tashkeelWAP: A Game With A Purpose For Digitizing Arabic Diacritics

Lin Kassem, Caroline Sabty, Nada Sharaf, Menna Bakry, Slim Abdennadher

Department of Computer Science and Engineering
The German University in Cairo, Egypt
lin.sara@student.guc.edu.eg,
{caroline.samy, nada.hamed, menna.nabil, slim.abdennadher}@guc.edu.eg

Abstract. Diacritics in Arabic language are the signs that are found above or under Arabic letters. Their main aim is to provide phonetic aid to readers as well as allowing them to understand the Arabic text in its intended and correct context. The existence of a diacritical mark can entirely change the meaning of Arabic text. Existing Optical Character Recognition (OCR) systems face accuracy difficulties when trying to read Arabic letters with diacritics. This affects the quality of the digitized Arabic text. We introduce “tashkeelWAP”, a web application with two games that allow the digitization of Arabic text by outsourcing it to native Arabic speaking players. As a bi-product of playing the games, we collect possible digitization of Arabic words with diacritics that were not recognized by OCR systems.

Keywords: Crowdsourcing, GWAP, Arabic, Digitization, Diacritics.

1 Introduction
Since the 7th century, the Arabic language has tremendously contributed in the fields of literature, science and art. Researchers state that the unique characteristics of the Arabic language add to its beauty and majesty [1].

With the rapid growth of the Internet in the past few years, the world is becoming dependent on digital content to a great extent. There is a need to boost the quantity and the quality of the digital Arabic content, not only to preserve the valuable Arabic language heritage but also to pass it on to the coming generations. One approach to enhance the digital Arabic content is digitizing Arabic text present in books, novels, scholarly articles, historical archives, etc.

A lot of research has been done in the field of text recognition in the past few years. Many new approaches for digitizing different languages has been introduced [2]. However Arabic text recognition is still limited due to the complex nature of the Arabic language script [3]. The overlapping characters and large number of dots and diacritics in Arabic text result in poor recognition outcomes from OCR systems [3]. Diacritics are the signs found above or under Arabic letters such as ìضی. The main aim of diacritics is to provide phonetic aid to readers as well as allowing them to understand the Arabic text in its intended context. Most OCR systems suggest removing diacritics before the digitization process in order to decrease the number of wrong digitizations produced [4]. Neglecting
that diacritics play a very essential role in the Arabic language, the addition or removal of a diacritical mark can entirely change the meaning of the Arabic text. For example the word دَفَّبُ means went (verb) in English and once the diacritic signs are changed to دَفَّبُ the meaning will change to gold (noun).

In this paper, we take advantage from crowdsourcing through Games With A Purpose (GWAPs) [5]. We introduce “tashkeelWAP”, a web application containing two games, a two-player game and a single-player game. Both games serve the purpose of digitizing Arabic text rich in diacritics. We intend to exploit the huge number of Arabic-speaking Internet users [6] and their knowledge of the Arabic language to digitize Arabic words with diacritics [7]. Thus a way to compensate for the poor digitization results produced by current OCR systems.

Our paper is structured as follows. In Section 2 the related GWAPs and techniques that tackle the digitization problem are listed. In Section 3 the concept of the two games in “tashkeelWAP” is explained, followed by a detailed evaluation of the games in Section 4. Finally our conclusion and future work is presented in Section 5.

2 Related Work

Human computation and algorithmic approaches are the most used approaches considered by researchers to improve the output obtained during the digitization process. Up till now, there are some tasks that are trivial for humans to perform yet are very challenging for computers to carry out; such as image and text recognition. In [5], von Ahn introduced the GWAP concept as an implementation of human computation. GWAPs rely on game elements and strategies to keep people motivated while performing computer challenging tasks[8].

