International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 12 Issue: 2
Article Received: *25 November 2023* ***Revised:*** *12 December 2023* ***Accepted:*** *30 January 2024*

Automatic Arabic Text Diacritisation Challenges: A Review

Iman Zubeiri*1 Adnan Souri2, Badr Eddine El Mohajir3
1, 2,3, NTTI Team, FS, Abdelmalek Essaadi University, Tetouan, Morocco
imane.zubeiri@gmail.com1adnan.souri@gmail.com2 B.elmohajir@ieee.ma3

Abstract— Arabic text diacritization poses a significant challenge within the realm of Natural Language Processing, driven by the intricacies of language modelization and the techniques employed to accurately add diacritics to the text. The complexity is significantly amplified when tackling the Arabic language, characterized by its unique linguistic features. The scarcity of tools and resources dedicated specifi-cally to Arabic diacritization, coupled with the necessity for algorithmic adapta-tion and modelization techniques, further intensifies the challenges. This paper undertakes the task of presenting multiple research endeavors dedicated to Arabic text diacritization, applying diverse algorithms across various datasets. The sub-sequent comparison of these research efforts aims to provide a comprehensive understanding of the field, offering conclusive insights that can guide researchers in their future work and endeavors in this specialized domain.

Keywords- ANLP, ATD, Deep Learning

I. INTRODUCTION

In Arabic, diacritics are not decorative flourishes – they are the secret sauce that unlocks the language's true richness and complexity. Imagine them as tiny musical notes guiding your pronunciation and revealing hidden layers of meaning.

Beyond their role in pronunciation, diacritics are important for disambiguating meanings: Arabic words often share the same root consonants, and the addition of diacritics helps to disambiguate their meanings.

For example, the root "s-l-m" in Table 1 can be associated with writing (‘silm’, Peace) (‘solam’, Stairs), and diacritics specify the intended meaning.

Additionally, diacritics contribute to language standardization: ensuring clear communication across learners, readers, and speakers, and guiding everyone toward smooth understanding and accurate articulation. This is especially crucial in formal and educational settings, where clarity and precision are paramount. Imagine a classroom without diacritics – it would be like deciphering a puzzle blindfolded! With these marks, learners can confidently navigate the nuances of written Arabic, and teachers can effectively bridge the gap between written and spoken forms.

TABLE I. ONE ARABIC WORD HAS MULTIPLE MEANINGS DEPENDING ON DIACRITICS

A written word without diacritics signs	Manning	Written word with diacritics signs
سلم	Peace	سِلْمٌ
سلم	Submit	سَلَّمَ
سلم	Stairs	سُلَّمٌ
سلم	Survive	سَلِمَ

This paper presents several research studies addressing Arabic text diacritization (ATD), each employing different algorithms on diverse datasets. A comprehensive comparison of these research efforts is provided, culminating in conclusions designed to offer valuable insights for researchers.

The structure of this paper is organized as follows: In the initial section, we delve into related works, exploring existing diacritization tools, their categories, and the relevant literature. Following that, we address the challenges associated with ATD in Section 4; we delve into ATD techniques, conducting an in-depth survey of recent works. In Section 5, focuses on datasets related to ATD, providing valuable resources for researchers. Finally, we present our conclusions and summarize the key findings of our work in the concluding section.

II. RELATED WORKS

Research on ATD has attracted significant attention in recent years. Various studies and approaches have emerged to tackle the challenges in diacritizing Arabic text, reflecting the ongoing efforts to enhance the accuracy and efficiency of diacritization processes.

In the study referenced [5], the authors integrate Arabic morphological rules and machine learning (ML) techniques to tackle diacritization, covering both morphological and case ending diacritics. The ongoing research emphasizes continuous improvement, intending to incorporate additional linguistic

rules, semantic features, various machine-learning techniques, and larger datasets to effectively handle Out-Of-Vocabulary cases.

In [6] researchers focus on diacritizing Modern Standard Arabic (MSA), a writing style that often omits short vowels. While experienced readers can usually deduce meaning from context, certain instances remain challenging, even for native speakers. Unlike existing algorithms that fully restore all diacritics, this system selectively adds diacritical marks where necessary to address ambiguity. By combining morphological analyzers and context similarities, the system generates potential diacritics for words and employs statistical methods to eliminate ambiguity.

