Innovative AI system of Arabic vowel signs can help learners and speakers read Arabic texts fluently, scientists say

Scientists develop an advanced automated system of vowel sounds that can inform learners and speakers how a letter should be pronounced in Arabic.

SHARJAH, EMIRATE OF SHARJAH, UNITED ARAB EMIRATES, December 17, 2024 /EINPresswire.com/ -- A newly developed automated system can add vowel sings to computerized Arabic texts, enabling learners and speakers to read them in an easy and accurate manner, scientists reveal.

In linguistic jargon, the signs are called diacritics. Adding the right diacritics manually is a time-consuming task which only linguists can master, and their absence from digital texts has been an issue for scientists to grapple with as it is even hard for native speakers to read Arabic texts properly without them.

But the scientists say their system can supplement all types of computerized texts with their proper diacritics automatically. Diacritics are an integral part of Arabic texts as they are placed below, above, and occasionally even through letters to help in pronouncing words correctly and grasping their meanings.

A creative Arabic calligraphy artwork showcasing the beauty of the language's script and diacritic intricacies. Credit: Ruba Kharsa

The details about the scientists' automated system are published in the journal Expert Systems with Applications. The research dubs the system "a state-of-the-art approach" that can improve the accuracy of Arabic texts and their pronunciation. (Original Source URL: https://www.sciencedirect.com/science/article/pii/S0957417424002811)

"In order to accurately represent the meaning and pronunciation of Arabic words and sentences, the presence of diacritics plays a crucial role," the scientists write. "Over the years, researchers have dedicated significant efforts to enhancing automated diacritization systems."

The diacritical marks or vowel sounds are called Harakat in the Arabic language. There are three primary symbols and five secondary ones. They are of paramount importance to correctly read Arabic texts, guess shades of meanings of different words, as well as their syntactical function in a sentence.

Arabic diacritics can even change the entire meaning of words. Crucial in shaping pronunciation, meaning and gender distinction, the signs are indispensable for obtaining correct Arabic language skills of reading, speaking, learning, and listening.

The Arabic alphabet comprises 28 letters, all representing consonants. Unlike English, consonant clusters are not common in Arabic. Thus, each of its 28-letter consonants comes with a diacritic or vowel sound that joins them together in a flowing manner both in writing and speech.
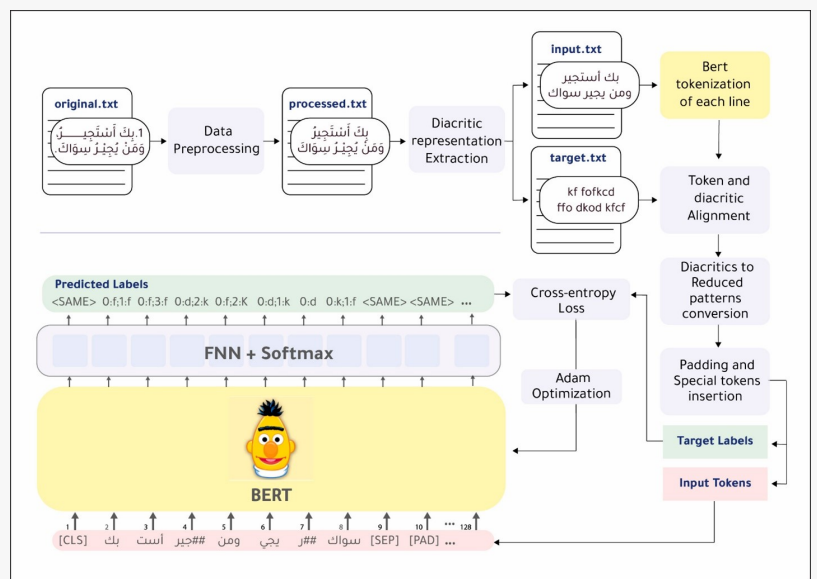
The scientists call their new system 'SUKOUN' in reference to an Arabic diacritic whose presence above a letter

Sukoun Diacritization	Input
كُلَّما تَقَدَّمْتِ بِكِ الْحَياةُ عَلَّمَتْكِ الْمَزِيدَ وَزادَتْ فَهْمَكِ وَوَعْيَكِ وَحِكْمَتَكِ وَبِما أَنَّها دُنْيا مَجْبُولَةً عَلَى الْأَلَمِ وَالْكَدَرِ وَلا راحَةَ فِيها لِمُؤْمِنٍ إِلَّا بِلِقاءِ رَبِّهِ فَمِنَ الْجَمِيلِ أَنْ تَحْتَفِظِي بِرُوحِ مِنْ طُفُولَةٍ تُخَفِّفُ جُزْءًا مِنْ ذَلِكَ الضَّغْطِ وَتُعادِلُهُ وَمِنَ الْأَجْمَلِ أَلَّا تَسْمَحِي لِأَحَدٍ بِسَرِقَةِ تِلْكَ الرُّوحِ مِنْكِ. 🌸	كلما تقدمت بك الحياة علمتك المزيد وزادت فهمك ووعيك وحكمتك وبما أنها دنيا مجبولة على الألم والكدر ولا راحة فيها لمؤمن إلا بلقاء ربه فمن الجميل أن تحتفظي بروح من طفولة تخفف جزءا من ذلك الضغط وتعادله ومن الأجمل ألا تسمحي لأحد بسرقة تلك الروح منك. 🌸

