practical guide to descriptive translation research*.

[Papineni et al., 2002] Papineni, K., Roukos, S., Ward, T., and Zhu, W. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

[Sennrich et al., 2016] Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

[Shterionov et al., 2018] Shterionov, D., Superbo, R., Nagle, P., Casanellas, L., O'dowd, T., and Way, A. (2018). Human versus automatic quality evaluation of nmt and pbsmt. *Machine Translation*, 32(3):217–235.

[Snover et al., 2006] Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *In Proceedings of Association for Machine Translation in the Americas*, pages 223–231.

[Templin, 1975] Templin, M. C. (1975). *Certain Language Skills in Children: Their Development and Interrelationships*. Greenwood Press, Westport, Connecticut, USA.

[Tiedemann, 2012] Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In Chair), N. C. C., Choukri, K., Declerck, T., Dogan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

[Tomás et al., 2003] Tomás, J., Mas, J. À., and Casacuberta, F. (2003). A quantitative method for machine translation evaluation. In *Proceedings of the EACL 2003 Workshop on Evaluation Initiatives in Natural Language Processing: are evaluation methods, metrics and resources reusable?*, pages 27–34.

[Toral, 2019] Toral, A. (2019). Post-editese: an exacerbated translationese. In Forcada, M. L., Way, A., Haddow, B., and Sennrich, R., editors, *Proceedings of Machine Translation Summit XVII Volume 1: Research Track, MTSummit 2019, Dublin, Ireland, August 19-23, 2019*, pages 273–281. European Association for Machine Translation.

[Vanmassenhove et al., 2019] Vanmassenhove, E., Shterionov, D., and Way, A. (2019). Lost in translation: Loss and decay of linguistic richness in machine translation. In Forcada, M. L., Way, A., Haddow, B., and Sennrich, R., editors, *Proceedings of Machine Translation Summit XVII Volume 1: Research Track, MTSummit 2019, Dublin, Ireland, August 19-23, 2019*, pages 222–232. European Association for Machine Translation.

[Varga et al., 2007] Varga, D., Halacsy, P., Kornai, A., Nagy, V., Nemeth, L., and Tron, V. (2007). Parallel corpora for medium density languages. In Nicolov, N., Bontcheva, K., Angelova, G., and Mitkov, R., editors, *Recent Advances in Natural Language Processing IV. Selected papers from RANLP-05*, pages 247–258. Benjamins, Amsterdam.

[Vaswani et al., 2017a] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017a). Attention is all you need. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008.

[Vaswani et al., 2017b] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017b). Attention is all you need. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008.

[Yule, 1944] Yule, G. U. (1944). *The Statistical Study of Literary Vocabulary*. Cambridge University Press, Cambridge, USA.

Division of Work:

Literature review and establishing research questions:

A total of twelve papers have been read from our end for practicum and made a habit of writing summary for the papers, thanks to our supervisor for the same. We divided the work of reading papers and writing the summary amongst ourselves. Once the summary was written for each paper one did a peer review of the other's work and then send it to our supervisor. This process helped us understand what the other person was reading and educate each self on unknown facts related to our final deliverable. Once all possible research areas were listed each would then question the other on why a given topic should be our area of research, this exercise helped us omit the redundant questions listed by both and narrowed down our research areas.

Data collection and data pre-processing:

We have used JW300 and Europarl corpora parallel texts in our study, apart from these we also collected data from other sources in the web which were done parallelly by both of us. Along with this we also created a list of ambiguous sentences for English to Bengali translations which were to be tested on the neural translation engines we developed to check how our systems fare in translating ambiguous text. Souradip worked on collecting the ambiguous sentence list and did all pre-processing for the list, on the other hand Aritra collected the data from the JW300 corpus and did the required pre-processing.

Shell script to calculate the translation scores:

In order to evaluate the results of our translations we developed this script to calculate the quality and lexical diversity scores for the same. Aritra developed the shell script as per the requirement and did the initial round of testing to check whether the script is working or not. Souradip took the work from there test it on actual data and modified the same as and when required.

Training the Neural Machine Translation engines:

A total of eight times the NMT engines were trained for the English - Bengali language pair. The NMT engines were trained using two MT systems (TRANS and LSTM), the training for English to Bengali translations was done by Aritra and for that of Bengali to English was done by Souradip. We had to train the engines twice as after the first training we realized that the data needs to be further pre-processed to get better translations and then again the process of training the engines was repeated with the new pre-processed data.

Writing the paper:

We started writing the paper long back whenever we completed a particular task, we had a shared document on overleaf with our supervisor which allow us to work parallelly on the paper. Souradip started by writing the abstract and the related works section. Aritra started by writing about the dataset and the empirical setup section. Later we both simultaneously wrote about the experiments that were done for our work. We individually carried out experiments on the language pairs after diving the work between us and inserted the graphs and tables for the same. The results of our respective experiment were properly analysed by each of us and shared with our supervisor. Once the results were accumulated and the results we intend to show was finalized, Aritra and Souradip inserted the respective tables and graphs retrieved from the experiment and explained the results.