From Analysis to Implementation: A Comprehensive Review for Advancing Arabic-English Machine Translation

Aichetou Mohamed Sidiya
Computer Science Department,
Effat College of Engineering,
Effat University
Jeddah, Saudi Arabia
asidiya@effat.edu.sa

Hanin Alzaher
Computer Science Department,
Effat College of Engineering,
Effat University
Jeddah, Saudi Arabia
habalzaher@effat.edu.sa

Razan Almahdi
Computer Science Department,
Effat College of Engineering,
Effat University
Jeddah, Saudi Arabia
raoalmahdi@effat.edu.sa

Passant Elkafrawy Ph.D
Effat College of Engineering,
Energy and Technology Research Center
Effat University
Jeddah, Saudi Arabia
pelkafrawy@effatuniversity.edu.sa

***Abstract*—In an increasingly interconnected world, the demand for accurate Arabic-English translation has surged, highlighting the complexities in handling Arabic's intricate morphology and diverse linguistic structures. This research delves into various translation models, including Convolutional Neural Networks (CNNs), LSTM, Neural Machine Translation (NMT), BERT, and innovative fusion architectures like the Transformer-CNN. Each model's strengths and limitations are scrutinized through comprehensive evaluations and comparisons, unveiling their potential to address translation challenges. The research then builds two models, the first based on LSTM and the second on BERT, and tests their performance in translating English to Arabic. The paper then conducts an in-depth analysis of the results. The comparative analysis provides insights into the landscape of Arabic-English translation models, guiding future research toward refining models, leveraging diverse datasets, and establishing standardized evaluation benchmarks to bridge existing gaps.**

Keywords—Arabic-English translation, Neural Machine Translation (NMT), LSTM, BERT

I. Introduction

In today's globalized world, the demand for accurate and nuanced translation between languages is ever-increasing, particularly in bridging the gap between Arabic and English. Arabic, a language with a rich history and complex linguistic structure, poses significant challenges for machine translation systems due to its morphology, syntax, and diverse regional variations. Meanwhile, English stands as a widely used language in academia, business, and global communication, necessitating proficient translation systems to facilitate seamless communication between the two languages [1][2].

Over the years, researchers have explored various approaches to Arabic-English translation, from rule-based systems and statistical machine translation (SMT) to the more recent advancements in neural machine translation (NMT) and attention-based models [3][4]. Machine translation models can be divided into three main categories: (1) rule-based, (2) data-driven, (3) hybrid. The division depends on the data source, for rule-based they depend on linguistic information such as semantics, morphological analysis, and human-crafted rules. Data-driven, on the other hand, uses algorithms and large language corpora, and hybrid combines both approaches [3]. While these models have shown remarkable progress in addressing translation challenges, the need for further enhancement in accuracy, context preservation, and fluency, especially in handling literary texts, remains a critical area of exploration [1].

The objectives of this research encompass a review of existing translation models and an assessment of their capabilities in handling the nuances of Arabic grammar and syntax. The goal of the paper is to leverage this analysis in testing the performance of the models on an Arabic-English dataset.

The rest of the paper is structured as follows: section 2 literature review, section 3 methodology, section 4 results and discussions, and section 5 conclusion.

II. Literature Review

A. Models Overview

With over 300 million native speakers, Arabic is one of the world's five most spoken languages and one of the United Nations' (UN) six official languages. It is a Semitic language with a rich and complex morphology that differs significantly from that of Indo-European languages (such as English and French). The addition of Arabic morphology to other linguistic aspects has made automatic translation from and into Arabic much more difficult [1].

Alhawarat and Aseeri [5] have developed a novel Multi-Kernel Convolutional Neural Network (CNN) model called A Superior Arabic Text Categorization Deep Model (SATCDM) Figure 1 for categorizing Arabic news documents. The primary problem tackled by the paper involves the efficient classification of the surging influx of Arabic text documents, encompassing web pages, news articles, and social media content. The model uses n-gram word embedding within a Multi-Kernel CNN architecture and outperforms conventional methods like bag-of-words or TF-IDF weighting in Arabic text categorization. The model's

979-8-3503-9356-9/24/$31.00 ©2024 IEEE

accuracy ranges from 97.58% to 99.90% across 15 datasets, making it a valuable tool for various industries. However, the paper has limitations, including limited comparative analysis and lack of in-depth discussion about the datasets used. The findings suggest further enhancement of the SATCDM model and exploring diverse datasets for Arabic text categorization.