The paper referenced in [7] introduces an automated system for ATD, employing a multi-layered approach with three levels of analysis granularity. The system utilizes diacritized language models at the surface form, morphological segmentation (prefix/stem/suffix), and character levels. Utilizing Viterbi search for each pass, the system systematically sequentially applies these models.

In [8] authors introduce a hybrid system to automatically diacritize Arabic sentences, incorporating both linguistic rules and statistical approaches. The method unfolds in four stages. Initially, morphological analysis is performed using Alkhalil Morpho Sys, followed by syntactic rule-based filtering. Subsequently, a discrete hidden Markov model and the Viterbi algorithm are employed to determine the most probable diacritized sentence, addressing unseen transitions using smoothing techniques. The final stage handles words not analyzed by Alkhalil, employing statistical treatments based on letters.

In [9] the study explores a hybrid approach for automatic diacritization of Arabic text, utilizing a recurrent neural network (RNN). The approach combines the capabilities of the MADAMIRA full morphological and syntactical analyzer with the RNN, using only the high-confidence diacritics and word segmentation output from the analyzer to generate fully diacritized text.

In this study [10], the attention is on two Maghrebi sub-dialects, namely Tunisian and Moroccan, utilizing Conditional Random Fields (CRF). In addition to character n-grams, character-level Brown clusters are employed as features.

In [11] authors continue the work on the two Maghrebi Arabic sub-dialects [10], using this time a character-level deep neural network architecture with two bi-LSTM layers stacked over a CRF output layer. The model demonstrates the ability to implicitly identify the sub-dialect of the input.

Authors in [12] introduce an innovative method for ATD using deep encode-decode RNN assisted by a morphological analyser. The proposed approach, along with text correction techniques, demonstrates superior performance on the Wikinews test set.

The study presented in [13] introduces the first openly available dataset, derived from the Tashkeela Corpus, comprising 55K lines and about 2.3M words. The work conducts a thorough review of existing systems and tools for ATD, along with an empirical study comparing the performance of six of these tools on the new dataset. Experimental results demonstrate that the neural Shakkala system surpasses traditional rule-based approaches and closed-source tools.

In this study [14], authors implemented a seq2seq model to create a comprehensive system for Arabic diacritic recovery. The model was trained on a fixed-length sliding window of n words, with each word represented by its characters. To enhance performance, they incorporated a voting mechanism to select the most common diacritized form of a word in various contexts.

In [15], the authors introduce a range of deep learning models trained for the automatic diacritization of Arabic text. Employing two primary approaches, namely Feed-Forward Neural Network (FFNN) and RNN, the models integrate several enhancements, including 100-hot encoding, embeddings, CRF, and Block-Normalized Gradient (BNG).

Furthermore, the authors highlight the broader implications of Arabic diacritics by showing how they can enhance Natural Language Processing (NLP) tasks such as Machine Translation (MT) through their proposed Translation over Diacritization (ToD) approach.

The work presented in [16], explored Arabic diacritization using a BiLSTM neural network with CRF, avoiding the use of morphological analyzers, dictionaries, or feature engineering. The approach's effectiveness depends on diacritization quality and dataset size. Compared to existing models, the method achieved state-of-the-art DER results on ATB parts 1, 2, and 3. Despite requiring more resources, the presented deep learning approach, combined with CRF, proved beneficial, emphasizing the importance of large, consistently diacritized training sets covering diverse topics.

The paper referenced in [17] conducts extensive experiments to optimize design and data encoding alternatives, addressing challenges in sequence lengths and proposing efficient diacritized output encodings. The results recommend a solution that utilizes four bidirectional LSTM layers on the larger new Tashkeela dataset.

The study in [18] introduces an innovative approach to address the ATD problem, using a three-component pipeline. The first component features a deep learning model, specifically a multi-layer RNN with LSTM and Dense layers. This model is complemented by a character-level rule-based corrector, which applies deterministic operations to prevent certain errors and a word-level statistical corrector that utilizes context and distance information to rectify diacritization issues. When evaluated the system outperforms all tested systems. When ignoring the

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 12 Issue: 2
Article Received: *25 November 2023* ***Revised:*** *12 December 2023* ***Accepted:*** *30 January 2024*

diacritization of the last letter of every word, the system still demonstrates strong performance.

