Machine Translation in Natural Language Processing by Implementing Artificial Neural Network Modelling Techniques: An Analysis

Fazeel Ahmed Khan, Adamu Abubakar Ibrahim
Dept of Information and Communication Technology, International Islamic University Malaysia, Kuala Lumpur, Malaysia
fazeelahmedkhan15@gmail.com
Dept of Information and Communication Technology, International Islamic University Malaysia, Kuala Lumpur, Malaysia
adamu@iium.edu.my

***Abstract*— Natural Language Processing is emerging with more efficient algorithms to perform detailed analysis and synthesis on different languages and speech translation with techniques from computer science. Machine translation is emerging from Statistical Machine Translation to a more efficient and robust oriented deep learning based Neural Machine Translation. The limitation in Statistical based MT opens a new spectrum of research in NMT to resolve the existing problems and explore NMT potential in MT research. This paper comprehensively analyses various NMT models proposed in recent years and their contribution in resolving language translation issues. It also discusses on some NMT based open-source toolkits introduced in recent year and the feature implemented in these toolkits. It also analyses the potential of these toolkits to comply with research in language translation particularly in NMT based techniques.**

Keywords— Statistical Machine Translation, Neural Machine Translation, Natural Language Processing, Artificial Neural Network, Machine Translation.

I. INTRODUCTION

In Natural Language Processing, the Machine Translation (MT) focuses on translation of textual data from one language to another by using some MT methods. Traditionally, the MT was performed by using various Statistical models with predictive algorithms to teach computers how to translate text from one language to another. These models are created from parallel bilingual corpora to create probable output based on different examples. With translated text output, the statistical model predicts how to translate foreign language text. The benefits of this method are automation in language translation systems, but it has some drawbacks mostly related to MT methods based on corpora translations to develop its own textual segments. The advancement in machine learning with especial effect to deep learning algorithms, the Neural Machine Translation (NMT) has arisen as a saviour to language translation system.

The NMT introduces state-of-the art algorithms in which massive amount of dataset with translated sentences have the capability to translate sentence between any two language quickly and effectively. The algorithms are based on human brain models with information transferred between multiple layers of processing before an output can be predicted. The NMT is designed on neural based structures making connection with each other's, learning new information and can access the comprehensive structure of these sentences instead of piecewise strategy. NMT doesn't directly predict the outcome but follows the two-step procedure of encoding and decoding. In encoding, each word from a source language is transformed into a vector which is inputted into the model [1]. The decoder then transformed the vector input to the target translated language as shown in Fig. 1.

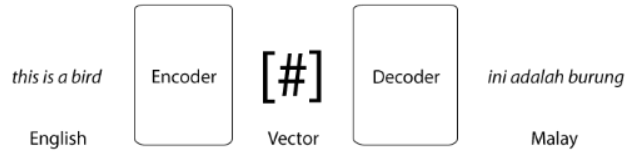


Fig. 1 Encoder and Decoder Method

This paper is organized with brief introductory section elaborating the idea of MT and NMT in computational linguistics. The second section comprise of brief overview on Artificial Neural Network and its various model functions and features respectively. The third section focuses on neural network recent studies focusing on compendious

analysis on recently proposed NMT models. Similarly, the following chapter explained the recently introduced related works (tools or platforms) in the application of NMT in MT. Furthermore, the conclusion will explain the nexus among various models as stated in this research.

II. **Artificial Neural Networks**

1. *Neural Networks*

Neural Network is a machine learning technique which draws a set of input from a relevant source and predicts its outcome by analyzing from training examples. A neural network comprises of processing nodes in large numbers which are connected densely and organized in layers with each other. An individual node can be connected to multiple nodes in bottom layer to receive data and multiple nodes on above layers to sends the data. [2]

a. Linear Model

Linear models are an important part of statistical machine translation having potential to translate a single sentence with certain set of features. Each feature is weighted by some parameter to obtain an overall score by ignoring exponential function to turn the linear model into log-linear model. The graphical representation of linear neural model is shown in Fig.2.

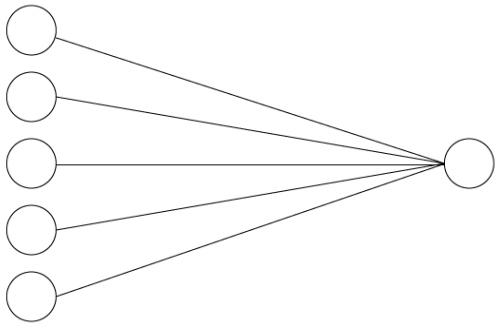


Fig. 2 Graphical representation for Linear Neural Model

b. Hidden Layers

The linear model was multiplied in two aspects in neural networks. Firstly, the use of multiple layers in contrast with obtaining output values directly from input source which introduces a hidden layer's mechanism in this domain. A linear combination of weighted input is computed to obtain hidden node values for each node. Secondly, the linear combination of weighted hidden nodes computed to obtain output values for each node. The concept of hidden layers is very similar to Markov models in which input and output were observed during training instances but not as a method to connect them. The graphical representation of neural network with hidden layers is shown in Fig.3.

c. Back-Propagation Model

Neural networks require the refinement of weighted values to develop a network for predicting correct output from a dataset. It is a cyclic process in which input is constantly feed to the network, compared with the computed output of the network with correct output from the training dataset example. There will be multiple cycles carried out in this process, each process carried out by passing on data is called as an *epoch*. The training method applies in neural network is commonly called as *back-propagation*. The process starts firstly from updating weights to output layers, propagates to find any back-error information to earlier layers. When the training process reaches to an end, each node in the network error term is computed for updating value for incoming weights.