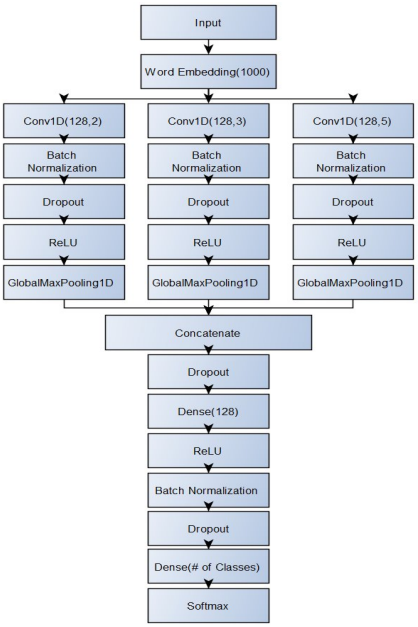

Figure 1. SATCDM Architecture [2].

Lulu and Elnagar's [6] study explored the challenges of Arabic dialectal languages in social media, focusing on the application of deep learning models for automatic classification. They use the Arabic Online Commentary dataset, which includes Egyptian (EGP), Levantine (LEV), and Gulf dialects (GLF). The research addresses the limitations of standard Natural Language Processing (NLP) tools when dealing with informal dialectal textual data, and its implications for NLP applications like sentiment analysis and machine translation. The study employs four deep learning models: long-short-term memory (LSTM), convolutional neural networks (CNN), bidirectional LSTM (BLSTM), and convolutional LSTM (CLSTM) for automatic classification of Arabic dialectal text. The findings demonstrate varying yet promising performances of the models for each dialect pair, illustrating the potential efficacy of deep learning techniques in classifying Arabic dialectal text. However, limitations, include a narrow focus on classification experiments and reliance on the AOC dataset, raising concerns about its representativeness and completeness of dialectal nuances. The findings suggest the need for further refinement and review of benchmark datasets, linguistic analysis, and the enhancement of deep learning models for Arabic dialectal text to create more accurate and robust NLP tools for social media content [13].

Gamal et al. [7] have developed a novel Neural Machine Translation (NMT) model for Arabic-English translation, addressing the scarcity of resources in this field. The paper's approach involves the development of an innovative NMT model that combines a transformer architecture with multi-head attention mechanisms. This fusion optimizes translation accuracy, as validated by empirical evaluations showcasing impressive performance metrics: notably high accuracy (97.68%), low loss (0.0778), and a near-perfect Bilingual Evaluation Understudy (BLEU) score of 99.95%. The model's strengths lie in its groundbreaking architecture and ability to address resource scarcity. However, the paper could benefit from a more comprehensive discussion of datasets and evaluation methodologies, comparative analysis against other models, and future research on enhancing evaluation methods and extending the model's success to other low-resource languages.

Baniata et al. [8] introduce a novel Reverse Positional Encoding Multi-Head Attention (RPE MHA) Neural Machine Translation (NMT) model specifically designed for translating Arabic dialects. The model addresses the linguistic challenges of Arabic, focusing on suffixes, affixes, complex sentence structures, and the free-word order. It uses subword embeddings, reverse positional encoding, and transformer architecture to improve translation accuracy. The study's main strength lies in its introduction of this groundbreaking solution, which effectively identifies and addresses the complex linguistic challenges of Arabic. The model's efficacy and practical utility are evaluated using quantitative and qualitative measures, suggesting potential real-world applicability. However, the study has limitations, including a lack of specifics about model architecture, training methodologies, and hyperparameters, and a lack of detailed information about the dataset. Additionally, the over-reliance on BLEU scores and human evaluations without integrating additional external metrics restricts a holistic evaluation of the model's performance[9].

