Enhancing Neural Arabic Machine Translation using Character-Level CNN-BILSTM and Hybrid Attention

Dhaya Eddine Messaoudi
ICOSI Laboratory, Abbes Laghrour University, Khenchela, Algeria
messaoudi.dhayaeddine@univ-khenchela.dz (corresponding author)

Djamel Nessah
ICOSI Laboratory, Abbes Laghrour University, Khenchela, Algeria
nessah_djamel@univ-khenchela.dz

Received: 12 July 2024 | Revised: 30 July 2024 | Accepted: 4 August 2024

 DOI: https://doi.org/10.48084/etasr.8383

ABSTRACT

Neural Machine Translation (NMT) has made significant strides in recent years, especially with the advent of deep learning, which has greatly enhanced performance across various Natural Language Processing (NLP) tasks. Despite these advances, NMT still falls short of perfect translation, facing ongoing challenges such as limited training data, handling rare words, and managing syntactic and semantic dependencies. This study introduces a multichannel character-level NMT model with hybrid attention for Arabic-English translation. The proposed approach addresses issues such as rare words and word alignment by encoding characters, incorporating Arabic word segmentation as handcrafted features, and using part-of-speech tagging in a multichannel CNN-BiLSTM encoder. The model then uses a Bi-LSTM decoder with hybrid attention to generate target language sentences. The proposed model was tested on a subset of the OPUS-100 dataset, achieving promising results.

Keywords-Arabic natural language processing; deep-learning; machine translation; deep CNN Bi-LSTM; hybrid attention; PoS-tagging; Arabic word segmentation

I. INTRODUCTION

Machine Translation (MT) systems are very important tools in human life today, as they make communication between people easier. People use such systems in different areas of their lives, such as professional, educational, or even travel settings. MT systems are considered one of the basic tasks in Natural Language Processing (NLP), where three principal approaches have been applied in their history:

- Rule-based approach.
- Statistical approach [1, 2].
- Neural approach [3, 4], which has become the leader in terms of performance and results, especially with the recent development of deep learning.

The actual development of Neural Machine Translation (NMT) began with the launch of the encoder-decoder model [5]. A critical and apparent disadvantage of this fixed-length context vector design is its inability to remember long sentences, which can cause a lot of context information to be lost in this process. Due to these limitations [6], this idea was extended by allowing a model called the Attention Mechanism (AM) to automatically search for parts of a source sentence that are relevant to predict a target word. On the other hand, a transformer was proposed in [7], which is the first sequence transduction model entirely based on attention, replacing the recurrent layers, which are most commonly used in encoder-decoder architectures, with multiheaded self-attention. Many researchers have worked on AM [8, 9]. On the other hand, a CNN was first introduced to NMT in [3], which was unsuccessful due to its limited receptive field. In [10], a solution was proposed to increase the receptive field through dilation. Another solution is to reduce the computations involved in CNN [11]. State-of-the-art performance has been achieved in multilingual NMT systems [12, 13]. On the other hand, most research on NMT relies on word-level translation [14] despite its weaknesses, especially with Out-Of-Vocabulary (OOV) and rare words. On the contrary, character-level NMT models are very effective in mitigating or vanishing the OOV problem [15, 16]. Methods to construct powerful encoder-decoder models can be classified into three categories: Recurrent Neural Network (RNNs, LSTMs, GRUs) based methods [6, 17], CNN-based methods [10, 11], and Attention Network (SAN) based methods [7, 18].

Many studies incorporated different linguistic knowledge sources into baseline NMT systems [19, 20], both in pre- and

post-processing [21, 22], to improve performance using segmentation and character-level encoding. Character-level NMT has become the new tendency in NMT. In [23], a character-based hybrid NMT model was proposed, combining both RNN and CNN. This model was trained on a small portion of the TED parallel corpus and its results in English-to-Arabic translation were improved by 10 BLEU points, while the word-based NMT model failed to train. In [14], a system was proposed to convert Arabic scripts to sub-word units and deal with the OOV problem on the MT task between Arabic and Chinese using various segmentation methods. This approach effectively tackled the OOV problem and improved the translation quality by up to 4 BLEU points. In [24], a hierarchical decoding method was proposed for NMT, considering both words and characters for the English-to-Arabic translation task, reporting an increase of up to 1.3 BLEU points over the baseline BPE subword-based NMT model. In [25], automatic Arabic translation was introduced using paraphrases with NMT, employing a bilingual corpus, where a paraphrase is an alternative surface form in the same language describing the same semantic content as the original. In [26], several tests were performed regarding the best possible morphological language-related representations.