Various GWAPs have been implemented to digitize the content of scanned documents and books as a bi-product of playing the game. DigitalKoot [9] is a game that was used by the National Library of Finland [10] to help in verifying the correctness of the previously digitized news papers archives. Another example is TypeAttack [11], where players get images containing small paragraphs of scanned text and the more a player participates in the typing game sessions, the more points they gain. Both games are optimized for digitizing content of documents written in Latin script. The games proposed in “tashkeelWAP” are optimized for digitizing documents with Arabic scripts. The Arabic script is different from the Latin script in terms of its cursive nature and the large number of dots and diacritics used. Kalema [12] and [13] are two different GWAPs for digitizing Arabic scanned documents for visually impaired people, where text can be converted to audio format or Braille. Unlike Kalema [12], the games presented in “tashkeelWAP” tend to focus specifically on Arabic words with diacritics.

re-CAPTCHA system [14] is another implementation of human computation. It is used in on-line web forms to make sure that the entities filling the forms are humans. AreCAPTCHA is the Arabic version of the re-CAPTCHA. Where users are presented with two distorted images of scanned Arabic words. One of the words is a test and the other one is a word that OCR systems failed to recognize. Consequently, AreCAPTHCA becomes a test that only humans can pass as well
as a source of crowdsourced digitizations. However, the AreCAPTCHA system considers the diacritic signs as one of the main factors causing wrong digitization outputs. The games presented in “tashkeelWAP” aim at obtaining digitizations from players while preserving the diacritic signs associated with the scanned Arabic words.

In [15] and [16], two statistical approaches for automatic restoration of Arabic diacritic signs are presented. In [15], the system uses Hidden Markov Model (HMM) approach combined with dynamic programming [17] to generate diacritics. However, in [16], the system is divided into two phases. The first phase creates a rich corpus of frequently used patterns of 4 consecutive diacritized letters. The second phase uses the created corpus to diacritize any Arabic text given to the system as an input. Authors stated that the system had limited accuracy as the final diacritization of any given Arabic word is highly dependent on the syntax and content where the word has been used.

It is worth mentioning that there are several implementations of Arabic GWAPs that tackle a variety of problems and topics. As in [18], a game was created to construct a well-structured corpus that can be used to train classification algorithms for the Arabic short text. Also in [19], a game was implemented to collect annotations from players on Arabic words and sentences. The collected annotations were then used by a sentiment analyzer for the reviews given in conversational Arabic language. Moreover, in [20] a game was proposed to collect Arabic dialect variations, making it possible to perform mapping between these dialects in the future. In [21] a set of games were introduced with the goal of helping young children with learning difficulties, to learn the Arabic language.

3 “tashkeelWAP”

“tashkeelWAP” is a web application consisting of two games. Its aim is to offer an enjoyable gaming experience to players and as a bi-product obtain Arabic digitizations of documents with diacritics. It can be used to create a corpus of digitized Arabic words with diacritics and possibly their Franco-Arabic version. Franco-Arabic is the representation of Arabic text but in the Latin script [22], where a combination of Latin characters and numbers is used to represent the diacritics in a certain word and therefore preserving the word’s phonetics. For example the Arabic word “مرحبا”, will be represented in Franco-Arabic as “Marhabaan”. Where the “an” at the end act as an indicator for the presence of the tanween Arabic diacritic sign.

The first game is a two-player game which allows players to compete against each other. The goal of the game is to use the knowledge of the players to digitize Arabic words with the correct diacritics. The second game is a single-player game which acts as a validation game. The players is given a scanned image of the word and its corresponding digitization, then they are asked to match the digitization to its corresponding word in the scanned image. In both games, elements recommended by von Ahn and Dabbish in [7] to increase the player’s sense of enjoyment were used, such as the timed response, highest-scores lists and the randomness of the game’s input.
3.1 Admin Side

In order to create a word bank to be used by the games during the testing phase, we manually scanned pages of an Arabic document, then cropped around the boundaries of each word making multiple smaller word images. Saved those images in a database, along with their corresponding diacritic-less digitization, certainty rate of 1 and a display repetitions value of 3.