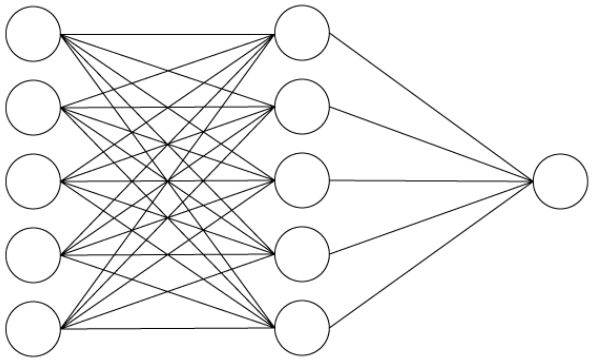


Fig. 3 Neural Network with Hidden Layers

The formula to compute the updated values for each weight applies with *gradient descent training* principle. To reduce the error of given function, the gradient of the error function was computed along with each weight and move against the gradient to overcome the errors. The graphical representation of backpropagation based gradient descent training is shown in Fig. 4.

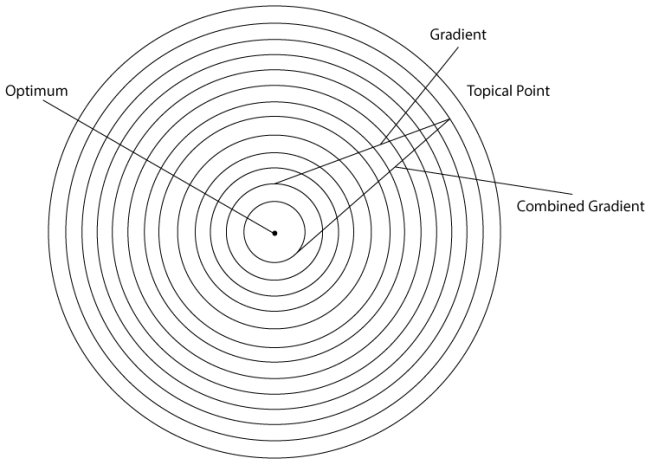


Fig. 4 Back-Propagating based Gradient Descent Training

III. Neural Network for Machine Translation

1. Multi-source NMT

Multi-source NMT is used to check multiple inputs with variety of languages to enhance the translation efficiency and accuracy. The study by [3] examine the NMT using incomplete multilingual corpus which have incomplete or missing translations. This research focuses on use of incomplete multilingual corpora based on multi-encoder NMT and mixture of experts based on NMT. It will be examined for an easy implementation in which source translation are replaced by special symbol. It was used to allow the incomplete corpora both at training and test period.

The experimental result shows that the simulated and incomplete multilingual corpora allow to accurately induct most available translation at same time. The performance of multi-source NMT rely on source and target languages. Similarly, it also depends on missing data and achieved greater translation accuracies measured with BLEU by one-on one NMT system.

2. Neural Summarization

Summarization condenses documents into short paragraph while maintaining core information. The study by [4] presents a neural summarization models with effective mechanism to allow the users to specify high level attributes to control the shape of concluding summaries to better comply with required needs. The works focuses on abstractive summarization to control the important aspects of generated summary.

3. Neural Automatic Post-Edition System

The study by [5] focuses on the interoperability of predictions develop by neural automatic post-edition (APE) system for correcting systematic errors generated by machine translators. The model is proposed to assemble contextual neural automatic post-edition systems that encodes source and machine translated sentences with two separate encoders that shared single and joined attention mechanism, having leverage on shared attention for effective learning on two inputs points. It will contribute to generation of post-edited sentences. [6]. The shared attention supports a key feature for the identification of selection shifts of either on source or machine translated inputs at every decoding step.

The proposed model has been trained and evaluated as shown in Fig. 5 with data gathered from WMT16 and WMT17 APE Information Technology domain English-German with an extra 500K of artificial data by applying round-trip translation for shared tasks. [7]. It also shows easy interpretability and competitive accuracy with knowing about each input derivative related to its prediction as shown in Table 1.

Table I RESULT ON WMT 17 INFORMATION TECHNOLOGY DOMAIN ENGLISH-GERMAN APE TEST SET (Source:[1])

Model	TER	BLEU
MT	24.48	62.49
SPE	24.69	62.97
Train 11K	41.58	43.05
Train 23K	30.23	57.14
Train 23K + 500K	22.60	66.21

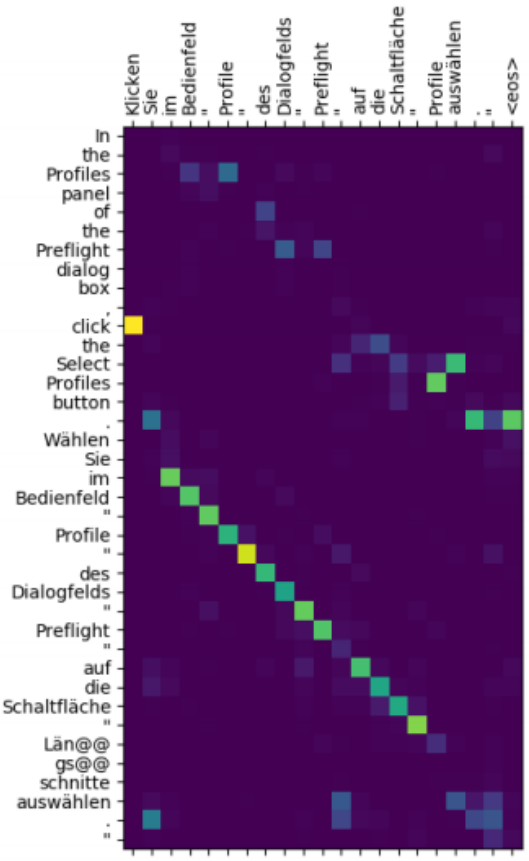


Fig. 5 An example of ideal correction for Machine Translation sentence (Source: [5])