More than 400 million people around the world speak the Arabic language, making it one of the five most used languages. Unfortunately, the peculiarities of Arabic, such as the form of characters, the morphology of words, and the richness and complexity of synonyms and rare words, represent a significant challenge to NMT, especially in solving the OOV and alignment problems. This study proposes a character-level hybrid AM model for Arabic-English NMT to improve translation performance by resolving the problems of rare words and word alignment. This is achieved by encoding the character, the Arabic word segmentation, and the PoS-Tag information in parallel with a multichannel encoder and an independent decoder with hybrid attention, which outputs the target language sentence. The model uses a Bi-LSTM network, CNN, and character-level representation.

II. BACKGROUND

A. Convolutional Neural Networks (CNNs)

CNN is a type of feedforward neural network designed to process array-formatted data [27] in various areas of machine learning, including image recognition, image classification [28], NLP [29], and medical image analysis [30], demonstrating significant efficiency. CNNs function by extracting features from inputs and processing them through multiple hidden layers, constituting the architecture of any CNN, with the ultimate objective of classification.

B. Long Short-Term Memory (LSTM)

LSTM was introduced to provide a solution to the vanishing gradient problem by making the weights on the network conditional on the context of the sequence input instead of fixing it. An input gate, an output gate, and a forget gate make up a typical LSTM unit. These three gates control the flow of information into and out of the cell, which remembers values across arbitrary time intervals. LSTMs have been applied to several areas of machine learning, such as speech recognition [31], machine translation [4], image captioning [32], and parsing [33].

C. Attention Mechanism (AM)

Attention is a popular concept and a useful tool in deep learning in recent years, proposed to solve the issue of memorizing long source sentences in NMT. The attention-based encoder-decoder framework [6] can be briefly described as consisting of a bidirectional RNN as an encoder to summarize not only the preceding but also the upcoming words. Suppose having two sequences X and Y of lengths n and m, where X is the input sequence and Y is the output sequence. A bidirectional RNN consists of a forward and a backward RNN, where each one reads the input sequence as it is ordered and calculates a sequence of forward and backward hidden states. Then, the annotation for each word x_t is obtained by concatenating the forward and the backward hidden states. In this way, the annotation h_j contains the summaries of both the preceding and following words, where this sequence of annotations is used by the decoder and the alignment model to compute the context vector. In the decoder architecture, each conditional probability is defined as [6]:

$$P\ (y_i|y_1, \dots\ ,\ y_{i-1},\ x)=g\ (y_{i-1}, s_i, c_i) \quad (1)$$

The decoder network has a hidden state s_i, calculated as

$$s_i = f(s_{i-1}, y_{i-1}, c_i) \quad (2)$$

where c_i is the context vector for the output y_t, which is calculated as:

$$c_i = \sum_{j=1}^{Tx} a_{i,j}\ h_j \quad (3)$$

where $a_{i,j}$ is the weight of each annotation, h_j calculated as

$$a_{i,j} = \frac{exp^{e_{i,j}}}{\sum_{t=1}^{Tx} exp^{e_{i,t}}} \quad (4)$$

where:

$$e_{i,j} = v_a tanh(Z_a s_{i-1} + U h_j\ ) \quad (5)$$

where v_a is the weight matrix to be learned in the alignment model.

III. PROPOSED MODEL

Figure 1 shows the proposed model. Since the approach depends on the segmentation of the word and POS-Tag as an addition to the source word, they are represented as inputs and pass through the encoder layers until they arrive at the proposed hybrid AM as a context vector, where the hybrid AM builds a context representation of the sequence.

