Deep Neural Transformer Model for Mono and Multi Lingual Machine Translation

Mohamed Islam Khaber
Department of computer sciences
University of Ferhat Abbas 1
Setif, Algeria
mohamedislam.khaber@univ-setif.dz

Abdelouahab Moussaoui
Department of computer sciences
University of Ferhat Abbas 1
Setif, Algeria
abdelouahab.moussaoui@univ-setif.dz

Mohamed Saidi
Department of computer sciences
University of Ferhat Abbas 1
Setif, Algeria
mohamed.saidi@univ-setif.dz

Nabila Frahta
Department of technology
University of Ferhat Abbas 1
Setif, Algeria
frahta.nabila@gmail.com

***Abstract*—In recent years, the Transformers have emerged as the most relevant deep architecture, especially machine translation. These models, which are based on attention mechanisms, outperformed previous neural machine translation architectures in several tasks. This paper proposes a new architecture based on the transformer model for the monolingual and multilingual translation system. The tests were carried out on the IWSLT 2015 and 2016 dataset. The Transformers attention mechanism increases the accuracy to more than 92% that we can quantify by more than 4 BLEU points (a performance metric used in machine translation systems).**

Keywords—Neural machine translation (NMT), deep learning, multilingual, transformer, monolingual.

I. INTRODUCTION

Deep learning is a subset of artificial neural network-based machine learning methods [1]. It allows computational models with multiple processing layers to learn different abstraction levels for data representations. These methods have improved the state-of-the-art research in language translation [2].

Neural Machine Translation (NMT) is a deep learning end-to-end machine translation approach that utilizes an extensive artificial neural network to predict a set of words' probabilities. Typically, entire sentences are modelled in a single integrated model. This approach has the advantage of being able to train a single system on both the source and target text and generates high-quality translation results. NMT has recently shown promising results on several language pairs [3],[4].

The end-to-end learning approach of NMT models consists of two essential components, the encoder, and decoder, which are usually built on similar neural networks of different types, such as recurrent neural networks [5], convolution neural networks [6], and more recently on transformer models, which are built entirely with attention layers, without convolution or recurrence [7].

In neural machine translation, multiple model variants and training procedures have been proposed and tested. NMT models were generally used in single language-pair settings, where a parallel corpus from a source language to a target language is required for the training process, and the inference process only involves those two languages in the same direction.

In our approach explained in this paper, we proposed a method deep transformer model for machine translation (DTMMT) with existing transformer architectures for analyses the translation outputs of multiple-languages and single-language models. We utilise the data collected in the IWSLT 2015 and IWSLT 2016 MT evaluation campaign [8]. A multilingual NMT system may be a factor in improving the final system, which improves by over 4 BLEU points over the monolingual NMT system [9].

The paper is organized as follows. In section II, we begin with a brief review of related work interested in the monolingual NMT and multilingual NMT of MT tasks. In Section III, we introduce our monolingual and multilingual NMT approaches to propose a good NMT system. In section IV, describes the transformer architecture. In section V, we describe our experimental set-up and discuss the results of our experiments. Section VI ends the paper with our conclusions.

II. RELATED WORK

In related works, Researchers have tried to build Multilingual NMT systems at Monolingual NMT systems expense in recent approaches. Bahdanau et al. [10] proposed an encoder-decoder architecture based on recurrent neural networks and attention in the Neural Machine Translation field, capable of translating between language pair consider one-to-one translation systems.

In a many-to-many translation system, Firat et al. [11] introduced a way to share the attention mechanism across multiple languages. In particular, multiple languages are applied to both sides. Luong et al. [12] used different encoders and decoders for each source and target language.

Dong et al. [13] proposed a multi-task learning approach for a one-to-many translation system by sharing hidden representations among related tasks to enhance generalization on the target language. They used a single language in the source and separate attention mechanisms and multiple languages on the target side.

Zoph et al. [14] employed a many-to-one translation system that considers multiple languages in the source and one language in the target side. Similarly, Gu et al. [15] propose a

Mixture-of-Language-Experts and a Universal Language Representation layer to improve a many-to-one model from different 5 languages into target English. Malaviya et al. [16] trained a many-to-one system from bible translation and used it to infer typology features for the different languages without evaluating the translation quality. In another related work, Artetxe et al. [17] trained a multilingual NMT model and used the learned representations to perform cross-lingual transfer learning.