4. Iterative Back-Translation

The study by [8] proposed an idea of implementing back-translation to build a better system with back-translated data which can be performed in a cyclic process since it doesn't allow to stop at single iteration of repeated back-translation. It can be iterated to multiple back-translation system numerous times referring it to an iterative back-translation. The proposed model validates this approach with high and low resource condition with sophisticated re-back translation as shown in Table 2 and Table 3.

Table II. WMT GERMAN-ENGLISH QUALITY COMPARISON TASK WITH DIFFERENT FINAL SYSTEM (Source: [5])

German-English	Back	Shallow	Deep	Ensemble
Back-translation	23.7	32.5	35.0	35.6
Re-back-translation	27.9	33.6	36.1	36.5
Best WMT 2017	-	-	-	35.1

Table III WMT ENGLISH-GERMAN QUALITY COMPARISON TASK WITH DIFFERENT FINAL SYSTEM (Source: [5])

English-German	Back	Shallow	Deep	Ensemble
Back-translation	29.1	25.9	28.3	28.8
Re-back-translation	34.8	27.0	29.0	29.3
Best WMT 2017	-	-	-	28.3

5. Noise impact on NMT

The study by [9] perform the analysis the various types of noise in parallel training data that impact the quality of NMT systems. The artificial noises were created to critically analyze the impact of these noises on the degrading performance in Statistical MT and NMT systems. It was found that neural model has high potential of affecting with these noises as compared with statistical models in particular with egregious noise which learn easily to duplicates the input sentences. The result have shown that NMT is less robust as compared with other Statistical MT's.

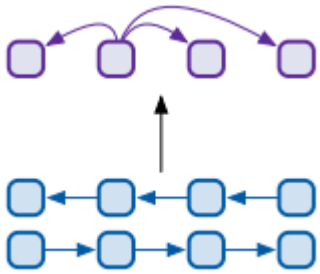
6. Bi-Directional NMT

The study by [10] propose a mechanism which combines back-translation and multilingual NMT to enhance efficiency in low-resource and out-of-domain scenarios of phrase based translation in MT. The mechanism trains a single model for multiple directions in pair of language making it easier to back-translate source and target monolingual data without entailing any auxiliary model. Furthermore, the training will be continuing augmented parallel data, authorizing a cyclic improvement for single model which can integrate any source and target as well as parallel data to improve both translation directions.

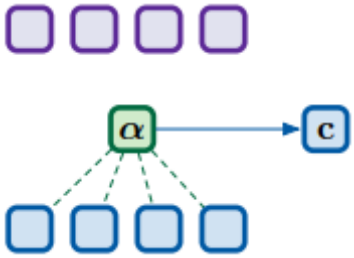
The model gives significantly reduce in training and deployment costs in contrast with uni-directional models. The experiment shows that this mechanism accomplishes standard back-translation in low-resource scenarios, enhance quality on cross-domain and enormously reduce the costs covering on board.

7. Syntactic Attention Model

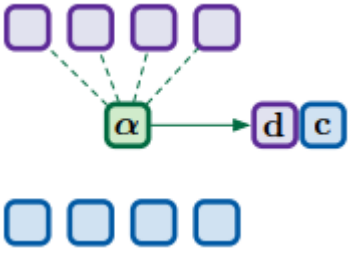
The study by [11] proposed a model based on syntactic attention as opposed to neural machine translation in two dimensions as shown in Fig. 6. Firstly, the encoder has two sets of annotation in output: content annotations based on standard BiLSTM (Bi-directional Long Short-Term Memory) and syntactic annotations based on Head Word selection layers. The syntactic annotation capture dependencies between the source words and supports transfer from source to target. The source dimension syntax transforms the standard attention from target to source in Neural Machine Translation (NMT) applies into both content and syntactic through shared attention layer as shown in Fig. 6.



a) Assembled Self Attention Encoder



b) Context Vector Computation using NMT model



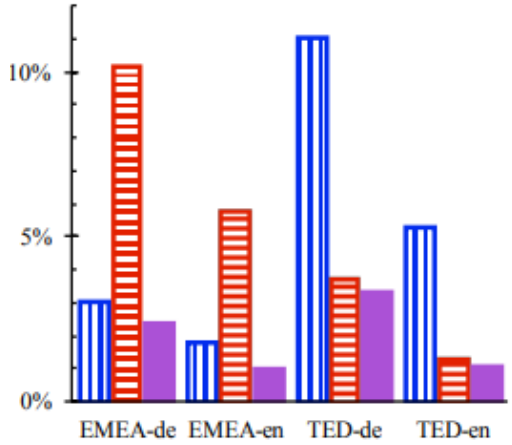
c) Calculate Syntactic Vector

Fig. 6 Proposed Shared Attention Model (Source: [11])