A. Part of Speech Tagging (PoS-Tag)

PoS-Tag is the process of assigning the morphosyntax of words. The principal challenges for this task are the ambiguity (when a word can take several possible tags) and the problem of OOV words (in particular, words that did not appear in the training examples) [34]. Table I contains the tags that mark the core part-of-speech categories used in the Stanza library.

TABLE I. CORE PART OF SPEECH (POS) TAGS

PoS-Tags	PoS-Tags coding shape
ADJ	Adjective
ADP	Ad position
ADV	Adverb
AUX	Auxiliary verb
CCONJ	Coordinating conjunction
DET	Determiner
INTJ	Interjection
NOUN	Noun
NUM	Numeral
PART	Particle
PRON	Pronoun
PROPN	Proper noun
PUNCT	Punctuation
SCONJ	Subordinating conjunction
SYM	Symbol
VERB	Verb
X	ForeignW

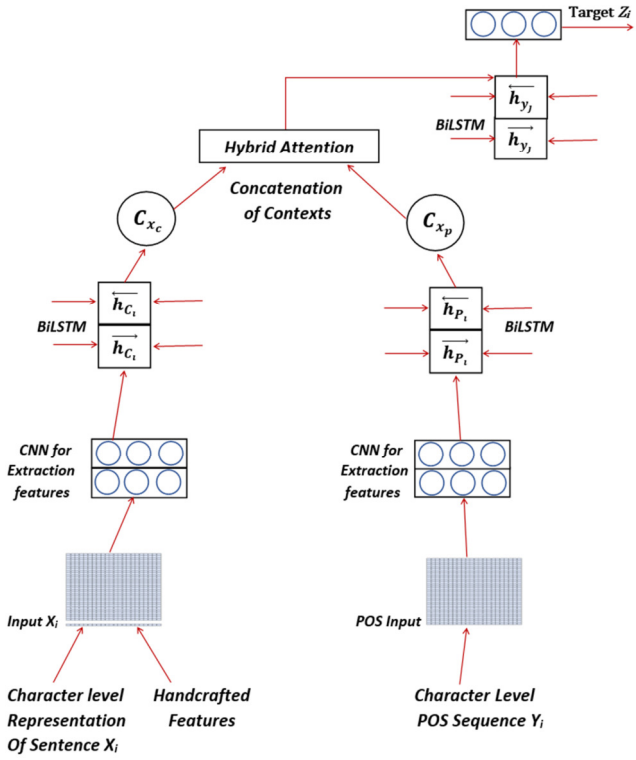

Fig. 1. The proposed model.

B. Word Representation

Input representations are constructed by dividing the input into three parts. The first is the character-level representation [35]

$$C = L_1, L_2, \ldots\ldots, L_N \quad (6)$$

$$L_1 = \begin{matrix}1\\ \vdots\\ 0\end{matrix} \quad L_2 = \begin{matrix}0\\ \vdots\\ 0\end{matrix} \quad L_N = \begin{matrix}0\\ \vdots\\ 1\end{matrix} \quad (7)$$

where C is the character-level matrix and the L vector represents the character in the word. Second, the PoS-Tag is obtained as words. For example, 'DET' = 'determiner'. The shape of the two encoder inputs is represented as follows:

X = [Corpus size, max sentence length, max word length, vocabulary size]

Y = [Corpus size, max POS sequence length, max word length, vocabulary size]

Max sentence length = Max POS sequence length + 1.

In the case of adding HCF, the third part is concerned with the segmentation of words as handcrafted features, where the vector is an N-hot encoding vector J_x. In this case, the vector consists of 7 attributes, as shown in Table II.