An example showcasing SUKOUN's capability to accurately diacritize female-targeted Arabic text, demonstrating its precision in handling gender-specific language. Credit: University of Sharjah

The pipeline of the proposed architecture Credit: Expert Systems with Applications. DOI: https://doi.org/10.1016/j.eswa.2024.123416

indicates that it is in a still position. Like other diacritics, it plays a key phonetic, semantic, and grammatical role. The diacritic is pronounced 'as-sokoun' and its correct pronunciation requires intensive training for correct recitations of the Quran, the Muslim holy book.

"This study introduces a real-time diacritization system called SUKOUN, which offers diacritized text through a user-friendly website. A comparison with existing automatic diacritization tools, using six example texts, reveals the superior prediction accuracy and preservation of input format provided by SUKOUN," the scientists write.

Ashraf Elnagar, Sharjah University's professor of computer science, described SUKOUN's performance as "groundbreaking," claiming to have "achieved a Diacritic Error Rate (DER) as low as 1.14% and a Word Error Rate (WER) of just 3.34% on the Arabic Diacritization (AD) dataset, and
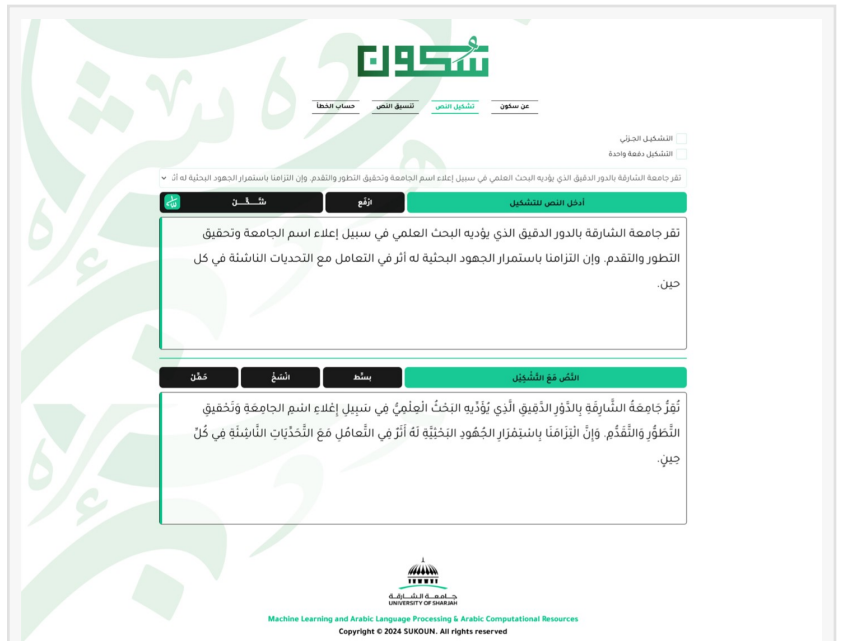
an even more remarkable DER of 1.11% on the Tashkeela Processed (TP) dataset. These results represent over a 30% reduction in error rates compared to the previous best systems.