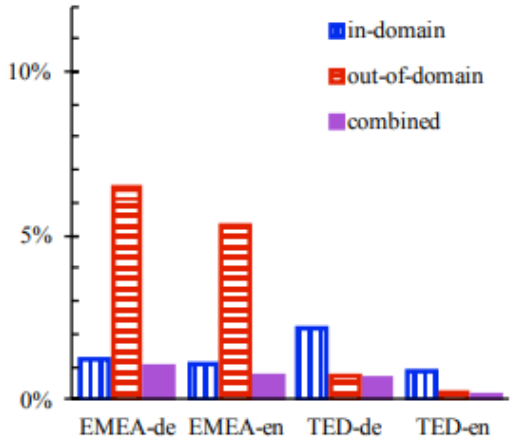
8. Supervised Domain Adaption

The study by [12] demonstrates the domain adaption in which there is not a compulsory requirement to have large datasets in order to apply NMT system in appropriate language pair. The proposed model applies the output

distribution of trained out-of-domain model to regularize the training of in-domain model as an adaption to a new domain. [13] The NMT system were trained using modified OpenNMT-py [2]. It is built on RNN (Recurrent Neural Network) – based encoders and decoders with bidirectional RNN. Both encoders and decoders were fitted with two-layer LSTM hidden size of 1024 [14]. The experiment shows wider improvements on the EMEA as compared to TED observations as shown in Fig. 7.



a) Out-of-Vocabulary (OVV) rates by type



b) Out-of-Vocabulary (OOV) rates by tokens

Fig. 7 Result Analysis on OOV by types and tokens (Source: **[12]**)

9. Dual Learning in NMT

The study by [15] develops a dual-learning mechanism which utilize the NMT system to automatically learn from unlabeled data. It is uplift from dual task mechanism from any MT with a pair of single language (English-to-French) translation to another language pair (French-to-English) translation. It can accomplish this task from closed loops and develops information feedback ladders for training translation models without any interference from any human labeler. The experiment on proposed structure shows that dual-NMT works most better on English-to-French translation more focus by learning from monolingual data with 10% bilingual data for better start. It finally achieves comparable accuracy to NMT trained from full bilingual data for French-to-English translation.

10. Target Monolingual Corpora

The study by [16] extended the methods proposed by [17] to enhance the encoder and attention using target monolingual corpora. The proposed model generates multiple sentences from various sources by sampling each target sentence in a translation back order. The multiple sources achieve the average of errors in individual synthetic sentences and ensure diversity in human translations which will reduce their harmful effects in countermeasure against machine translated sentences having less variety. The graphical representation of model is shown in Fig. 8.

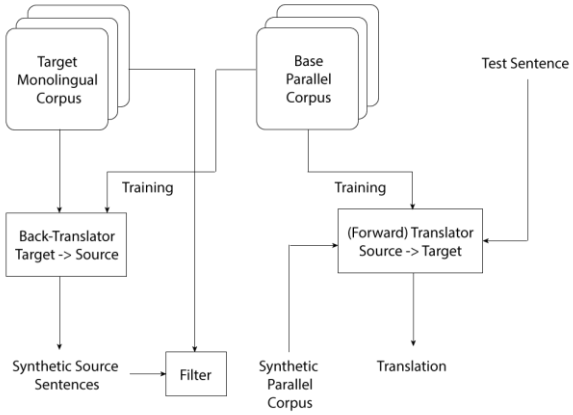


Fig 8. Approach Model (Source: **[16]**)

11. Document-Level Adaption

The study by [18] proposed two complementary approaches to treat with adaption in machine translation system. Firstly, the Single-Sentence Adaption which iteratively adapted over the previous translated sentences along with its references and the model is updated to use with next sentence translation. Secondly, the Dictionary Training approach aims to adapt models with particular goals of effective translation of novel words. With the given words, the approach used to identify words that are novel having not appeared in any training or adaption data before and the outcome will be to obtain a single translation for each of these words [19]. The comprehensive analysis on result are shown in Table. 3.

12. Rare Words Translation in NMT

Since NMT works with fixed vocabulary, because of this limitation it lacks the feature to translate with open-vocabulary [20]. The study by [21] introduces a simple and effective approach to enhance the capacity of NMT model in open-vocabulary translation by encoding rare and unknown words in sequences of their subword. The model is based on intuition of word classes into various translatable segments with smaller units. It discusses the acceptable word

segmentation technique into multiple units with simple characters *n-gram* models. Similarly, it will also include *byte-pair* encoding compression algorithm and shows the improvement in subword dictionary models for baseline for WMT 15 translation task e.g. English-German or English-Russian with BLEU up to 1.1 and 1.3 respectively.

Table IV RESULT ANALYSIS OF BASELINE AND DICTIONARY TRAINING WITH FULL WMT 17 TEST DOCUMENTS (Source: **[18]**)

Model	BLEU	Nov. Acc.
WMT Baseline	25.1	48.9%
Single Sen. Adapt.	26.7	58.4%
Lex. Const. Decoding	25.0	76.9%
Dictionary Training	25.1	71.7%
Dict. + Single-Sent.	26.9	72.2%

13. Semi-Supervised Learning for NMT

The semi-supervised learning for NMT proposed a semi-supervised approach for jointly train the NMT models on labelled parallel corpora and unlabelled monolingual corpora datasets. The study by [22] focuses on introducing the remodelling techniques for observed monolingual corpora by applying autoencoders where the source-to-target and target-to-source translation layers serves in terms of encoders and decoders respectively. The study also introduces full search space method for improving the efficiency of the model. Similarly, the proposed model doesn't exclusively depend on the network architecture and can be implemented with any arbitrary end-to-end NMT system. Furthermore, the model also applies to both the source and target monolingual corpora as opposed to conventional way of focusing on target monolingual corpora only.