TABLE II. HANDCRAFTED FEATURES

Feature	Type
The word length	Integer
The stem length	Integer
The root length	Integer
The suffix length	Integer
The prefix length	Integer
The suffix end position	Integer
The prefix end position	Integer

C. Proposed Encoders

The proposed model depends on the multichannel principle, where the model shown in Figure 1 relies on two encoders. The first is for encoding the word and the second is for the "PoS-Tag + handcrafted features" encoding. The two sequences pass through a CNN layer to extract the properties and then pass through a Bi-LSTM layer, which gives two context representations: C_{X_c} and C_{X_p}, respectively.

D. Proposed Decoder with Hybrid Attention

In hybrid attention, both content-based and location-based AMs were used, which were combined to leverage the benefits of both approaches. By incorporating both content and positional information, the model can effectively attend to relevant parts of the input sequence based on their content and relative positions, leading to improved performance in various sequence-to-sequence tasks.

There is a difference between [6] and the proposed model. The classic character-level NMT reads the encoder's hidden variables and estimates the target sequence during decoding. In the proposed model, C_{X_p} and C_{X_c} are calculated as follows:

$$C_{X_c} = \sum_{i=1}^{T_x} h\, c_i \quad (8)$$

$$C_{X_p} = \sum_{i=1}^{T_w} h\, p_i \quad (9)$$

Then, C_{X_c} and C_{X_p} are concatenated to create one context representation C_X.

The second difference is in the AM, where the decoder's hidden state of the network H is calculated as:

$$h_t^{att} = tanh(Z_{hh}^{att} h_{t-1}^{att} + Z_{xh}^{att} W_t) \quad (10)$$

where the context vector C_x for the output Y_t is calculated as:

$$C_{xt} = \sum_{s=1,n} a_{ts} h_s \quad (11)$$

where a_{ts} is the weight of each annotation h_t, computed as:

$$a_{ts} = softmax(\ score\ [h_t, h_t^{out}]) \quad (12)$$

where:

$$score\ [h_t, h_t^{out}] = 1 - \frac{h_t\ h_t^{out}}{h_t + h_t^{out} - h_t\ h_t^{out}} \quad (13)$$

where:

$$h_t^{att+} = sigmoid(Z_{ch}^{att} C_{xt} + W_{hh}^{att} h_t^{att}) \quad (14)$$

The output of the decoder is calculated as

$$O_t = Softmax(\ Z_{hz}^{att} h_t^{att+}) \quad (15)$$

Binary cross entropy is used as the loss function.

$$Loss = -\frac{1}{N} \sum_{i=1}^{N} y_i\ .$$

$$log\big(p(y_i)\big) + (1 + y_i)\ .\ log\big(1 - p(y_i)\big) \quad (16)$$

where y_i represents the actual class and $log\big(p(y_i)\big)$ is its probability.

IV. EXPERIMENTS

A. Dataset

A part of the OPUS-100 dataset [36] was used to train and test the proposed model. The dataset consists of 1 million pairs of sentences. The corpus was divided into an 80:20 ratio for training and testing. Tables III and IV show details on the dataset and some samples.

TABLE III. DATASET USED ON SENTENCES

	Number of sentence pairs	Number of tokens	Max sequence length	Vocabulary size
Source Language	~727k	~4.0M	15	30
Target Language	~727k	~5.7M	15	30
PoS-Tag sequence	~727k	~3.9M	15	30

TABLE IV. SAMPLES FROM THE DATASET

Arabic	English	PoS-Tag
وهذه؟	And this?	CoordinatingConj Determiner
لقد كان	It was, um...	Adverb Verb
ما الذي تفعله هناك؟	What is she doing here?	Determiner Verb Pronoun Adverb

For data extraction, the Stanza library was used to tag Arabic words, and the pyarabic lib was used to extract features from them. The BLEU metric was used to assess translation quality.

B. Results

Python and TensorFlow were used for coding and deep learning development. Table V details the setup, outlining the hardware and software configurations employed. Table VI presents the experimental results, using a consistent model architecture across all trials, with variations only in the architecture.