Ahammad et al.'s [10] study on deep learning algorithms and encoding-decoding strategies aims to improve machine translation systems' performance. They highlight the limitations of traditional sequential encoder-decoder frameworks and the importance of syntactic information in improving translation quality. The paper's strengths lie in its focus on Neural Machine Translation (NMT) and deep learning algorithms, offering valuable insights. However, limitations include the lack of specific methodologies and empirical results, and the absence of experimental evaluations. The findings suggest exploring innovative approaches that integrate syntactic information to improve translation quality, prompting further research on novel methodologies and advancements in NMT.

Abdurahimov's [11] paper explores the use of BERT (Bidirectional Encoder Representations from Transformers) to enhance machine translation accuracy and reliability between English and Arabic. The paper addresses the challenges faced in English-Arabic machine translation, such as the rich morphology and syntactic differences in Arabic [2]. The authors use BERT as the baseline model, preprocessing data using Unicode normalization, orthographic normalization, diacritization for Arabic, and lowercase and normalized punctuation for English. BPE tokenization is used for both languages. The paper contributes by leveraging BERT as the baseline model for English-Arabic machine translation, providing detailed preprocessing steps and tokenization techniques for

reproducibility. A comprehensive evaluation is included, comparing the performance of the BERT-fused model with baseline models on separate test sets. However, the paper lacks specific details about the experimental setup, including architecture and hyperparameter settings. The findings suggest that leveraging BERT can lead to improved machine translation performance for the English-Arabic language pair [12].

Bensalah et al. [13] introduce a novel Deep Learning architecture, "The Transformer-CNN," for improving Arabic-English Machine Translation. The authors aim to enhance translation quality and performance by combining Convolutional Neural Networks (CNNs) with the transformer model. They emphasize the importance of preprocessing Arabic sentences using the Farasa segmenter and BPE tokenizer. The paper addresses the challenges in Arabic Natural Language Processing (NLP) tasks, particularly in Machine Translation, due to its rich vocabulary and complex morphology. The proposed architecture combines CNNs and the transformer model to leverage their respective strengths. Experimental results show that the proposed approach outperforms previous state-of-the-art Arabic MT systems. However, the paper lacks extensive analysis or comparison with other existing approaches, limited information about the dataset used for the experiments, and a comprehensive discussion of potential limitations or implementation challenges.

A. Model Gap Analysis

While there are similarities in the emphasis on the complexity of Arabic, the effectiveness of deep learning models, and the need for specialized techniques, there are also notable distinctions in the reviewed studies. For example, Alhawarat and Aseeri's study concentrates on Arabic text categorization using the Multi-Kernel Convolutional Neural Network (CNN) model, addressing the challenge of efficiently classifying Arabic news documents [4]. On the other hand, Lulu and Elnagar's research delves into the obstacles of Arabic dialectal text classification and highlights the limitations of standard Natural Language Processing (NLP) tools. They utilize various deep learning models, such as LSTM, CNN, bidirectional LSTM (BLSTM), and convolutional LSTM (CLSTM), to classify Arabic dialectal text and underscore the necessity for accurate sentiment analysis and machine translation in this particular context.

Moreover, Gamal et al. and Baniata et al. concentrate on Neural Machine Interpretation (NMT) models particularly planned for Arabic-English interpretation. Gamal et al. propose a transformer-based NMT show with consideration instruments, tending to the shortage of assets within the Arabic interpretation space. On the other hand, Baniata et al. introduce a reverse positional encoding multi-head attention (RPE MHA) NMT model that focuses on translating Arabic dialects, taking into account the linguistic challenges posed by suffixes, affixes, complex sentence structures, and free-word order [13].

This literature review discussed the challenges and progress of Arabic-English translation models, highlighting the complexity of Arabic morphology, regional variations, and linguistic structures [2]. Deep learning models like CNNs, LSTM, and Transformer have shown significant improvement in translation accuracy and performance. Innovative techniques like multi-head attention mechanisms, reverse positional encoding, and n-gram word embedding have also improved translation quality. However, studies vary in their focus, including text categorization, dialectal text classification, and NMT models [2]. Despite similarities, there are disparities in proposed techniques and areas for improvement.

B. Model Performance Analysis

The evaluation of machine translation models for Arabic-English language pairs yielded varying degrees of accuracy and translation quality. The SATCDM model demonstrated high accuracy, achieving an impressive range of 97.58% to 99.90% in translating Arabic to English [5]. Its performance surpassed traditional methods, showcasing its effectiveness in accurately translating between the two languages.