The certainty rate corresponds to how confident we are regarding the correctness of the digitization of the scanned Arabic word image. A certainty rate of 3 means that we are highly confident. A certainty rate below 3 means that we are not confident about the correctness of the digitization and there is a possibility that the digitization of the Arabic word is wrong. The display repetitions value decides how many times a scanned Arabic word image needs to be used as an input to the two-player game in order to add digitized diacritic signs to it. A word will not be used in the two-player game if it has a display repetition value of 0.

3.2 Homepage

Once a player goes to “tashkeelWAP” website, a welcome page appear to them. It contains a leader board showing the names and scores of the top seven players ranked according to their score. In addition to a carousel of images clarifying the flow of the games to give a general idea about the games to the players. The player starts by either registering, in case of a new player, or logging in, in case of a returning player. This is to keep track of their score. Once successfully logged in, the player is redirected to the home page. It contains a mini-profile about the player in the sidebar showing their name, current score and current rank. The main section of the home page consists of buttons linking to either the two-player game or the single-player game in addition to the individual score board of each game which shows the top three players.

3.3 Two-Player Game

The main aim of this game is to obtain digitizations for diacritic signs found in the scanned images of Arabic words. According to von Ahn’s classification of GWAP structure templates [7], this game falls under inversion-problem category.

Whenever a player chooses to play the two-player game, a random partner from the on-line players is picked. The pair will both be redirected to a page showing the rules of the game in its main section while the side bar of the page will contain the names of the two players, the round number and the game timer. After ten seconds the game rules will disappear to be replaced by the game itself. If there were no enough players on-line, the unassigned player is directed to a waiting page till another player logs in and decides to play the game.

The game session consists of two rounds, each having a time limit of 120 seconds. In each round, one player is assigned to be a hinter who gives hints to the other player who is assigned to be a solver. This allows the satisfaction of the alteration element suggested by von Ahn and Dabbish in [7] to increase the enjoyment of the players during the game.
The hinter is provided with an image of a scanned Arabic word with diacritics as can be seen in Figure 1. The hinter can provide up to three hints, whenever requested by the solver, to help him out in choosing the appropriate diacritics. The first hint is the number of diacritic signs that appear in the provided Arabic word. The second hint is the actual diacritic signs that are used in the word. Where the hinter is provided with a set of eight buttons, each representing a diacritic sign, allowing the sending of the diacritic signs to the solver. Finally, the last hint is the Franco-Arabic version of the word.

The hinter is notified about the solver’s guess once the solver submits it. The hinter is then asked to verify whether the solver’s guess is identical to the Arabic word shown in the provided scanned image or not, as illustrated in Figure 2. This is done to obtain an instant verification of the solver’s guess, thus reflecting on the certainty rate of the Arabic word. The hinter’s score is affected by how quickly a hint is provided after the solver requested it and it increases with every hint provided.

The solver is provided with the actual digitized version of the same Arabic word provided to the hinter, most probably without any diacritics as seen in Figure 3. During the round, the solver should guess and submit the correct form of the digitized word after adding the appropriate diacritics to it. The solver is allowed to request up to three hints from the hinter, in case help is needed while guessing the diacritics of the given word. A hint can be requested by clicking on the magic wand shown in Figure 3. The score of the solver is affected by how quickly a guess is submitted, whether the guess is correct or not and it decreases with every hint requested.
The round ends in two cases. The first case is when the time for the round ends without the solver submitting any guesses. The second case is when the solver submits a guess before the time of the round ends. In this case the solver’s guess is sent to the hinter and the hinter is requested to mark the solver’s guess as correct or incorrect. After the hinter verifies the solver’s guess the round ends. The winning condition is satisfied in case the hinter verifies the digitization provided by the solver is correct and identical to the one in the provided image.

3.4 Single-Player Game

The implementation and usage of a validation module during the digitization process is of great importance. It is the job of the validation module to verify that the digitized output is equivalent and identical to the original scanned images of the data in the physical form. The single-player game was designed with the aim of it acting as a validation module. The idea is inspired from the matching cards memory games. Where initially all the cards used in the game are placed face down and during the game session, players are required to figure out and turn over the pairs of matching and identical cards.