TABLE V. LAB ENVIRONMENT

Software and hardware	Configuration
GPU	A100 80 GB VRAM
CPU	8v CPU
RAM	62 GB RAM
OS	Linux
DE	Python

TABLE VI. BLEU RESULTS

Models	BLEU scores
CNN-BiLSTM Encoder-Decoder (only words) +Hybrid Attention	19.12
CNN-BiLSTM 2Encoder-Decoder (words + PoS-Tag) +Hybrid Attention	19.77
CNN-BiLSTM 2Encoder-Decoder (words+PoS-Tag+HCF) +Hybrid Attention	19.87

C. Discussion and Analysis

This study builds on the existing body of research by addressing key challenges in NMT, specifically focusing on Arabic-English translation. The primary contributions of this work are threefold:

- Character-Level Representation: Unlike traditional word-level models, the proposed approach uses character-level encoding to handle the intricacies of the Arabic language, including its rich morphology and the problem of rare words. This level of granularity helps in better capturing the sub-word structures, which is crucial for translating languages with complex word formations.
- Multi-Channel Encoder with Handcrafted Features: A multichannel CNN-BiLSTM encoder was employed to process character-level representations alongside handcrafted features, such as word segmentation and PoS-Tag. This combination allows the model to utilize both raw character data and linguistic information, resulting in a more comprehensive understanding of the source text.
- Hybrid AM: The proposed model incorporates a hybrid AM that blends global and local attention strategies. This enables the decoder to focus on relevant parts of the input sequence more effectively, thus improving the alignment and translation quality. This study parallels the work in [16], which also uses a hybrid AM to improve machine translation.

Previous studies [24] explored character-level models, but this implementation uniquely combines this with additional linguistic features. The experimental results, derived from testing on a subset of the OPUS-100 dataset, underscore the effectiveness of the proposed model, which can be attributed to the harmonic integration of character-level information and handcrafted features. The addition of these features makes the model more resilient to common translation issues, such as rare words and alignment issues.

The models were trained using loss binary cross-entropy, Adam optimizer, stochastic gradient descent, dropout = 0.2, and mini-batches of size 100 for improved generalization, where each training instance consisted of one Arabic sentence and one English sentence. The translation quality of the model

was quantitatively evaluated in terms of BLEU. Table VII presents results and comparisons of the proposed with other similar approaches.

TABLE VII. BLEU SCORES FOR NMT METHODS ON VARIOUS TASKS

Approach	BLEU	Dataset	Descriptions
[23] AR → EN	18.67	IWSLT16	Character NMT method that incorporates RNN and CNN
[24] AR → EN	12.72	IWSLT16	Character NMT method that incorporates RNN and Bi RNN and Attention
Proposed model AR → EN	19.87	OPUS-100	CNN-BiLSTM 2Encoder-Decoder (words + PoS-Tag+HCF) + Hybrid Attention
[25] AR → Ch	19.63	UN corpus	Word-level NMT using GRUs

V. CONCLUSION

This study proposed a novel character-level hybrid attention model for Arabic-English machine translation. This approach integrates character-level representation, Arabic word segmentation, and PoS-Tag information, processed in parallel using a multichannel encoder. This encoder utilizes both CNN and Bi-LSTM, while the decoder operates independently with a hybrid AM to generate target language sentences. The motivation behind the proposed model lies in addressing significant challenges in NMT, particularly those related to rare words, word alignment, and the need for extensive training data. Experimental results in a subset of the OPUS-100 dataset demonstrated the efficacy of the proposed model. Character-level information, combined with PoS-Tag and word segmentation, proved to be both compatible and complementary, leading to competitive performance improvements. The hybrid AM, which merges aspects of both global and local attention, was shown to effectively capture context and dependencies within sentences, thus enhancing translation quality. In conclusion, the proposed character-level hybrid attention model represents a significant step forward in Arabic-English machine translation. Future work should focus on expanding and refining these techniques, developing richer datasets, and exploring new word representation methods to further advance the field of NMT. Future work will also focus on multimodal text-image Arabic NMT.

REFERENCES

[1] P. Koehn, F. J. Och, and D. Marcu, "Statistical phrase-based translation," in Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - NAACL '03, Edmonton, Canada, 2003, vol. 1, pp. 48–54, https://doi.org/10.3115/1073445.1073462.