In [19], authors propose an approach to enhance ATD by incorporating auto-generated knowledge through regularized decoding and adversarial training. Specifically, the method treats the auto-generated knowledge, such as diacritization labels from another criterion, as distinct "gold" labels during decoding. This process allows the model to learn and predict these additional labels. Additionally, adversarial training is employed to ensure that shared information between gold and auto-generated labels contributes effectively to the diacritization process. By leveraging regularized decoding and adversarial training, the approach enables the main tagger to intelligently utilize auto-generated knowledge from an existing diacritization tagger.

The paper referenced in [20] analyzes diacritic recognition performance in Arabic Automatic Speech Recognition (ASR) systems, addressing the challenge of limited diacritical marks in existing speech corpora. It experimentally explores whether input diacritization affects ASR quality and compares diacritic recognition performance with text-based diacritization as a post-processing step. The study reveals that ASR diacritization notably outperforms text-based diacritization, especially when the ASR model is fine-tuned with manually diacritized transcripts.

Authors in [21] emphasize the utility of automatic Arabic diacritization across various applications and introduce a novel approach named 2SDiac. Unlike previous works focusing on non-diacritized text, this model supports optional diacritics, enhancing accuracy by incorporating human annotations for ambiguous words.

A recent study in [22] introduced a two-phase training methodology named PTCAD for Arabic Text Diacritization. It begins with a pre-finetuning phase that integrates learning from linguistically relevant tasks to enhance contextual understanding, followed by a finetuning phase using pre-trained BERT-like models as token classifiers. This method demonstrated significant improvements in Word Error Rate (WER) reduction compared to state-of-the-art models.

The study in [23] explored global and local attention mechanisms for automatic Arabic text diacritization. These attention-based models aim to enhance the diacritization process by focusing on both the overall context of the text and the specific local features of each word, which helps in improving the accuracy of diacritization predictions

ARABIC TEXT DIACRITIZATION CHALLENGES

While automatic diacritisation holds immense promise for unlocking the richness of Arabic texts, it faces several challenges:

A. Limited Resources:

Compared to other languages, the field of ATD still suffers from the sparsity of high-quality training data. This lack of robust datasets can hinder the effectiveness of ML models, restricting their ability to generalize and adapt to diverse writing styles and contexts.

B. Lexical-Level Ambiguity:

POS Ambiguity: refers to words that share the same spelling and part-of-speech tag but have different meanings. In addition, homographs are words with the same spelling but different parts-of-speech and meanings.

This lexical ambiguity can pose challenges in NLP and understanding, as determin-ing the intended meaning becomes context-dependent in such cases. This challenge demands the algorithm to consider both lexical and grammatical cues for accurate disambiguation [1] [2].

Challenge: Without proper diacritization, a sentence containing "كتب" can be ambiguous. Automated systems often struggle to correctly apply diacritics based on broader textual context, leading to errors that change the intended meaning of a text.

Case Study: In one documented instance, a translation tool mistakenly interpreted "كتب في الرف" as "He wrote on the shelf" instead of the correct "Books on the shelf" due to incorrect diacritization of "كتب". This type of error can lead to misunderstandings in practical applications such as automated translation or content retrieval systems.

C. Grammatical-Level Ambiguity:

Sentential Ambiguity: This occurs when entire sentences or phrases can be inter-preted in multiple ways. Diacritics can play a crucial role in resolving this ambiguity, clarifying verb conjugations, noun cases, and other grammatical structures that af-fect sentence meaning. For instance, consider the sentence "يَكْتُبُ الْكِتَابَ" – depending on the diacritics, it could translate to "He writes the book" or "The book is written." The algorithm needs to understand the grammatical context to resolve this ambiguity correctly.

Challenge: Diacritization errors in such cases can lead to significant misinterpretations, particularly in educational materials or legal documents where the exact meaning is crucial.

Case Study: An e-learning platform encountered issues with automated content generation where sentence meanings were altered due to improper diacritization, affecting the learning process. Corrective measures involved manual review, highlighting the need for more sophisticated AI that understands grammatical contexts.

The integration of ML techniques for ATD represents a significant advancement, offering a data-driven and context-aware approach to address the complexities and challenges inherent in diacritizing Arabic text.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 12 Issue: 2
Article Received: *25 November 2023* ***Revised:*** *12 December 2023* ***Accepted:*** *30 January 2024*

ML techniques, particularly those based on deep learning models like RNNs or Transformers, offer promising solutions to address challenges in ATD.

These models specialize in pattern recognition and context modeling, crucial for understanding the intricacies of Arabic text. RNNs, processing text sequentially, excel in capturing long-term dependencies and comprehending the contextual shifts in word meanings based on neighboring words and overall sentence structure.