In contrast, the LSTM model achieved an accuracy of 71.4% with a cross-validation accuracy of 84.5%[6], showing moderate proficiency in handling Arabic-English translation tasks. Although LSTM models excel in capturing long-term dependencies in sequential data, specific challenges faced by LSTM in accurately translating Arabic to English require further investigation for potential improvements. Furthermore, the CNN model obtained an accuracy of 68.0%, and significantly a cross-validation accuracy of 96.0% in Arabic-English translation [6]. Despite being primarily known for image processing tasks, CNNs exhibited promising capabilities in capturing relevant textual features and patterns, contributing to accurate translations in Arabic. The BLSTM and CLSTM models achieved accuracies of 70.9% and 71.1%, respectively [6], with cross-validation going as high as 84.4% and 90.4% respectively, showcasing their effectiveness in handling Arabic text translation. The bidirectional approach of the BLSTM and the hybrid architecture of the CLSTM combining CNNs and LSTMs illustrate their potential for capturing dependencies and local patterns in Arabic sentences.

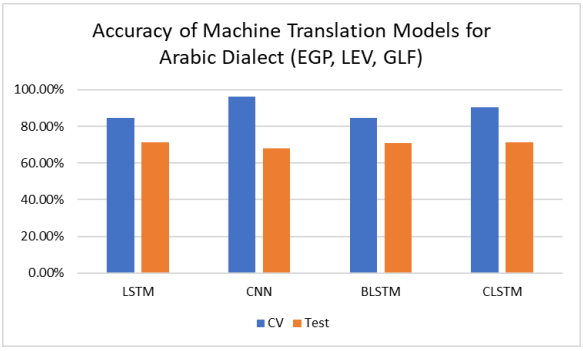

Figure 2. Accuracy of Machine Translation Models on Arabic Dialect (EGP, LEV, GLF) [3].

Conversely, the NMT model, showcasing an accuracy of 97.68% and an impressive BLEU score of 99.95, outperformed other models in translation quality [7][12], including RPE MHA-based NMT which scored 66.87 in BLEU [8], and transformer CNN which scored 68.60 [13][14].

III. Methodology

This research aims to build two machine translation models, to test their effectiveness on a single dataset. The following section provides a detailed overview of the models and the dataset.

A. Dataset

The dataset used in this research is a text file containing up to 20 thousand records of English text mapped to its Arabic translation. The data set consists of everyday sentences and snippets from news articles. The data was extracted from Kaggle: https://www.kaggle.com/code/ahmedgamal12/english-arabic-nmt/input

Table 1. Sample of the Data

english	arabic
Hi.	.مرحبًا
Run!	!اركض
Help!	!النجدة
Jump!	!اقفز

B. LSTM

The first model we tested is an LSTM model [6], as shown in Figure 4 [13][14] is composed of an input layer of size 50, followed by an embedding layer, which produced 2606200 parameters. The next layer is an LSTM encoder of 256 units, this layer produced 731136 parameters. The fourth layer is the decoder composed of 256 units and outputting 787456 parameters. The last two layers are the dense layer and the output layer. The final Dense layer also included a softmax activation function. The model had in total 18991471 trainable parameters. The loss was calculated using sparse categorical cross-entropy, and Root Mean Squared Propagation (RMSProp) was used for optimization. The model was built in Google Colab for faster processing, and it was built using Keras tensorflow.

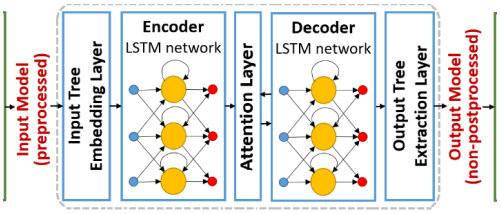

Figure 3. LSTM model structure [7]

This architecture of LSTM allows it to capture dependencies, and be able to process longer sequences.