[2] D. Chopra, N. Joshi, and I. Mathur, "A Review on Machine Translation in Indian Languages," *Engineering, Technology & Applied Science Research*, vol. 8, no. 5, pp. 3475–3478, Oct. 2018, https://doi.org/10.48084/etasr.2288.

[3] N. Kalchbrenner and P. Blunsom, "Recurrent Continuous Translation Models," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle, WA, USA, Oct. 2013, pp. 1700–1779.

[4] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to Sequence Learning with Neural Networks," in *Advances in Neural Information Processing Systems*, 2014, vol. 27.

[5] K. Cho *et al.*, "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, 2014, pp. 1724–1734, https://doi.org/10.3115/v1/D14-1179.

[6] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate." arXiv, May 19, 2016, https://doi.org/10.48550/arXiv.1409.0473.

[7] A. Vaswani *et al.*, "Attention Is All You Need." arXiv, Aug. 01, 2023, https://doi.org/10.48550/arXiv.1706.03762.

[8] T. Luong, H. Pham, and C. D. Manning, "Effective Approaches to Attention-based Neural Machine Translation," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, 2015, pp. 1412–1421, https://doi.org/10.18653/v1/D15-1166.

[9] J. Cheng, L. Dong, and M. Lapata, "Long Short-Term Memory-Networks for Machine Reading." arXiv, Sep. 20, 2016, https://doi.org/10.48550/arXiv.1601.06733.

[10] N. Kalchbrenner, L. Espeholt, K. Simonyan, A. van den Oord, A. Graves, and K. Kavukcuoglu, "Neural Machine Translation in Linear Time." arXiv, Mar. 15, 2017, https://doi.org/10.48550/arXiv.1610.10099.

[11] L. Kaiser, A. N. Gomez, and F. Chollet, "Depthwise Separable Convolutions for Neural Machine Translation." arXiv, Jun. 15, 2017, https://doi.org/10.48550/arXiv.1706.03059.

[12] Y. Wu *et al.*, "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation." arXiv, Oct. 08, 2016, https://doi.org/10.48550/arXiv.1609.08144.

[13] B. Zhang, P. Williams, I. Titov, and R. Sennrich, "Improving Massively Multilingual Neural Machine Translation and Zero-Shot Translation." arXiv, Apr. 24, 2020, https://doi.org/10.48550/arXiv.2004.11867.

[14] F. Aqlan, X. Fan, A. Alqwbani, and A. Al-Mansoub, "Arabic–Chinese Neural Machine Translation: Romanized Arabic as Subword Unit for Arabic-sourced Translation," *IEEE Access*, vol. 7, pp. 133122–133135, 2019, https://doi.org/10.1109/ACCESS.2019.2941161.

[15] J. Chung, K. Cho, and Y. Bengio, "A Character-Level Decoder without Explicit Segmentation for Neural Machine Translation." arXiv, Jun. 20, 2016, https://doi.org/10.48550/arXiv.1603.06147.

[16] F. Wang, W. Chen, Z. Yang, S. Xu, and B. Xu, "Hybrid Attention for Chinese Character-Level Neural Machine Translation," *Neurocomputing*, vol. 358, pp. 44–52, Sep. 2019, https://doi.org/10.1016/j.neucom.2019.05.032.

[17] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Aug. 1997, https://doi.org/10.1162/neco.1997.9.8.1735.

[18] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens, "Stand-Alone Self-Attention in Vision Models," in *Advances in Neural Information Processing Systems*, 2019, vol. 32.

[19] S. Ding, A. Renduchintala, and K. Duh, "A Call for Prudent Choice of Subword Merge Operations in Neural Machine Translation." arXiv, Jun. 24, 2019, https://doi.org/10.48550/arXiv.1905.10453.

[20] D. Ataman, W. Aziz, and A. Birch, "A Latent Morphology Model for Open-Vocabulary Neural Machine Translation." arXiv, Feb. 26, 2020, https://doi.org/10.48550/arXiv.1910.13890.