The experiment model is evaluated on Chinese-English dataset which include both the parallel corpus and two monolingual corpora as training set. The NIST 2006 dataset is used for hyper-parameter optimization and model selection in terms of validation for combine datasets. The NIST 2002, 2003, 2004 and 2005 datasets serve as the test set for the experiment model. The approach is being compared with two Statistical Machine Translation (SMT) and NMT system i.e. *MOSES: Phrase-based SMT* [23] and *RNNSEARCH: Attention-based NMT systems* [24] respectively. For MOSES, the experiment uses default configuration to train the phrase-based translation on parallel corpus and log-linear models as minimum rate error algorithm for training as an optimization parameter. It also used SRILM toolkit for training 4-gram model. Similarly, the RNNSEARCH uses parallel corpus to train the attention-based NMT having configured the vocabulary size of word embedding to 30K for both Chinese and English. There is a bidirectional attention based NMT on top of RNNSearch on the merge of both parallel and monolingual corpora respectively.

The effect on sample size shows that by increasing the sample size, the approximate search space improved the BLEU scores. However, increasing sample size does not bring major improvements and decrease the efficiency of the training set. Similarly, the effect on Out-of-Vocabulary (OOV) ratio shows that by using monolingual corpus brings with a lower OOV ratio and higher BLEU scores. Furthermore, the comparison with SMT and NMT summarizes the results by showing that applying the target monolingual corpus improves translation for source-to-target end-to-end system, applying source monolingual corpus improves source-to-target translation which will be further improve by adding target monolingual corpus and lastly, applying source and target monolingual corpora doesn't improve any significant changes in the end-to-end system.

14. Agreement-based joint training for bidirectional attention-based end-to-end NMT

The study by [25] introduces the method on agreement-based joint training for bidirectional attention-based end-to-end NMT. It transforms the agreement-based learning method into attention-based NMT method. The study also focuses on applying the source-to-target and target-to-source models to merge on word-based alignment matrices on the current dataset. The idea is based on defining new objectives for training in bidirectional context as an agreement which measures the level of consensus among word aligned matrices in bidirectional way.

The agreement-based joint training shows the ability to effectively remove the unlikely attention resulting into an extension on focus orientation and effective alignment of matrices in bidirectional medium. During the experiment there is a loss function which measures the disagreement between the two matrices and a hyper-parameter to ensure the balance between the likelihood and agreement respectively. In addition to that, the loss function comprises of three major types including, *Square of Addition (SOA)* which is the addition of matrix cells using square of element-wise to corresponding cells, *Square of Subtraction (SOS)* which is the subtraction of matrix cells using square of element-wise to corresponding cells , *Multiplication (MUL)* which is the multiplication of element-wise to corresponding matrix respectively.

The experiment comprises of Chinese – English and English – French datasets. The NIST 2006 dataset is used as a set for validation in particular aspect for hyper-parameter optimization and model selection. Similarly, the NIST 2002, 2003, 2004, 2005 and 2008 dataset apply for testing purposes respectively. Furthermore, for English – French

dataset, the WMT 2014 training corpus is used consisting of 12.07M sentence pairs with 303.88M English words and 348.24M French words respectively. The approach of this study is compared with MOSES [23] and RNNSEARCH [24] approaches respectively. For MOSES, the experiment uses parallel corpus to train phrase-based model and target-side of parallel corpus to train 4-gram SRILM model. Similarly, for RNNSEARCH, the parallel corpus is used to train the attention based NMT and the size of vocabulary is set of 30k for all languages.

The experiment results in comparison of loss functions shows for attention-based NMT that by moving from Chinese – English and English – Chinese directions, the RNNSEARCH outperforms MOSES in exception to Chinese – English direction on NIST 2008 test set only. Similarly, the agreement-based joint training improves the quality in translation in bidirectional way in exception to English – Chinese direction on NIST 2004 test set only. The results on Chinese – English word alignment using TSINGHUA ALINGER dataset for evaluation and *"force decode"* method [26] on jointly trained models for generating translations compare with the reference finds a significant decrease on the alignment errors for bidirectional in comparison with independent training case. The results on English – French translation shows that independent training by RNNSEARCH achieves higher performance over MOSES and agreement-based joint learning improves to a higher extent over both approaches.

15. Massively Multilingual NMT

Multilingual NMT performs single model training from multiple source languages into multiple target languages. The study by [27] performed experiments on different ways of training the multilingual model and analyse the translation quality and the correlation between such models. The experiment comprises of a low-resource setting with 59 languages and high-resource setting with 102 languages respectively. The factors for model evaluation include, model capacity, number of trained tasks (directions), low-resource in comparison with high-resource setting.