Recent works propose different parameter sharing methods between language pairs in a multilingual NMT system. Blackwood et al. [18] propose sharing all parameters and shows improvements in over-sharing all parameters. Sachan et al. [19] explore sharing various components in Transformers models. Platanios et al. [20] propose to share the entire neural network while using a contextual parameter generator that learns to generate the system's parameters given the desired source and target language.

For the Arabic language, Almahairi et al. [21] proposed NMT for Arabic translation in both directions (Arabic-English and English-Arabic) and compared NMT and SMT and showed that NMT better than SMT, which is the first result on Arabic neural machine translation. Preprocessing Arabic texts increase the performance of the system, especially normalization. The morphology of Arabic languages is complex and productive, with a primary word-formation mechanism known as root-and-pattern. For example, from the Arabic word "وسوف يكتبونه "(wasawf yaktwubunahu) and its English translation "and they will write it". A possible analysis of these complex words defines the stem as "aktub" (write), with an inflectional circumfix, "y-uwna", denoting masculine plural, an inflectional suffix, "ha" (it), and two prefixes, "sa" (will) and "wa" (and) [22].

NMT has many challenges, such as; One model trained to translate many languages instead of one model per language [23]. This paper deals with this problem.

III. Deep Transformer Model for Machine Translation (DTMMT)

A. Monolingual NMT System

The proposed approach is performed in two steps. In the first step, a monolingual NMT system which is the simplest and most effective one, trains a single neural network on parallel data, including French-to-English and Arabic-to-English, as shown in Fig 1.

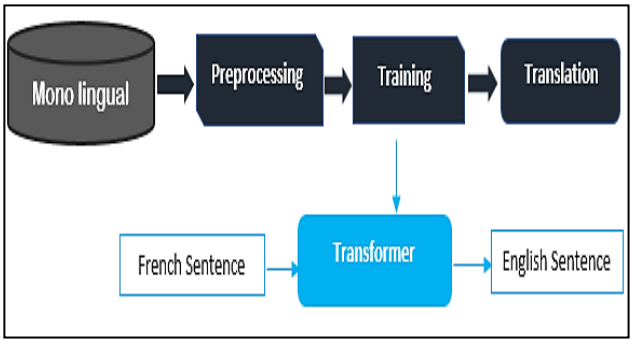

Fig. 1. Generic graph for the monolingual NMT system

B. Multilingual NMT System

In the second step, a multilingual NMT system is trained on the available data from different languages such as French and Arabic after concatenation L1 and L2, as shown in Fig 2.

We follow the method of Zoph et al. [14]. We add a target-language token to each source sentence to enable a many-to-one translation system. This different setup enables us to examine translation quality.

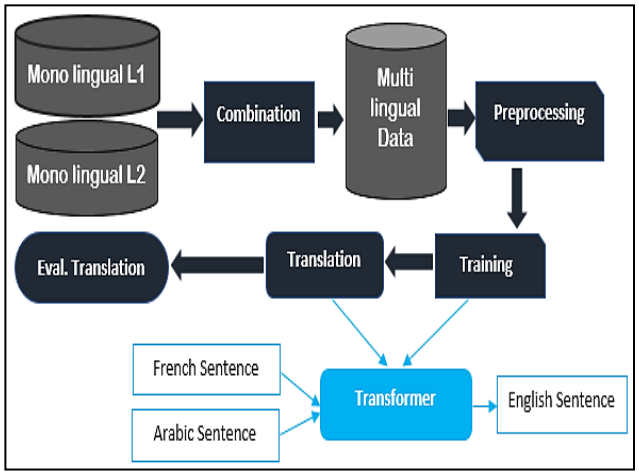

Fig. 2. Generic graph for the multilingual NMT system

We use this approach in our work with the Transformer model, which distinguishes multilingual NMT training and inference from a single language pair NMT to improve translation performance with minimal complexity. In addition to reducing Training of several single language pair systems. Transformer model, Preprocessing and Training, we will describe it in sections IV, V.

IV. The Transformer Model

Transformers are deep learning models introduced in 2017 [7], used for the first time in natural language processing (NLP). The first transduction model relies entirely on a self-attention mechanism to compute representations of its input and output with seq2seq modelling and without using sequence-aligned (RNNs) architecture or convolution architecture (CNN).

A transformer is composed of an encoder and a decoder. The encoder's role is to encode the inputs (i.e. sentence) in a state, which often contains several tensors. Then the state is passed into the decoder to generate the outputs. In machine translation, the encoder transforms a source sentence, e.g., "The Black Cat.", into a state, e.g., a vector, that captures its semantic information. The decoder then uses this state to generate the translated target sentence, e.g., "Le chat noir.". both the encoder and decoder are composed of two main components: Multi-Head Self-Attention and Feed Forward Network.