During the initiation of the game, twelve cards are created, and six scanned images of Arabic words along with their digitizations are randomly selected from the database.

The game session consists of a 90 seconds round. During the round the player should find and flip over the pair of matching cards. Where a pair of cards is considered to be matching if one card contained the image of a scanned Arabic word and the other card contained the potential digitization of this word. Upon finding and flipping the pair of matching cards a tool-tip, a check-mark and an x-mark will appear on one of the the cards, as shown in Figure 4. The tool-tip asks the player whether the two words inside the pair of cards are identical taking into consideration the diacritic signs. The player needs to press one of the marks in order to be able to continue playing the game. If the player clicks
Fig. 4: Upon finding and flipping the pair of matching cards a tool-tip, a check-mark and an x-mark will appear on one of the cards. On the check-mark then we increment the certainty rate of the digitized word. Otherwise, if the player clicks on the x-mark then we will decrement the certainty rate of the digitized word.

The round ends in two cases. The first case is when the 90 seconds pass without the player finding all the matching pairs in the twelve cards deck of the game. The second case is when the player finds all the matching pairs, making verifications when needed, before the time of the round ends. When the round ends, the player is redirected back to the home page.

The winning condition is satisfied when the player finds all the matching pairs in the twelve cards deck of the game, making verifications when needed. The player’s score depends on the time taken to finish the round. The more time spent, the lower is the score obtained.

4 Evaluation and Results
In order to predict the success and effectiveness of the two games introduced in tashkeelWAP, it was necessary to evaluate the correctness of the collected data. Along with calculating the evaluation metrics suggested by von Ahn and Dabbish in [7], in addition to investigating and analyzing how players perceived the games. “tashkeelWAP” was hosted online on https://tashkeel.herokuapp.com/ for data collection.

The GWAP evaluation metrics [7] are the throughput, the average lifetime play and the expected contribution. The throughput gives an overview of how many instances of the problem, digitizations in our case, were solved per human hour. The average life time play shows the amount of time a player is expected to spend playing the game. The expected contribution is the number of problem instances a player can be expected to solve by playing the game.
Two-Player Game

During the testing phase, 22 players played 56 sessions of the two-player game, giving a total of 112 played rounds. Around 112 diacritic-less digitizations of Arabic words were displayed and players were expected to guess and add the appropriate missing diacritic signs to the digitized words.

After analyzing the output data from the 112 rounds, it was noticed that 53.5% of the displayed words obtained correct diacritic digitizations. On the other hand, 2.8% of the words obtained wrong diacritic digitizations. The remaining 43.7% of the words did not get any digitizations, in other words the digitization entry text box was left empty. This could have occurred either because the player did not have time to enter his guess, or the player left the game in the middle of the round.

As for the GWAP evaluation metrics [7], the throughput was equal to 3 digitization per human hour. Hence, playing the two-player game for 24 hours would enable us to obtain digitizations for 72 Arabic words. The average life time shows that a player is expected to play the two-player game for approximately 10 minutes, which is equivalent to the time needed for playing two game sessions. The expected contribution was approximately equal to 0.5 digitizations per player. The equations used are illustrated in (1), (2), and (3).

\[
Throughput = \frac{\text{collected data}}{\text{total human game-play time}} = \frac{10 \text{ correct digitizations}}{\sim 3.7 \text{ hours}} \approx 3 \text{ digitizations per human hour}
\]  

\[
\text{Average Lifetime Play} = \frac{\text{total human game-play time}}{\text{number of players}} = \frac{\sim 3.7 \text{ hours}}{22 \text{ players}} \approx 10 \text{ minutes per player} = \text{Time needed for 2 game sessions}
\]  

\[
\text{Expected Contribution} = \text{Average Lifetime Play} \times \text{Throughput} = \sim 10 \text{ minutes} \times 3 \text{ digitizations per human hour} \approx 0.5 \text{ digitizations per player}
\]

Single-Player Game

The single-player game was played by 35 players during the testing phase. A total of 165 words were displayed by the game to be validated by players.
After analyzing the output data from the single-player game, 50% of the digitized Arabic words were validated. Where a digitized Arabic word is considered to be validated as correctly digitized if it had a certainty rate equal to 3 and as incorrectly digitized if it had a certainty rate of -3. The remaining 50% of the words had certainty rate values ranging from 2 to -2, which means that they are not yet validated and thus they were excluded from any further investigations.