Moreover, ML models contribute significantly to handling ambiguity in diacritiza-tion. By training on large and diverse datasets that include examples of ambiguous words, these models can learn contextual cues and effectively disambiguate mean-ings. Leveraging the contextual information surrounding words, ML models make informed predictions about diacritic placement, thereby reducing ambiguity in chal-lenging cases. This approach not only enhances diacritization accuracy but also sup-ports the development of efficient models suitable for deployment on low-power devices, which holds the potential to democratize access to diacritization technology.

Transformers, on the other hand, analyze the entire text simultaneously, allowing them to discern complex linguistic patterns and make context-aware predictions for diacritic placement. This capability to understand contextual relationships enables ML models to recognize and predict diacritics with sensitivity to the surrounding context.

III. ARABIC TEXT DIACRITIZATION TECHNIQUES

A. Traditional Rule-Based Approaches for ATD:

Traditional rule-based approaches rely on manually crafted linguistic rules to predict the placement of diacritics. These rules are designed based on features such as word morphology, context, and lexical knowledge.

Examples of techniques include finite-state automata, representing language rules as transitions between states, and decision trees, classifying words based on specific criteria to assign diacritics.

The strengths of traditional rule-based approaches lie in their high accuracy for common cases and explainability, as the reasoning behind diacritic choices is transparent. However, these approaches may have limited generalizability to rare words or exceptions, and maintaining or modifying rules can be complex and time-consuming.

To address the inherent limitations of these systems, integrating more complex linguistic theories could significantly enhance their performance.

Dependency Grammar: which focuses on the dependency relations between words in a sentence rather than on a phrase structure, can provide a more nuanced understanding of sentence construction. By implementing dependency rules, diacritization systems can more accurately reflect the syntactic roles of words, improving the precision of diacritics placement in complex sentences.

Automation in Rule Generation

While traditional rule-based diacritization systems are highly interpretable, they require extensive linguistic expertise and continuous updates, which are both time-consuming and costly. Leveraging machine learning to semi-automate the generation and updating of these rules could offer a substantial improvement in maintaining these systems.

Collaborative Filtering: By analyzing corrections made by users in real-time, such as through feedback mechanisms embedded in text editors or reading applications, the system can learn common deviations from existing diacritization rules and adjust accordingly. This method not only improves the system's performance but also engages the user community in the system's refinement process.

B. Hybrid Approaches for ATD:

Hybrid approaches combine traditional rule-based methods with other techniques such as statistical models and ML to enhance diacritization accuracy.

In hybrid approaches, rule-based systems are integrated with statistical models like n-grams or Hidden Markov Models, or ML algorithms such as neural networks.

Examples of techniques represented in [5], [6], [7], [8], and [9] include the integration of Arabic morphological rules and ML techniques, Incorporating linguistic rules and statistical approaches, and combining MADAMIRA's morphological and syntactical analyzer with an RNN.

Hybrid approaches offer improved accuracy and generalization, handling both common and rare cases effectively. They are adaptable, learning from data and adjusting to changes in language usage. However, designing and training hybrid models can be challenging, and ML models may lack clear explanations for their predictions.

C. Deep Learning Approaches for ATD:

Building upon the foundation of traditional and hybrid approaches, deep learning techniques have emerged as powerful tools for tackling the complexities of ATD. These methods leverage large datasets and sophisticated neural network architectures to automatically predict the placement of diacritics. The summarized research delves into a variety of ML models for ATD. The studies leverage diverse ML techniques, displaying the versatility and effectiveness of these models in enhancing diacritization accuracy.

CRFs are prominently featured, as seen in research focused on Maghrebi sub-dialects [10] and in conjunction with deep learning models [15]. RNNs play a crucial role, with bidirectional LSTM layers recommended for optimization [17]. Additionally, the Seq2Seq model [14] and innovative multi-layer RNN approach [18] underscore the significance of diverse

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 12 Issue: 2
***Article Received:** 25 November 2023 **Revised:** 12 December 2023 **Accepted:** 30 January 2024*

ML architectures in pushing the boundaries of diacritization capabilities. These ML models collectively contribute to the continuous improvement and advancement of ATD techniques.
Deep learning models often outperform hybrid techniques in terms of accuracy, especially on large datasets. However, the difference in performance may not be significant on smaller datasets. Hybrid techniques can be more robust to limited data and may be preferred in specific domains where interpretability is crucial.
The choice between hybrid and deep learning techniques for ATD depends on several factors, including the size and quality of available data, the desired level of accuracy, and the need for interpretability.