Attention mechanism (Scaled Dot-Product Attention): The attention mechanism's primary goal is to estimate the relative importance of the keys term in comparison to the query term for the same concept. To that end, the attention mechanism takes query Q that represents a vector word, the keys K which

are all other words in the sentence, and value V represents the vector of the word. the attention mechanism gives us the importance of the word in a specific sentence.

The Transformer model uses the Multi-Head Attention mechanism; it is simply a projection of Q, K and V in h Linear Spaces. Perform the attention function in parallel on each of these projected versions of queries, keys, and values, producing DV-dimensional output values. The final values are calculated by concatenating these and projecting them again.

The Multi-Head Attention mechanism's output, h attention matrix for each word, is then concatenated to create one matrix per word. This Attention architecture enables more complex dependencies between words without adding any training time thanks to the linear projection, which reduces each word vector's size.

Input Embedding aims at creating a vector representation of words. Words with the same meaning will be close in terms of Euclidian distance. The authors decided to use a 512 size embedding for the encoder [7].

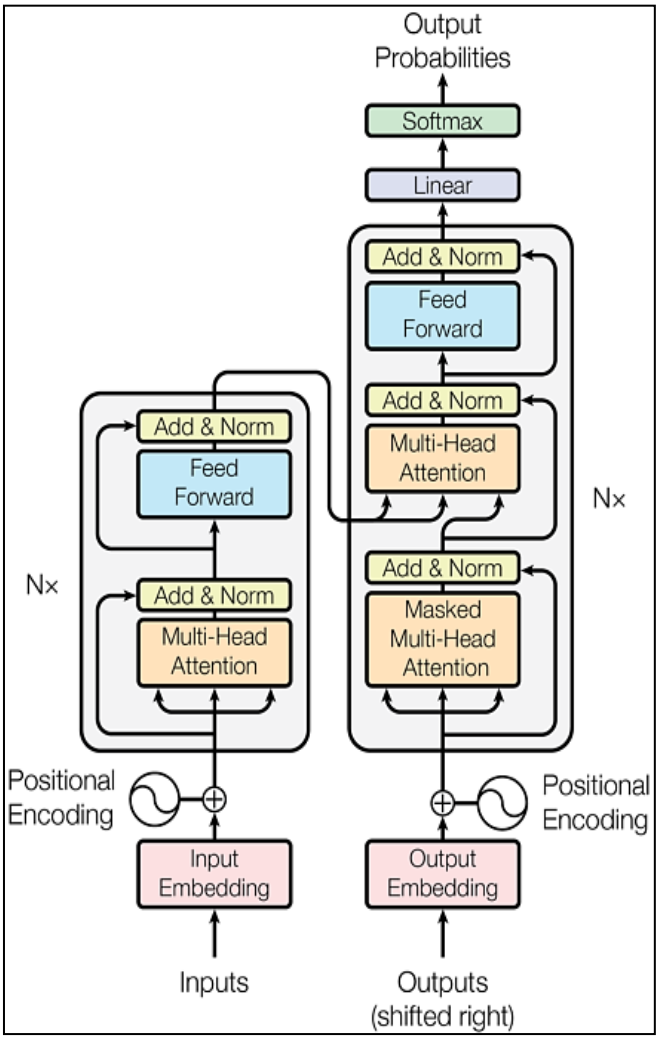

Fig. 3. The Transformer model architecture (from Vaswani et al.)

In the encoder phase, the Transformer first generates initial inputs creating by Input Embedding and Position Encoding for each word in the input sequence. Then, for each word, self-attention aggregates information from all other words in the context of the sentence, thus creating a new representation for each word, which is an attended representation of all other words in the sequence; this is repeated for each word in a sequence, successively building newer representations on top of previous one's multiple times.

The decoder generates one word at a time from left to right. The first word is based on the encoder's final representation (offset by one position). Each word predicted subsequently attends to the decoder's previously generated words at that layer and the final representation of the encoder (Multi-Head Attention) similar to a typical encoder-decoder architecture.

V. Experiment and Results

A. *Dataset and Preprocessing*

We trained and evaluated both monolingual and multilingual NMT systems based on the transformer models, with the relevant training data in the IWSLT 2016[(1)] evaluation campaign [8], which represent transcriptions from TED talks.