C. BERT

The second model tested was a transformer model [7]. The transformer model as shown in Figure 4 is based on BERT [15] which is a bidirectional transformer. The benefit of this bidirectional approach is that instead of just processing words from left to right or right to left, BERT uses the context of the word by analyzing the previous and following words. This approach is particularly effective in translation from English to Arabic given that English is a left-to-right language and Arabic is right-to-left.

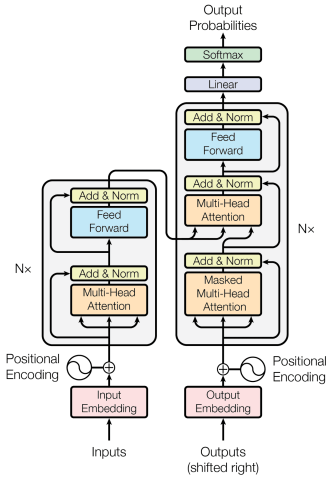

Figure 4. BERT model architecture [7].

In order to test BERT on the provided dataset, the MarianMT pre-trained model from Hugging Face was used. The model is a multilingual machine translation model based on transformer architecture, it employs a sequence-to-sequence approach, meaning it maps input sequences to the output sequences. The advantage of using this model to translate is its ease of use and its high performance, along with its versatility.

IV. Results and Discussion

A. Results Analysis

The following section provides a summary of the results as shown in Table 2. Moreover, Figures 5 and 6 provide a visual for the evolution of the loss and accuracy of LSTM between training and testing. The loss evolution is quite similar, while the accuracy seems to drop at the 4th epoch then increase again, suggesting discrepancies with the model, and the data.

The LSTM model demonstrated a notable accuracy of 72%, surpassing the BERT MarianMT model, which achieved an accuracy of 60.25%. It is crucial to highlight that the MarianMT model was solely employed for testing without fine-tuning, suggesting the potential for further performance improvement through fine-tuning.

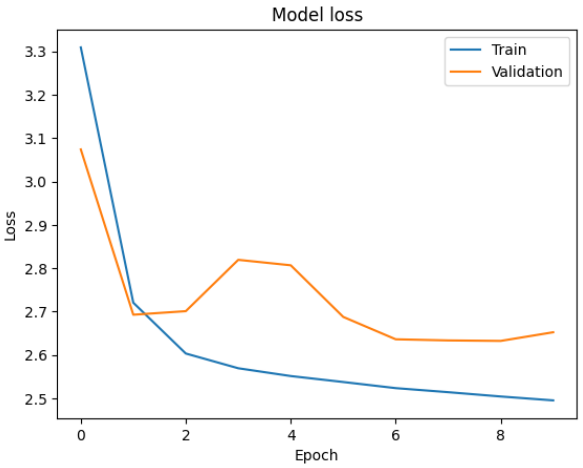

Figure 5. Model loss LTSM Train vs Validation

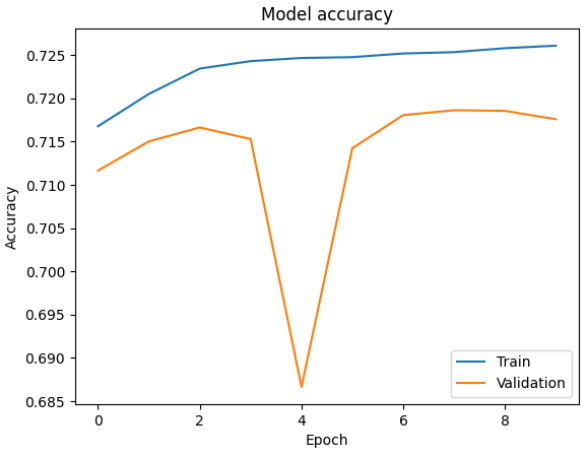

Figure 6. Model accuracy LTSM Train vs Validation

Additionally, the BLEU score analysis demonstrated the comparative effectiveness of the models. The MarianMT model, despite its low accuracy, produces a relatively high BLEU score of 29%, compared to the LSTM model which scored a BLEU of 0, emphasizing the nuanced nature of evaluation metrics and the importance of considering multiple criteria in model assessment.