TABLE 2. COMPARISON OF HYBRID AND DEEP LEARNING TECHNIQUES

Feature	Hybrid Techniques	Deep Learning Techniques
Accuracy	Moderate to high, depending on the complexity of the rules and the quality of the training data.	High-on large datasets, may struggle with unseen data.
Generalizability	Limited to the data used to develop the rules.	Can generalize well to unseen data if trained on a diverse dataset.
Interpretability	Rules are easily interpretable.	Difficult to interpret the learned patterns.
Computational cost	Lower computational cost.	Higher computational cost, especially for training.
Development effort	Requires linguistic expertise for rule development.	Requires less manual effort, but needs large datasets.

IV. DATASETS

The availability and quality of datasets play a crucial role in the development and evaluation of ML models, especially for ATD. Below, we provide a brief overview of common datasets that have been utilized in the previous research.
The Tashkeela Corpus has been utilized in multiple research papers, including [8], [13], [15], [17], and [18]. Boasting a substantial collection of over 75 million fully diacritized words, this dataset is comprehensive in its coverage of both modern and classical Arabic. The dataset comprises 55,000 lines, totaling around 2.3 million words, making it a substantial resource consistently referenced across various studies for its significant contributions to ATD research.
The ATB Datasets, encompassing ATB1, ATB2, and ATB3, are utilized in research referenced in [7], [9], [16], [19], and [20]. These datasets consist of Arabic text with diacritics, strategically segmented into distinct parts. Renowned for their widespread usage, they serve as fundamental resources for the evaluation of diacritization models. Researchers systematically employ experiments on these datasets, focusing on the optimization of designs and bidirectional LSTM layers to glean insights into model performance. Notably, several studies explicitly reference the ATB datasets, underlining their significance in conducting empirical investigations and advancing diacritization research.
The WikiNews corpus, as referenced in the study [14], has a crucial role in the research on ATD employing deep encode-decode RNNs. This corpus is specifically utilized to evaluate the performance of innovative ATD approaches, including the assessment of deep learning models assisted by morphological analyzers.
The Maghrebi Sub-Dialect Datasets, as mentioned in references [10] and [11], consist of translations of the New Testament in two Maghrebi sub-dialects: Moroccan and Tunisian. Each translation is fully diacritized, encompassing 8,200 verses. These datasets offer opportunities for exploration and application of techniques such as CRF and character-level deep neural networks.
The best dataset for ATD depends on the specific needs. Consider the size, coverage, and applications required. If you need a large and comprehensive dataset for general research, Tashkeela Corpus might be ideal. If you need a dataset for evaluating and optimizing your model, ATB Datasets could be a good choice.

V. CONCLUSION

In conclusion, this paper provides a comprehensive exploration of ATD, emphasizing the challenges, techniques, and datasets associated with this intricate task. By examining diverse research endeavors employing traditional rule-based, hybrid, and deep learning approaches, the paper elucidates the strengths and limitations of each method. Furthermore, the paper highlights the critical role of diacritics in unlocking the richness of Arabic texts, facilitating pronunciation, disambiguating meanings, and contributing to language standardization. In essence, this paper serves as a valuable resource for researchers and practitioners engaged in ATD, providing insights, comparisons, and references to guide future researchers in this specialized domain.

REFERENCES

[1] Zubeiri, I, Souri, A, EL Mohajir, BE, Neural Network for Arabic Text Diacritization on a New Dataset. In: Proceedings of the 6th International Conference on Big Data and Internet of Things. BDIoT 2022. Lecture Notes in Networks and Systems, vol 625. Springer, Cham. https://doi.org/10.1007/978-3-031-28387-1_17

[2] Thompson B, Alshehri A, Improving Arabic Diacritization by Learning to Diacritize and Translate. In: https://arxiv.org/ftp/arxiv/papers /2109/2109.14150.pdf

[3] Abandah GA, Graves, Al-Shagoor B, Arabiyat, A, Jamour, F, Al-Taee, M, Automatic diacritization of Arabic text using recurrent neural networks. In: Int. J. Document Anal. Recognit. vol. 18, no. 2, pp. 183_197, Jun. 2015.