The experimental setting comprises two languages: French-to-English and Arabic-to-English; for each language pair, we use the training data of approximately 200,000 parallel sentences. For the models' development and evaluation, we use the corresponding sets from the IWSLT2010, which are composed of 887 sentence pairs. We also use IWSLT2015 and IWSLT2016 datasets as the test set for both language pairs is composed of 1080 and 1133 sentence pairs, respectively. Details about the used data sets are reported in Table I.

TABLE I. The Total Number of Parallel Sentences Used for Training, Development, and Test

Language Pair	Train	Dev10	Test15	Test16
French-to-English	218081	887	1080	1133
Arabic-to-English	211726	887	1080	1133
All-to-English	429807	887	1080	1133

At the preprocessing, we applied word segmentation for each training condition (i.e., monolingual and multilingual) by learning a sub-word dictionary via Byte-Pair Encoding (BPE) [24]. We use byte pair encoding (BPE) to learn a variable-length encoding of the text's vocabulary; unlike the original BPE, it does not compress the plain text. Still, it can reduce the text's vocabulary to a configurable number of symbols,

(1) https://wit3.fbk.eu/

with only a small increase in tokens. BPE is considered the best preprocessing method for Arabic [25]; the number of BPE segmentation rules is set to 6000, following Denkowski et al. [26] for experiments with small training data condition. We removed sentence pairs longer than 100 words.

B. Models and Parameters

The transformer models are trained using the open-source Open-NMT in PyTorch(2) toolkit [27]. The two systems types, monolingual and multilingual, were trained with the same model and parameters.

The hyperparameters for both systems were set as follows: a dropout of 0.1 is used, 4 attention blocks in the encoder and decoder and 8 attention heads were used, and the embedding size was 512, feed-forward dimension 2048. Adam and Noam decay were used for optimization [28].

We trained each of the two monolingual NMT systems separately; for the multilingual NMT system, we combined the two parallel corpora. We stopped training when the validation accuracy became constant or increased very slowly from the previous steps. Training time and training steps are shown in table II.

TABLE II. TRAINING TIME AND TRAINING STEPS FOR DIFFERENT SYSTEMS

Systems	Training time (Hours)	Training steps
French-to-English	3	50000
Arabic-to-English	2.5	45000
All-to- English	6	70000

C. Results

We compare the translation performance of two independently trained single-language models against one multiple-languages model trained on combining the two language pairs. The experiments show that a multilingual NMT system outperforms the monolingual NMT systems. The performance of both types of systems is evaluated on IWSLT2015, IWSLT2016 and reported in Table III.

TABLE III. BLEU RESULTS FOR IWSLT2015 AND IWSLT2016 TEST SETS

Dataset /Systems	**French-to-English**		**Arabic-to-English**	
	IWSLT2015	***IWSLT2016***	***IWSLT2015***	***IWSLT2016***
Monolingual	20.78	19.27	19.03	20.35
Multilingual	**25.31**	**24.65**	**23.06**	**25.88**

The improvements observed in the multilingual NMT system are likely due to increased data, even if it is not from the same source language.

Table IV and V shows some examples of translations with long sentences and short sentences.

TABLE IV. FRENCH/ARABIC TO ENGLISH EXAMPLES FOR IWSLT2016 WITH LONG SENTENCES

French	Rien qu'au cours des dernières années, nous avons beaucoup appris sur la façon dont la Terre s'intègre dans le contexte de notre univers.
Single-Language	In the last few years, we learned a lot about how the Earth is in the context of our universe.
Multi-Language	In the last few years, we have learned a lot about how the Earth fits into the context of our universe.
Arabic	لأكثر من 20 قرناً، كان الأطباء يسردون قصة واحدة عن معنى التوحد وكيف تم إكتشافه، ليتبين فيما بعد أن هذه القصة خاطئة وأن ما نتج عنها ترك أثر مدمر على الصحة العامة.
Single-Language	For more than 20 centuries, doctors telling a single story of a autism and how discovered, only to find out later that this story was wrong and made up of a global impact on the health.
Multi-Language	For more than 20 centuries, doctors have been telling a single story about the meaning of autism and how it was discovered, and it turns out that this story is wrong and that what has resulted has had a devastating impact on public health.