The experiment for low-resource setting initialize with TED Talks parallel corpus formed by [28] which has 59 languages parallel data. The model deployed in the experiment comprises of three different types which include, many-to-many model having 116 translation directions to-and-from English with 58 languages, one-to-many model starts from English into 58 language, many-to-one model from 58 languages to English respectively. Similarly, there is a target-language prefix token to each source sentence for ensuring many-to-many translations. The experiment is restricted to Transformer models proposed by [29] due to their efficiency in context of multilingual models. The model applies with 6 layers both encoder and decoder having model dimension set to 512 and hidden dimension with size of 2048 as well as 8 attentions heads. The model also applies with dropout at dropping out rate of 0.2 on input embeddings, positional embeddings, output of every sub-layer before residual connection, inner output layer after ReLU activation for every feed-forward sub-layer and lastly, attention weights on every sub-layer. The results have shown that the model can scale on 59 languages with following conclusion including, outperformance of multilingual many-to-many models over many-to-one, bilingual models having similar ability and identical conditions of training. Similarly, many-to-many model performance is inferior when out-of-English words are applies as compare with one-to-many model. Lastly, the low-resource settings in models gives outstanding results as compare with any other previous models having the ability to support for up to 59 languages respectively.

The experiment of high-resource setting initializes with create an in-house dataset comprise of 102 languages pairs with to-and-from English having almost one million examples per pair of language. The dataset was tokenize using an in-house tokenizer and applies to join the subword segmentation for open-vocabulary purposes. Similarly, the vocabulary used is 64K instead of 32K. Furthermore, the model applies for the experiment comprises of a larger Transformer model having 6 layers both encoder and decoder along with model dimension set to 1024, the size of hidden dimension equals 8192 with 16 attention models. The results have shown that massively multilingual NMT is possible in large scale dimensions and can improve the performance of bilingual baselines. However, it appears to damage the performance for German-English language pair which shows that there is a significant need to further explore the trade-offs between the accuracy, model capacity and training configuration of the model for these languages.

16. XLNet: Generalized autoregressive pretraining for language understanding

The study proposed by [30] is based on generalized autoregressive pretraining for NMT method which enables bidirectional learning context by estimating the expected likelihood for all the permutations based on factorized order and control the effect on the limitation of BERT (a pretraining approach) [31]. The successful implementation of unsupervised learning in the area of natural language processing open new horizons for research. Within this high-level idea different pretraining objectives have been explored. The most notable unsupervised pretraining objectives include, autoregressive (AR) language modelling and autoencoding (AE) respectively. The XLNet proposed study leverages both the AR [32] and AE [33] modelling while considering their limitations also. The application of maximizing the log likelihood with respect to all the possibility of permutations for factorization order. The

XLNet proposed model doesn't rely on data corruption which gives an edge to avoid the pretrain-finetuning discrepancy which was originated by BERT mostly. Similarly, the XLNet proposed model also improves the architectural designs for pretraining objectives influence by recent advancement in AR language modelling. XLNet introduces the segment recurrence mechanism and relative encoding scheme of Transformer-XL into pretraining [34] which shows empirical improvement in performance for longer text sequence. It also proposed a solution to Transformer (-XL) architecture in the which permutation-based language modelling doesn't work due to arbitrary factorization order and ambiguity in target. The solution for the stated problem includes removal of ambiguity by reparametrizing the Transformer (-XL) network.

The experiment for the proposed model has followed with BERT, to BookCorpus [35] and Wikipedia for English as an initializing point for pretraining. Similarly, other text includes, Giga5 (16Gb) ClueWeb 2012-B and Common Crawl respectively. The architecture model for XLNet is the same with hyperparameter as BERT-Large model size. The XLNet-Large-wikibooks is also trained with BookCorpus and Wikipedia having reutilized all the parameters as elaborated in BERT. Furthermore, the experiment also used bidirectional pipeline for data input in which every forward and backward direction consumes half in terms of its batch size. The results have shown that XLNet achieves enormous significance in terms of performance improvement in comparison with other pretraining tasks. Finally, the proposed models were analysed based on ablation study which results in determining the performance superiority among various models i.e. Transformer-XL performance on XLNet over BERT.

17. Efficient 8-bit quantization of transformer NMT translation model

The study proposed by [36] introduces the quantization of deep learning-based Transformer model first time in the industry. The study proposed a quantization technique to using TensorFlow to replace the 32-bit floating point (FP32) with 8-bit integers (INT8). Similarly, it also computationally transforms the graph using TensorFlow library. The study also proposed a batching technique in parallel to maximize the utilization of CPU power during inference process. The resulting outcome on optimization has improved significant performance for FP32 and INT8 model on quantization over net scale of 1.5X among the best model on quantization for FP32 performance. The improvement inference on performance scale include, quantized optimization on MatMuls for the size and tensor shape in Transformer model, decrease in overhead problem due to the quantization operations in the compute graph of Transformer model, pipeline optimization on input by sentence ordering on token length and apply parallel batches for execution having increasing inference on throughput.

The experiment for the proposed study includes, applying FP32 and INT8 evaluation on 2S Intel® Xeon® Platinum 8168 (24 cores per socket) processors and Intel® Xeon® Platinum 8268 (24 cores per socket) processors. The TensorFlow library for the inference of performance on Transformer model in both FP32 and INT8 are evaluated with a *newstest2014* dataset respectively. Similarly, a mini batch is applied throughout the experiment having the size of 64 in scale. The results have shown on the quantization of Transformer ML translation model using TensorFlow library having maintained the BLEU score accuracy below 0.5 drop. It shows that the quantization of MatMuls produces certain overheads on the Dequantize and QuantizeV2 on the graph for INT8 values. Similarly, the major learning model using non-linear i.e. Softmax layer and normalization on layer appeared to be between layers intensifying the process for the purpose of quantization efforts. Furthermore, the conducted experiment optimizes the compute graph in terms of reducing the operation numbers, kernels on the key operation improvement i.e. MatMuls and GatherNd (the operations on N-dimensional tensors having input indices to fulfil Gather on input tensor to produce an output tensor), the order optimization of sentences within input pipeline and parallel batching for achieving higher throughput gains within 1.5X scale.