[4] Devlin, J, Chang K, Lee K, Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In arXiv:1810.04805v2 [cs.CL] 24 May 2019

[5] Fashwan, A, Alansary,S, SHAKKIL: An Automatic Diacritization System for Modern Standard Arabic Texts. In: Proceedings of The Third Arabic Natural Language Processing Workshop (WANLP), pages 84–93, Valencia, Spain, April 3, 2017. Association for Computational Linguistics

[6] Alnefaie, R, Azmib, AM, Automatic minimal diacritization of Arabic texts. In: The 3rd International Conference on Arabic Computational Linguistics. Procedia Computer Science 117 (2017) 169–174

[7] Al-Badrashiny, M, Hawwari,A, M, Diab, A Layered Language Model based Hybrid Approach to Automatic Full Diacritization of Arabic. In: Proceedings of The Third Arabic Natural Language Processing Workshop (WANLP), pages 177–184, Valencia, Spain, April 3, 2017.

[8] Chennoufi, A, Mazroui , A, Morphological, syntactic and diacritics rules for automatic diacritization of Arabic sentences. In: Journal of King Saud University – Computer and Information Sciences 29 (2017) 156–163.

[9] Alqudah, S, Abandah, G, Arabiyat, A, Investigating hybrid approaches for Arabic text diacritization with recurrent neural networks. In: 2017 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)

[10] Darwish, K, Abdelali, A, Mubarak, H, Samih, Y, Attia, M, In : The 3rd Workshop on Open-Source Arabic Corpora and Processing Tools.

[11] Abdelali, A, Attia, M, Samih, Y, Darwish, K, Hamdy, M, Diacritization of Maghrebi Arabic Sub-Dialects. In: arXiv:1810.06619v2 [cs.CL] 29 Oct 2018

[12] Noaman, HM, Sarhan, SS, Rashwan, MAA, A Hybrid Approach for Automatic Morphological Diacritization of Arabic Text. In : Mansoura Journal of Computers and Information Sciences.

[13] Fadel, A, Tuffaha, I, Al-Ayyoub, M. Arabic Text Diacritization Using Deep Neural Networks. In: arXiv:1905.01965v1 [cs.CL] 25 Apr 2019.

[14] Mubarak, H, Abdelali, A, Sajjad, H, Samih, Y, Darwish, K. Highly Effective Arabic Diacritization using Sequence to Sequence Modeling. In : Proceedings of NAACL-HLT 2019, pages 2390–2395.

[15] Fadel, A, Tuffaha, I, Al-Ayyoub, M. Neural Arabic Text Diacritization: State of the Art Results and a Novel Approach for Machine Translation. In: arXiv:1911.03531v1 [cs.CL] 8 Nov 2019.

[16] Al-Thubaity, A,Alkhalifa, A, Almuhareb, A, Alsani, W. Arabic Diacritization Using Bidirectional Long Short-Term Memory Neural Networks With Conditional Random Fields. In: Received August 5, 2020, accepted August 19, 2020, date of publication August 24, 2020, date of current version September 3, 2020.

[17] Abandah, G, Abdel-Karim, A. Accurate and fast recurrent neural network solution for the automatic diacritization of Arabic text. In: Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 06, No. 02, June 2020.

[18] Abbad, A, Xiong, S, Multi-components System for Automatic Arabic Diacritization. In: ECIR 2020: Advances in Information Retrieval pp 341–355.

[19] Qin, H, Chen, G, Tian, Y, Song, Y, Improving Arabic Diacritization with Regularized Decoding and Adversarial Training. In : Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)

[20] Aldarmaki, H, Ghannam, A, Diacritic Recognition Performance in Arabic ASR. In : https://arxiv.org/ftp/arxiv/papers/2302/2302.14022.pdf

[21] Baha, P, Di Gangi, M, Rossenbach1, N, Zeineldeen, M, Take the Hint: Improving Arabic Diacritization with Partially-Diacritized Text. In : arXiv:2306.03557v2

[22] I.Berrada, A.Skiredj,"Arabic Text Diacritization In The Age Of Transfer Learning: Token Classification Is All You Need," ar5iv. Accessed: Apr. 29, 2024. [Online]. Available: https://ar5iv.labs.arxiv.org/html/2401.04848

[23] A. Mijlad and Y. E. Younoussi, "The Global and local attention for automatic Arabic text diacritization," Int. J. Eng. Appl. Phys., vol. 3, no. 1, pp. 653–662, Jan. 2023, doi: https://n2t.net/ark:/15735/IJEAP.v3i1.113.