TABLE V. FRENCH/ARABIC TO ENGLISH EXAMPLES FOR IWSLT2016 WITH SHORT SENTENCES

French	Je regarde le comportement des gens et leur réponse au son.
Single-Language	I look at the behavior of people and the answer to sound.
Multi-Language	I look at people's behavior and their response to sound.
Arabic	لقد ولدت صماء، وعلموني أن الصوت ليس جزء من حياتي.
Single-Language	I was born with the audiences, and they that sound is not part of my life.
Multi-Language	I was born deaf, and they taught me that sound is not part of my life.

(2) https://github.com/OpenNMT/OpenNMT-py

Finally, we analysed the results using the IWSLT data, Fig 4 and 5 shows the breakdown of BLEU in the test data, separating the results for French to English and Arabic to English.

When all data are present in training, multilingual NMT system for both dataset 2015 and 2016 has better performance than monolinguals NMT systems. However, we observe test results when training with a single-language pair is low.

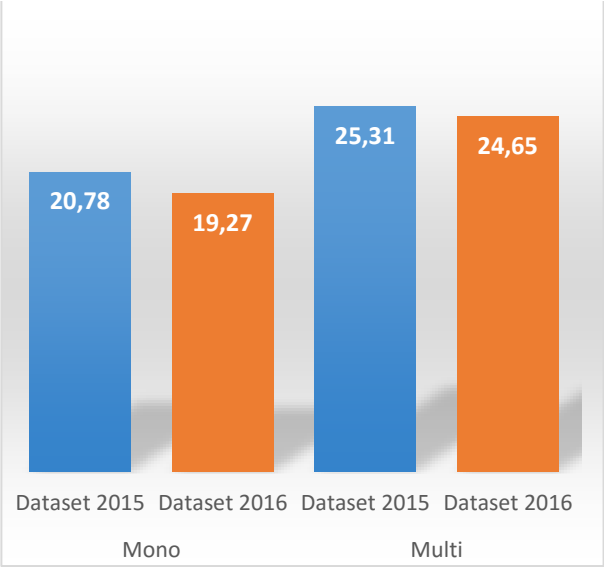

Fig. 4. Detailed comparison of BLEU in IWSLT dataset for French to English

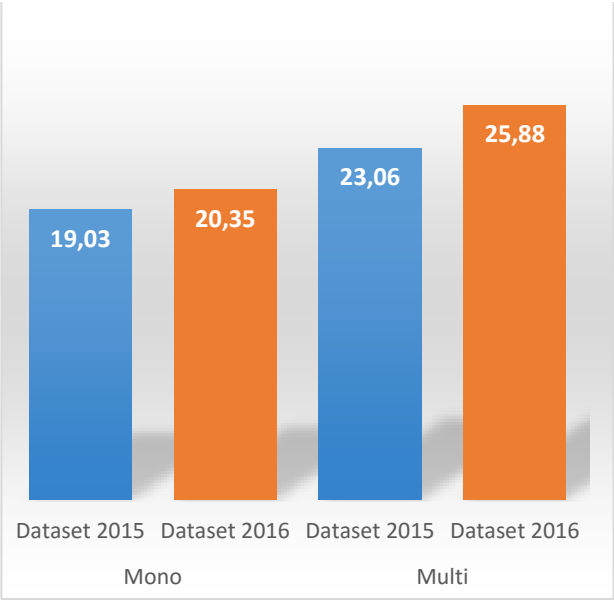

Fig. 5. Detailed comparison of BLEU in IWSLT dataset for Arabic to English

D. Evaluation methods

The accuracy of every translation result was compared based on the BLEU score [9] as implemented in multi-bleu.perl[(3)].

(3) https://github.com/moses-smt/mosesdecoder

VI. CONCLUSIONS

This work showed that a single multiple-languages model outperforms single-languages models applied to the Transformer architecture while avoiding the need for multiple language pairs to be trained. This model shows improvements in the final translation quality with over 4 BLEU points. In future work, we plan to explore our approach across many language varieties using a multilingual model and experiment with different architecture.

It's possible that if we designed a better strategy for multilingual NMT system like add more data or add more languages we may be able to obtain better results.

REFERENCES

[1] A Oppermann, "What is Deep Learning and How does it work?", [Online]. Available: https://towardsdatascience.com/what-is-deep-learning-and-how-does-it-work-2ce44bb692ac ,2019.

[2] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, p. 436, 2015.

[3] 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 , Vol. 1, 2016, pp. 1965–1974.

[4] F. Hieber, T.Domhan, M. Denkowski, D. Vilar, A. Sokolov, A. Clifton, and M. Post, "The SOCKEYE Neural Machine Translation Toolkit at AMTA ", In Proceedings of AMTA, vol. 1, 2018, pp. 200-207.