IV. **Related Works**

1. OpenNMT - Open-Source NMT toolkit

The study by [37] introduces an open-source toolkit namely OpenNMT for NMT and perform analysis on its multiple features related to efficiency, modularity, extensibility, modalities with reference to its NMT architecture. The open-source toolkit also focuses on feature representations, source data modalities by maintaining its competitive requirements related to performance and effective training. The design goals for this toolkit was increasing system efficiency, memory sharing which works with GPU based NMT models, integrate CPU/GPU/Mobile based translation, tokenization and word embedding features. The toolkit was designed to improve research in NMT related problems, and it is considered to be more feature oriented in this domain in future as propounded and maintained by its designers.

2. Nematus - Open-Source NMT toolkit

The study by [38] introduces an open-source toolkit namely Nematus for NMT based MT. The toolkit prioritizes translation accuracy, usability, extensibility and performance. The toolkit was implemented with decoder hidden state along with source annotation rather position based encoder with backward RNN. Similarly, it also includes novel conditional GRU including attention mechanism. The

decoder was initialized with feedforward hidden layer with *tanh* non-linearity instead of softmax layer. The word embedding is equipped with encoder and decoder layers and no further layers were introduced. Furthermore, the toolkit was implemented with *Look*, *Update* and *Generate* decoder techniques which improves the implementing efficiency of decoder. It is also experimented with recurrent Bayesian based dropout [39]. Lastly, instead of single word embedding at every source point, the toolkit was implemented with multiple factors of input representation at every timestamp with effect of concatenation with final embedding for each feature. [40]

V. **Conclusion**

Neural Machine Translation transform the Machine Translation into an emerging discipline. Unlike Statistical MT which lack many feature dimensions in translation, the NMT opens a new horizon of research and support wider spectrum for the implementation of Machine Learning algorithms into Natural Language Processing. In this paper, we discussed upon various NMT models proposed particularly in areas related to efficiency in Machine Translation through shared task. Similarly, the experimental observations of these model were also elaborated with effect to their results.

REFERENCES

[1] Q. Lanners, "Neural Machine Translation," Towards Data Science, [Online]. Available: https://towardsdatascience.com/neural-machine-translation-15ecf6b0b.

[2] P. Koehn, "Neural Machine Translation," in *Statistical Machine Translation*, 2017.

[3] Y. Nishimura, K. Sudoh, G. Neubig and S. Nakamura, "Multi-Source Neural Machine Translation with Missing Data," *Association for Computational Linguistics,* vol. Proceedings of the 2nd Workshop on Neural Machine Translation and Generation, pp. 92-99, 2018.

[4] A. Fan, D. Grangier and M. Auli, "Controllable Abstractive Summarization," *Association for Computational Linguistics,* vol. Proceedings of the 2nd Workshop on Neural Machine Translation and Generation, pp. 45-54, 2018.

[5] I. J. Unanue, E. Z. Borzeshi and M. Piccardi, "A Shared Attention Mechanism for Interpretation of Neural Automatic Post-Editing Systems," *Association for Computational Linguistics,* vol. Proceedings of the 2nd Workshop on Neural Machine Translation and Generation, pp. 11-17, 2018.

[6] H. Bechara, Statistical Post-editing and Quality Estimation for Machine Translation System, Dublin City University School of Computing, 2013.

[7] O. Bojar, R. Chatterjee, C. Federmann, Y. Graham, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, V. Logacheva, C. Monz, M. Negri, A. Névéol, M. Neves, M. Popel, M. Post and R. Rubino, "Findings of the 2016 Conference on Machine Translation," *Association for Computational Linguistics,* Vols. Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers, pp. 131-198, 2016.

[8] V. C. D. Hoang, P. Koehn, G. Haffari and T. Cohn, "Iterative Back-Translation for Neural Machine Translation," *Association for Computational Linguistics,* vol. Proceedings of the 2nd Workshop on Neural Machine Translation and Generation, pp. 18-24, 2018.

[9] H. Khayrallah and P. Koehn, "On the Impact of Various Types of Noise on Neural Machine Translation," *Association for Computational Linguistics,* vol. Proceedings of the 2nd Workshop on Neural Machine Translation and Generation, pp. 74-83, 2018.

[10] X. Niu, M. Denkowski and M. Carpuat, "Bi-Directional Neural Machine Translation with Synthetic Parallel Data," *Association for Computational Linguistics,* vol. Proceedings of the 2nd Workshop on Neural Machine Translation and Generation, pp. 84-91, 2018.

[11] Y. Bisk and K. Tran, "Inducing Grammars with and for Neural Machine Translation," *Association for Computational Linguistics,* vol. Proceedings of the 2nd Workshop on Neural Machine Translation and Generation, pp. 25-35, 2018.

[12] H. Khayrallah, B. Thompson, K. Duh and P. Koehn, "Regularized Training Objective for Continued Training for Domain Adaptation in Neural Machine Translation," *Association for Computational Linguistics,* vol. Proceedings of the 2nd Workshop on Neural Machine Translation and Generation, pp. 36-44, 2018.

[13] I. J. Goodfellow, M. Mirza, D. Xiao, A. Courville and Y. Bengio, "An Empirical Investigation of Catastrophic Forgetting in Gradient-Based Neural Networks," 2015.