After investigating the 50% validated words, we found that 65.05% of the validations obtained from players were indeed correct. While the remaining 34.94% of the validations obtained were false.

As for the GWAP evaluation metrics [7], the throughput was equal to 22 validations per human hour. Hence, playing the single-player game for 24 hours would enable us to obtain validations for 528 digitized Arabic words. The average life time shows that a player is expected to play the single-player game for approximately 6 minutes, which is equivalent to the time needed for playing six game sessions. The expected contribution was approximately equal to 2 validations per player. The equations used are illustrated in (4), (5), and (6).

\[ \text{Throughput} = \frac{\text{collected data}}{\text{total human game-play time}} = \frac{83 \text{ validated digitizations}}{\sim 3.68 \text{ hours}} \approx 22 \text{ validations per human hour} \] (4)

\[ \text{Average Lifetime Play} = \frac{\text{total human game-play time}}{\text{number of players}} = \frac{\sim 3.68 \text{ hours}}{35 \text{ players}} \approx 6 \text{ minutes per player} \equiv \text{Time needed for 4 game sessions} \] (5)

\[ \text{Expected Contribution} = \text{Average Lifetime Play} \times \text{Throughput} = \sim 6 \text{ minutes} \times 22 \text{ validations per human hour} \approx 2 \text{ validations per player} \] (6)

4.1 Feedback From Players

In order to know how players perceived our two games, two questionnaires were prepared, one for each game. Players were asked to fill the questionnaire corresponding to the game they have played.

The two-player game questionnaire was administered to 22 players, while the single-player game questionnaire was administered to 29 players. All native Arabic speakers with different backgrounds, and age groups between 18 to 40 years old. Other than demographic questions, each questionnaire consisted of
six 5-level likert scale questions, followed by one open-ended question where players were asked to write their suggestions. After collecting the answers, some statistics were calculated as illustrated in Table 1 and Table 2.

**Two-Player Game**

<table>
<thead>
<tr>
<th>Statement</th>
<th>Agree</th>
<th>Neutral</th>
<th>Disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>“I feel the game rules were clear and easy to understand.”</td>
<td>45.4%</td>
<td>18.2%</td>
<td>36.3%</td>
</tr>
<tr>
<td>“The hints provided, when I asked for, were helpful.”</td>
<td>72.8%</td>
<td>18.2%</td>
<td>9%</td>
</tr>
<tr>
<td>“I was interested in seeing the highest scores on the home page.”</td>
<td>45.4%</td>
<td>18.2%</td>
<td>36.3%</td>
</tr>
<tr>
<td>“I enjoyed the game.”</td>
<td>54.6%</td>
<td>18.2%</td>
<td>17.2%</td>
</tr>
<tr>
<td>“I will play the game again in the future.”</td>
<td>59.1%</td>
<td>22.7%</td>
<td>18.2%</td>
</tr>
<tr>
<td>“I feel that 120 seconds for each round of the game were...”</td>
<td>13.6%</td>
<td>63.6%</td>
<td>22.7%</td>
</tr>
</tbody>
</table>

Table 1: Results of the two-player game questionnaire answered by 22 players

It can be noticed that the number of players who thought the game concept is unclear is approximately equal to the number the players who thought the game concept and rules were clear. The duration of 120 seconds for each round in the game session was found suitable and players were interested in looking at the scoreboard of the game. Players found the hints provided by the other party (hinter), when needed, helpful.