Table 2. Results Summary

Metrics	LSTM	BERT MarianMT
Accuracy	0.72	0.60
BLEU	0	0.29
Error	2.55	0.69

A qualitative inspection of the results, as illustrated in Tables 3 and 4, reveals intriguing patterns. The BERT model exhibits the ability to predict entire sentences, although the predictions included repetitions of reference values, which suggested a potential area for refinement in handling redundancy. In contrast, the LSTM model results where only propositions, this not only underscores the model's distinctive methodology but also implies a nuanced variation in translation strategies compared to BERT. The emphasis on propositions could be indicative of a more granular understanding of language or a specific approach to translation tasks.

These differences observed between the BERT and LSTM models in our study may be linked to the data cleaning and preprocessing approach. The use of embeddings resulted in discrepancies in data final form, which affected the quality and integrity of the data it is trained on. Therefore, the nuances in the model behavior and performance.

Table 3. BERT Sample Results (add more)

english	Arabic	arabic_translation
Hi.	.مرحبًا	مرحباً.. مرحباً..
Run!	اركض!	!أركض! أركض !
Help!	النجدة!	النجدة، النجدة، المساعدة، المساعدة، المساعدة، ...
Let's begin.	.لنبدأ	لنبدأ
Tom is big.	توم كبير	.(توم) كبير

Table 4. LSTM Sample Results (add more)

english	arabic	arabic_ translation
"he's", 'my', 'brother'	'هو', 'أخي'	'أنا'
'at', 'the', 'end', 'of', 'the', 'day', 'riot', 'police', 'evicted', 'the', 'square', 'with', 'violence'	'في', 'نهاية', 'اليوم', ' 'افرغت', 'شرطة', 'مكافحة', 'الشغب', "الميدان', 'بالعنف',	'هل'
'stay', 'away', 'from', 'me'	'إبقَ', 'بعيداً', 'عني'	'أنا'
'an', 'indonesian', 'university', 'student', 'was', 'detained', 'and', 'charged', 'with', 'defaming', 'a', 'police', 'officer', 'after', 'uploading', 'a', 'video', 'of', 'the', 'office',	'تم', 'اعتقال', 'طالب', 'اندونيسي', 'وتوجيه', 'اتهامات', 'له', 'بالتشهير', 'بضابط', 'شرطة', 'بعد', 'ان', 'حمل', 'شريطا', 'مصورا',	'في'
'can', 'your', 'mother', 'drive', 'a', 'car'	'هل', 'تستطيع', 'أمك', 'أن', 'تقود', 'سيارة؟'	'هل'

While the LSTM model's performance is commendable, the detailed analysis of BERT's output suggests its potential

for capturing more comprehensive semantic structures. Fine-tuning the MarianMT model and exploring additional evaluation metrics could further enhance the robustness of the findings. This comparison highlights the trade-offs and strengths inherent in LSTM and BERT architectures, contributing valuable insights to the field of machine translation.

B. Future Work

Future research in Arabic-English machine translation should prioritize dataset expansion and diversification to include various Arabic dialects, thereby improving model robustness. Efforts to improve existing architectures or create new models tailored to the complexities of the Arabic language are essential. Standardizing evaluation metrics for Arabic-English translation nuances, as well as encouraging transparency in model documentation and methodologies, will improve reproducibility. Addressing the limitations of current state-of-the-art models, investigating linguistic complexities, and leveraging pre-trained models via transfer learning techniques are all critical pursuits. Furthermore, broadening research to include NLP tasks other than translation, encouraging collaboration among researchers, and emphasizing real-world applicability by engaging end users will all contribute to the development of more accurate and contextually adept Arabic-English machine translation systems.

V. CONCLUSION

This research delved into the landscape of state-of-the-art Arabic-English machine translation models, encompassing Convolutional Neural Networks (CNNs), LSTM, NMT, BERT, and a Transformer-CNN fusion architecture. Furthermore, the research utilized the explorative analysis to build and test two models, LSTM and BERT MarianMT, on a dataset comprising 20 thousand records of everyday English and Arabic translations sourced from Kaggle. The LSTM model, characterized by its sequential processing abilities, achieved a commendable accuracy of 72%, surpassing the BERT MarianMT model, which attained an accuracy of 60.25%. The BLEU score analysis underscored the nuanced nature of evaluation metrics, with the MarianMT model, despite its lower accuracy, exhibiting a relatively high BLEU score of 29% compared to the LSTM model's score of 0. Qualitative examination of results revealed that the BERT model demonstrated the capability to predict entire sentences, while the LSTM model predominantly produced individual propositions. Future research directions include expanding and diversifying datasets, refining existing models, and developing new architectures tailored to Arabic's linguistic intricacies.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the support of Effat College of Engineering at Effat University, Jeddah, Saudi Arabia.