[14] D. Bahdanau, K. Cho and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," *Proceedings of the International Conference on Learning Representations (ICLR),* 2015.

[15] D. He, Y. Xia, T. Q. L. Wang, N. Yu, T.-Y. Liu and W.-Y. Ma, "Dual Learning for Machine Translation," vol. Advances in Neural Information Processing Systems 29 (NIPS), 2016.

[16] K. Imamura, A. Fujita and E. Sumita, "Enhancement of Encoder and Attention Using Target Monolingual Corpora in Neural Machine Translation," *Association for Computational Linguistics,* vol. Proceedings of the 2nd Workshop on Neural Machine Translation and Generation, pp. 55-63, 2018.

[17] R. Sennrich, B. Haddow and A. Birch, "Improving Neural Machine Translation Models with Monolingual Data," *Association for Computational Linguistics,* vol. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 86-96, 2016.

[18] S. S. R. Kothur, R. Knowles and P. Koehn, "Document-Level Adaptation for Neural Machine Translation," *Association for Computational Linguistics,* vol. Proceedings of the 2nd Workshop on Neural Machine Translation and Generation, pp. 64-73, July.

[19] J. Wuebker, S. Green, J. DeNero, S. Hasan and M.-T. Luong, "Models and Inference for Prefix-Constrained Machine Translation," *Association for Computational Linguistics,* vol. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 66-75, 2016.

[20] P. Koehn and R. Knowles, "Six Challenges for Neural Machine Translation," *Association for Computational Linguistics,* vol. Proceedings of the First Workshop on Neural Machine Translation, pp. 28-39, 2017.

[21] R. Sennrich, B. Haddow and A. Birch, "Neural Machine Translation of Rare Words with Subword Units," *arXiv:1508.07909v5.*

[22] Y. Cheng, W. Xu, Z. He, W. He, H. Wu, M. Sun and Y. Liu, "Semi-Supervised Learning for Neural Machine Translation," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers*), Berlin, Germany, 2016.

[23] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin and E. Herbst, "Moses: Open Source Toolkit for Statistical Machine Translation," in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume*

Proceedings of the Demo and Poster Sessions, Prague, Czech Republic, 2007.

[24] D. Bahdanau, K. Cho and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," in *Proceedings of ICLR*, 2015.

[25] Y. Cheng, S. Shen, Z. He, W. He, H. Wu, M. Sun and Y. Liu, "Agreement-based Joint Training for Bidirectional Attention-based Neural Machine Translation," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)*, New York, USA, 2015.

[26] T. Luong, I. Sutskever, Q. Le, O. Vinyals and W. Zaremba, "Addressing the Rare Word Problem in Neural Machine Translation," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Beijing, China, 2015.

[27] R. Aharoni, M. Johnson and O. Firat, "Massively Multilingual Neural Machine Translation," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, 2019.

[28] Y. Qi, D. Sachan, M. Felix, S. Padmanabhan and G. Neubig, "When and Why Are Pre-Trained Word Embeddings Useful for Neural Machine Translation?," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, New Orleans, Louisiana, 2018.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, "Attention Is All You Need," in *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, California, United States, 2017.

[30] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov and Q. V. Le, "XLNet: Generalized Autoregressive Pretraining for Language Understanding," in *Advances in Neural Information Processing Systems 32 (NIPS 2019)*, California, USA, 2019 .

[31] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, 2019.

[32] B. Uria, M.-A. Côté, K. Gregor, I. Murray and H. Larochelle, "Neural autoregressive distribution estimation," *The Journal of Machine Learning Research,* vol. 17, no. 1, pp. 7184-7220, 2016.

[33] M. Germain, K. Gregor, I. Murray and H. Larochelle, "MADE: Masked Autoencoder for Distribution Estimation," in *Proceedings of the 32nd International Conference on Machine Learning*, 2015.

[34] Z. Dai, Z. Yang, Y. Yang, W. W. Cohen, J. Carbonell, Q. V. Le and R. Salakhutdinov, "Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context," in *arxiv.org/abs/1901.02860*, 2019.

[35] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba and S. Fidler, "Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books," in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV'15)*, Washington, DC, United States of America, 2015.

[36] A. Bhandare, V. Sripathi, D. Karkada, V. Menon, S. Choi, K. Datta and V. Saletore, "Efficient 8-Bit Quantization of Transformer Neural Machine Language Translation Model," in *arxiv.org/abs/1906.00532*, 2019.

[37] G. Klein, Y. Kim, Y. Deng, J. Senellart and A. M. Rush, "OpenNMT: Open-Source Toolkit for Neural Machine Translation," *arXiv:1701.02810v2* , 2017.

[38] R. Sennrich, O. Firat, K. Cho, A. Birch, B. Haddow, J. Hitschler, M. Junczys-Dowmunt, S. Läubli, A. V. M. Barone, J. Mokry and M. Nădejde, "Nematus: a Toolkit for Neural Machine Translation," *arXiv:1703.04357v1.*

[39] Y. Gal and Z. Ghahramani, "A Theoretically Grounded Application of Dropout in Recurrent Neural Networks," *arXiv:1512.05287v5.*

[40] R. Sennrich, B. Haddow and A. Birch, "Edinburgh Neural Machine Translation Systems for WMT 16," *Association for Computational Linguistics,* Vols. Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers, pp. 371-376, 2016.