As for the open-ended question, 2 players commented that they didn’t understand the game, hence clearer instructions were needed. One player recommended developing a mobile version of the game as well as adding the option of inviting friends to play the game. Another player stated that three hints are too much and suggested to increase the number of rounds per game session.

**Single-Player Game**

<table>
<thead>
<tr>
<th>Statement</th>
<th>Agree</th>
<th>Neutral</th>
<th>Disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>“I feel the game rules were clear and easy to understand.”</td>
<td>75.8%</td>
<td>13.8%</td>
<td>10.4%</td>
</tr>
<tr>
<td>“I was interested in seeing the highest scores on the home page.”</td>
<td>82.8%</td>
<td>13.8%</td>
<td>3.4%</td>
</tr>
<tr>
<td>“I enjoyed the game.”</td>
<td>79.3%</td>
<td>18.2%</td>
<td>17.2%</td>
</tr>
<tr>
<td>“I will play the game again in the future.”</td>
<td>59.1%</td>
<td>22.7%</td>
<td>18.2%</td>
</tr>
<tr>
<td>“I feel that 90 seconds for each round of the game were...”</td>
<td>3.4%</td>
<td>86.2%</td>
<td>10.3%</td>
</tr>
</tbody>
</table>

Table 2: Results of the single-player game questionnaire answered by 26 players

It can be seen from the results that for most players the game concept was clear and easy to follow. The 90 seconds for each game session was found suitable and players were interested in looking at the scoreboard of the game. Additionally, the majority found the game enjoyable and showed interest in playing it again in the future.

As for the open-ended question, players suggested improving the interface, with animations, music, larger decorative Arabic fonts. They also suggested adding clearer instructions to the game. One player said he did not get the purpose of the tool-tip asking whether the two cards are identical or not. Another player stated that the timer made the game very enjoyable and recommended adding the functionality of inviting a friend to play the game.
5 Conclusion and Future Work

In this paper, we propose “tashkeelWAP”, a web application with two games to help in digitizing the Arabic documents without losing the precious diacritic signs. In general, a positive feedback was received regarding both games. Nevertheless, there were some comments that shed light on possible improvements. GWAP evaluations metrics, such as the throughput, average lifetime play and the expected contribution of players were analyzed from both games. Results for the single-player game were promising. However the two-player game needs improvements, as still most players neglect the diacritic signs while digitizing the Arabic words.

In the future, the interface for both games could be enhanced to be more engaging and appealing. More attractive colors as well as associating music and sound effects with certain events can be added to capture the player’s attention. We would like to work on integrating our game with Facebook in order to reach a larger number of players. Additionally, developing mobile-friendly versions of both games could have a positive effect on the number of times the games are played. Also, an animated video can be created to illustrate the rules of the games thus making them clearer to the players.

Specific improvements for the two-player game include displaying a recap page, for 10 seconds, between the two rounds of the game. This will allow the solver to see the scanned Arabic word image and the correctness of his submitted guess. The solver can also notice the diacritic signs missed, which might help in avoiding missing the same diacritic signs again in the future game sessions. Moreover, the situation where odd number of players want to play the game occurs frequently. Therefore, instead of losing this potential player, von Ahn and Dabbish suggested in [7] that this single player can be paired with a set of prerecorded actions.

We could add extra levels for the single-player game to increase the players willingness to play the game more often and reach higher levels. Also, a two-player version of the game could be implemented; where two players will be chosen randomly by the game and paired to compete against each other. The certainty rate needed for a word to be considered as correctly digitized could be increased from 3 to 6 to give more accurate results. Hence, for a word to be considered as correctly digitized, it has to be at least validated by six different players. In addition, machine learning algorithms could be used to assist in determining the certainty rate and the validity of the obtained digitizations.

Acknowledgment

The authors would like to thank the testers, who volunteered to play the games in “tashkeelWAP”, as well as the participants who gave us their time filling the questionnaire. Part of the work was based on a prototype done by the (back then) GUC students: Menna Bakry, Mohamed Raafat and Essam Hafez.

References