REFERENCES

[1] M. S. Ameur, F. Meziane, and A. Guessoum, “Arabic machine translation: A survey of the latest trends and challenges,” Computer Science Review, vol. 38, p. 100305, 2020. doi:10.1016/j.cosrev.2020.100305

[2] Aldawsari, M., Kolhar, M., & Dawood Omer, O. S. (2023). Within-document Arabic event coreference: Challenges, datasets, approaches and future direction. Applied Sciences, 13(19), 11004. https://doi.org/10.3390/app131911004

[3] Harrat, S., Meftouh, K., & Smaili, K. (2019). Machine translation for Arabic dialects (survey). Information Processing & Management, 56(2), 262–273. https://doi.org/10.1016/j.ipm.2017.08.003

[4] Ahammad, S. H., Kalangi, R. R., Nagendram, S., Inthiyaz, S., Priya, P. P., Faragallah, O. S., Mohammad, A., Eid, M. M., & Rashed, A. N. (2023). Improved neural machine translation using Natural Language Processing (NLP). Multimedia Tools and Applications. https://doi.org/10.1007/s11042-023-17207-7

[5] M. Alhawarat and A. O. Aseeri, “A superior Arabic text categorization deep model (SATCDM),” IEEE Access, vol. 8, pp. 24653–24661, 2020. doi:10.1109/access.2020.2970504

[6] L. Lulu and A. Elnagar, “Automatic arabic dialect classification using Deep Learning Models,” Procedia Computer Science, vol. 142, pp. 262–269, 2018. doi:10.1016/j.procs.2018.10.489

[7] D. Gamal, M. Alfonse, S. M. Jiménez-Zafra, and M. Aref, “Case study of improving English-arabic translation using the transformer model.,” International Journal of Intelligent Computing and Information Sciences, vol. 23, no. 2, pp. 105–115, 2023. doi:10.21608/ijicis.2023.210435.1270

[8] L. H. Baniata, S. Kang, and Isaac. K. Ampomah, “A reverse positional encoding multi-head attention-based neural machine translation model for Arabic dialects,” Mathematics, vol. 10, no. 19, p. 3666, 2022. doi:10.3390/math10193666

[9] Zakraoui, J., Saleh, M., Al-Maadeed, S., & Alja’am, J. M. (2021). Arabic machine translation: A survey with challenges and future directions. IEEE Access, 9, 161445–161468. https://doi.org/10.1109/access.2021.3132488

[10] S. H. Ahammad et al., “Improved neural machine translation using Natural Language Processing (NLP),” Multimedia Tools and Applications, 2023. doi:10.1007/s11042-023-17207-7

[11] M. Abdurahimov, “Leveraging BERT for English-Arabic Machine Translation” 2023.

[12] Purohit, A., Yogi, K. K., & Sharma, R. (2023). A comparison of machine translation methods for Natural Language Processing and their challenges. Innovations in Computational Intelligence and Computer Vision, 475–487. https://doi.org/10.1007/978-981-99-2602-2_36

[13] N. Bensalah, H. Ayad, A. Adib, and A. I. Farouk, “Transformer model and convolutional neural networks (cnns) for Arabic to English machine translation,” Proceedings of the 5th International Conference on Big Data and Internet of Things, pp. 399–410, 2022. doi:10.1007/978-3-031-07969-6_30

[14] Lola BurgueñoLola Burgueño is an Associate Professor at the University of Malaga (UMA), “A LSTM-based neural network architecture to infer model transformations,” Modeling Languages, https://modeling-languages.com/lstm-neural-network-model-transformations/ (accessed Dec. 30, 2023).

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv preprint arXiv:1810.04805, 2019. [Online]. Available: https://arxiv.org/abs/1810.04805