[5] I. Sutskever, O. Vinyals, and Q. V Le, "Sequence to sequence learning with neural networks", In Advances in neural information processing systems, 2014, pp. 3104–3112.

[6] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. Dauphin, "Convolutional sequence to sequence learning", In Proceedings of the 34th International Conference on Machine Learning, 2017, pp. 1243-1252.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, " Attention is all you need", In Advances in Neural Information Processing Systems, 2017, pp. 6000–6010.

[8] M. Cettolo, J. Niehues, S. Stüker, L. Bentivogli, R. Cattoni, and M. Federico, "The IWSLT evaluation campaign", In Proceedings of the International Workshop on Spoken Language Translation (IWSLT), Seattle, WA, 2016.

[9] K. Papineni, S. Roukos, T. Ward, and W. Zhu., "Bleu: a method for automatic evaluation of machine translation", In Proceedings of the 40th annual meeting on association for computational linguistics, 2002, pp. 311–318.

[10] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate", In Proceeding of the 3rd International Conference on Learning Representations, ICLR ,USA, May 7-9, 2015.

[11] O. Firat, K. Cho, and Y. Bengio, "Multi-way, multilingual neural machine translation with a shared attention mechanism", In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics, 2016, pp. 866–875.

[12] M. Luong, Q. V Le, I. Sutskever, O. Vinyals, and L. Kaiser, "Multi-task sequence to sequence learning", international conference on learning representations, 2016.

[13] D. Dong, H. Wu, W. He, D. Yu, and H. Wang, "Multi-task learning for multiple language translation". In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, vol. 1, 2015, pp. 1723–1732.

[14] B. Zoph and K. Knight, "Multi-source neural translation", In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics, 2016, pp. 30–34.

[15] J. Gu, H. Hassan, J. Devlin, and O. K. V. Li, "Universal neural machine translation for extremely low resource languages", In Proceedings of

the Conference of the North American Chapter of the Association for Computational Linguistics, 2018, pp.344–354.

[16] C. Malaviya, G. Neubig, and P. Littell, "Learning language representations for typology prediction", In Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2017.

[17] M. Artetxe and H. Schwenk, "Massively multilingual sentence embeddings for zeroshot cross-lingual transfer and beyond", Transactions of the Association for Computational Linguistics, 2019, pp.597--610.

[18] G. Blackwood, M. Ballesteros, and T. Ward, "Multilingual neural machine translation with task-specific attention", In Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, New Mexico, USA. Association for Computational Linguistics, 2018, pp.3112–3122.

[19] D. Sachan and G. Neubig, " Parameter sharing methods for multilingual self-attentional translation models", In Proceedings of the Third Conference on Machine Translation, Belgium, Brussels, 2018.

[20] E. A. Platanios, M. Sachan, G. Neubig, and T. Mitchell, "Contextual parameter generation for universal neural machine translation", In Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2018.

[21] A. Almahairi, K. Cho, N. Habash, and A. Courville," First result on Arabic neural machine translation", 2016.

[22] I. Gashaw and H. L. Shashirekha, "AMHARIC-ARABIC NEURAL MACHINE TRANSLATION". Computer Science & Information Technology (CS & IT) Computer Science Conference Proceedings (CSCP), 2019, pp. 55-68.

[23] R. Aharoni, M. Johnson, and O. Firat, " Massively Multilingual Neural Machine Translation". In Proceedings of NAACL-HLT, 2019, pp. 3874–3884.

[24] R.Sennrich, B. Haddow, and A.Birch, "Neural machine translation of rare words with subword units", In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Vol. 1, 2016, pp. 1715–1725.

[25] H. Sajjad, F. Dalvi, N. Durrani, A. Abdelali, Y. Belinkov, S. Vogel, "Challenging Language-Dependent Segmentation for Arabic: An Application to Machine Translation and Part-of-Speech Tagging", In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vol. 2, 2017, pp. 601–607.

[26] M. Denkowski and G. Neubig, "Stronger baselines for trustable results in neural machine translation", in Proceedings of the First Workshop on Neural Machine Translation, 2017, pp. 18-27.

[27] G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. Rush, "OpenNMT: Open-Source Toolkit for Neural Machine Translation", In Proceedings of association for computational linguistics, 2017, pp. 67–72.

[28] D. P. Kingma, J. Ba, " Adam: A Method for Stochastic Optimization", 3rd International Conference for Learning Representations, San Diego, 2015.