[21] H. Sajjad, F. Dalvi, N. Durrani, A. Abdelali, Y. Belinkov, and S. Vogel, "Challenging Language-Dependent Segmentation for Arabic: An Application to Machine Translation and Part-of-Speech Tagging." arXiv, Sep. 02, 2017, https://doi.org/10.48550/arXiv.1709.00616.

[22] M. Oudah, A. Almahairi, and N. Habash, "The Impact of Preprocessing on Arabic-English Statistical and Neural Machine Translation," in *Proceedings of MT Summit XVII,* Dublin, Ireland, Aug. 2019, vol. 1, pp. 214–221.

[23] E. H. Almansor and A. Al-Ani, "A Hybrid Neural Machine Translation Technique for Translating Low Resource Languages," in *Machine Learning and Data Mining in Pattern Recognition*, New York, NY, USA, Jul. 2018, pp. 347–356, https://doi.org/10.1007/978-3-319-96133-0_26.

[24] D. Ataman, O. Firat, M. A. Di Gangi, M. Federico, and A. Birch, "On the Importance of Word Boundaries in Character-level Neural Machine Translation." arXiv, Oct. 21, 2019, https://doi.org/10.48550/arXiv.1910.06753.

[25] M. Alkhatib and K. Shaalan, "Paraphrasing Arabic Metaphor with Neural Machine Translation," *Procedia Computer Science*, vol. 142, pp. 308–314, Jan. 2018, https://doi.org/10.1016/j.procs.2018.10.493.

[26] Y. Belinkov, N. Durrani, F. Dalvi, H. Sajjad, and J. Glass, "What do Neural Machine Translation Models Learn about Morphology?," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada, 2017, pp. 861–872, https://doi.org/10.18653/v1/P17-1080.

[27] Y. LeCun *et al.*, "Handwritten Digit Recognition with a Back-Propagation Network," in *Advances in Neural Information Processing Systems*, 1989, vol. 2.

[28] J. Yim, J. Ju, H. Jung, and J. Kim, "Image Classification Using Convolutional Neural Networks With Multi-stage Feature," in *Robot Intelligence Technology and Applications 3*, 2015, pp. 587–594, https://doi.org/10.1007/978-3-319-16841-8_52.

[29] R. Collobert and J. Weston, "A unified architecture for natural language processing: deep neural networks with multitask learning," in *Proceedings of the 25th international conference on Machine learning*, Helsini, Finland, Apr. 2008, pp. 160–167, https://doi.org/10.1145/1390156.1390177.

[30] S. S. Yadav and S. M. Jadhav, "Deep convolutional neural network based medical image classification for disease diagnosis," *Journal of Big Data*, vol. 6, no. 1, Dec. 2019, Art. no. 113, https://doi.org/10.1186/s40537-019-0276-2.

[31] A. Graves, "Generating Sequences With Recurrent Neural Networks." arXiv, Jun. 05, 2014.

[32] R. Kiros, R. Salakhutdinov, and R. Zemel, "Multimodal Neural Language Models," in *Proceedings of the 31st International Conference on Machine Learning*, Jun. 2014, pp. 595–603.

[33] O. Vinyals, Ł. Kaiser, T. Koo, S. Petrov, I. Sutskever, and G. Hinton, "Grammar as a Foreign Language," in *Advances in Neural Information Processing Systems*, 2015, vol. 28.

[34] D. E. Messaoudi, D. Nessah, and A. Siam, "Intelligent system for part-of-speech tagging using convolutional neural network on arabic language," in *The 2nd International Conference on Distributed Sensing and Intelligent Systems (ICDSIS 2021)*, Jul. 2021, vol. 2021, pp. 207–219, https://doi.org/10.1049/icp.2021.2677.

[35] X. Zhang, J. Zhao, and Y. LeCun, "Character-level Convolutional Networks for Text Classification," in *Advances in Neural Information Processing Systems*, 2015, vol. 28.

[36] "OPUS-100 Corpus." [Online]. Available: https://opus.nlpl.eu/opus-100.php.