"What makes SUKOUN exceptional is not just its accuracy but also its efficiency and practicality. It requires less computational power to train and deploy, thanks to innovations in data preprocessing and transfer learning. Additionally, it operates in real-time, allowing users to input Arabic text and receive a fully diacritized version instantly via a user-friendly web interface."
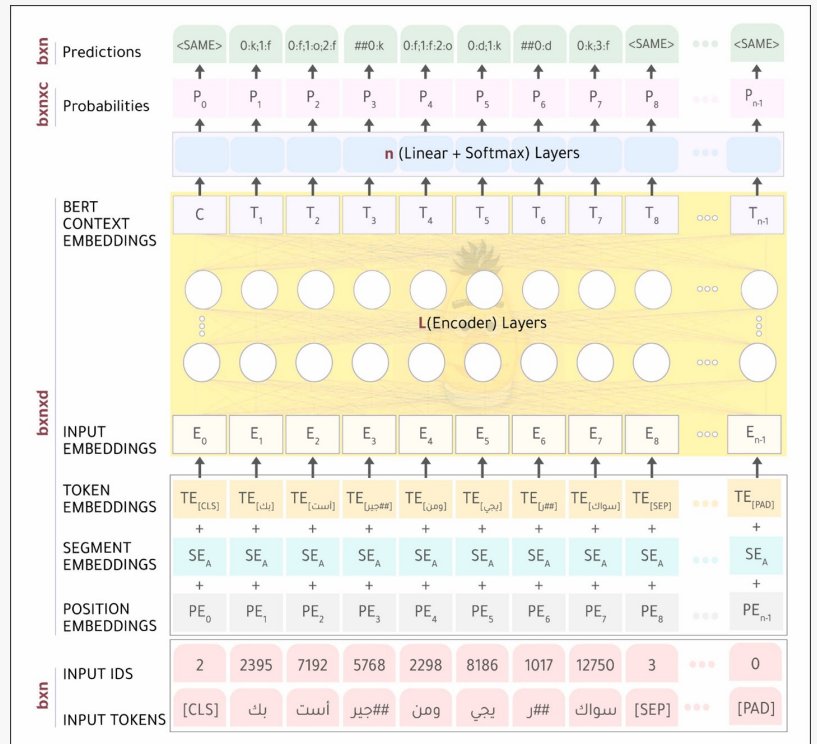
Arabic has both long and short vowels. While long vowels are distinguishable as they are represented by separate letters, the short ones are only recognized by diacritics or vowel marks written above or under the letter in a process called Tashkeel or TP in scientific jargon.

The system's success is due to its ability to bridge the gap between the linguistic complexity of the Arabic language, particularly in morphology, and the technological capability of machine learning. "SUKOUN has the potential to revolutionize applications in education, text-to-speech systems, translation, and beyond, making the Arabic language more accessible to all," added Prof. Elnagar.

SUKOUN website, showcasing its real-time Arabic diacritization feature developed for enhancing text readability and accuracy. Credit: University of Sharjah

The proposed research methodology Credit: Expert Systems with Applications. DOI: https://doi.org/10.1016/j.eswa.2024.123416

The authors showcase their system not merely as an AI tool but rather as a practical and user-friendly application, allowing anyone to add Arabic text without diacritical symbols instantly and get a version with all the correct diacritics, keeping the original text intact.

Prof. Elnagar goes on, "Beyond its accuracy and ease of use, SUKOUN has wide-ranging applications. It can improve education by helping students read and learn Arabic more effectively, support the visually impaired through better text-to-speech systems, and enhance translation services and other natural language processing tools."

If it is successfully deployed on a large scale, the automated system could change the perspective of Arabic learning and teaching, said lead author Ruba Kharsa. "SUKOUN has the potential to revolutionize Arabic education. Teachers and students can use the tool to easily diacritize texts, aiding in the learning of proper grammar, pronunciation, and meaning. This is particularly important for non-native learners and children developing their language skills.

"By enabling accurate diacritization, SUKOUN improves the effectiveness of text-to-speech systems and other accessibility tools, especially for the visually impaired. It also supports better language learning and interaction for users who rely on assistive technologies.

"SUKOUN showcases how cutting-edge AI, particularly BERT-based models, can solve complex linguistic problems efficiently. Its success demonstrates the power of AI in processing and enhancing underrepresented languages, paving the way for similar advancements in other domains."

The research underscores the power of AI to transform language learning and teaching as it ensures that "Arabic texts are accessible and comprehensible for speakers and learners worldwide," maintained Sane Yagi, Sharjah University's professor of linguistics and a co-author.

"SUKOUN is more than a diacritization tool—it's a gateway to improving education, accessibility, and cultural preservation in the Arabic-speaking world. Rooted in collaboration between the Departments of Computer Science and Foreign Languages, SUKOUN reflects the interdisciplinary innovation and commitment to excellence at the University of Sharjah."

While the industry has yet to engage with the new automated diacritical system, Prof. Elnagar predicts "significant practical applications" in education, accessibility, and language learning, providing "accurately diacritized texts to help students and teachers improve pronunciation, grammar, and comprehension."

Other implications, according to Prof. Elnagar, include enhancement of text-to-speech systems "for the visually impaired by ensuring accurate pronunciation, (and) making Arabic content more user-friendly. In automated translation services, SUKOUN reduces ambiguities in undiacritized texts, improving the quality of machine translations.

"Additionally, SUKOUN aids (Arabic) linguistic research by offering precise diacritization for large-scale text analysis and facilitates cultural preservation by making classical and historical Arabic texts accessible to future generations."

LEON BARKHO
University Of Sharjah
+971   50 165 4376
email us here

This press release can be viewed online at: https://www.einpresswire.com/article/769687501