Collecting Data for Automatic Speech Recognition Systems in Dialectal Arabic using Games With a Purpose

Dayna El-Sakhawy, Slim Abdennadher, and Injy Hamed

Media Engineering and Technology Faculty,
German University in Cairo
{dayna.el-sakhawy}@student.guc.edu.eg
{slim.abdennadher,injy.hamed}@guc.edu.eg

Abstract. Building Automatic Speech Recognition (ASR) systems for spoken languages usually suffer from the problem of limited available transcriptions. Automatic Speech Recognition (ASR) systems require large speech corpora that contain speech and their corresponding transcriptions for training acoustic models. In this paper, we target the Egyptian dialectal Arabic. As other spoken languages, it is mainly used for spoken rather than writing purposes. Transcriptions are usually collected manually by experts. However, this proved to be a time-consuming and expensive process. In this paper, we introduce Games With a Purpose as a cheap and fast approach to gather transcriptions for Egyptian dialectal Arabic. Furthermore, Arabic orthographic transcriptions lack diacritizations, which leads to ambiguity. On the other hand, transcriptions written in Arabic Chat Alphabet are widely used, and include the pronunciation effects given by diacritics. In this work, we present the game Ma5ame5o (pronouced as makhamekho) that aims at collecting transcriptions in Arabic orthography, as well as in Arabic Chat Alphabet. It also gathers mappings of words from Arabic orthography to Arabic Chat Alphabet.

Keywords: Dialectal Arabic, Speech Recognition, Egyptian Arabic Dialect, GWAP

1 Introduction

Accurate manual transcriptions of speech is an essential ingredient in constructing reliable *Automatic Speech Recognition* (ASR) systems. Both speech corpora and text corpora are needed to train acoustic and language models respectively. Transcriptions are typically done by trained transcribers or experts. This approach has three drawbacks:

1. It is time-consuming. As transcription is done by few people, it is difficult to gather huge amount of data in limited time. Moreover, in the case of using trained transcribers, it takes hours to days to train the transcribers on the transcription guidelines.

2. It is expensive. According to [11], the cost of one hour of transcribed speech can reach $100.
3. It may not generalize to the data at hand. This may happen due to the fact that experts or trained transcribers follow strict guidelines. Valid differences in transcriptions would not be present is such gathered transcriptions.

In order to avoid these problems, researchers have shifted their methodology to gather data using crowdsourcing. Crowdsourcing is the act of outsourcing a task that is computationally difficult to be solved not by experts but rather by the crowd. Crowdsourding has evolved as a successful tool in many fields, such as Natural Language Processing [13] and Machine Translation [14, 15]. It has recently received attention in the field of speech recognition. It has proven to be a reliable and inexpensive way to collect speech transcriptions. In [11, 12], speech transcriptions were obtained using Amazon's MTurk[1]. The gathered transcriptions achieved high levels of agreement with the experts' transcriptions. From a cost perspective, the average cost of collecting speech transcriptions using MTurk is one order of magnitude less than that using traditional methods, as stated in [8].

ASR systems for Egyptian dialectal Arabic suffer from the lack of speech corpus. Egyptian Dialectal Arabic is mostly spoken and not written. As the majority of spoken languages, it has limited available text. Moreover, it does not have a standardized way of writing. It may be written with *Arabic Orthography* (AO) or with *Arabic Chat Alphabet* (ACA), known as *Franco-Arabic*, which uses the English Alphabet with numbers to compensate for extra letters. All those factors affect the availability of speech corpora. The first goal of this work is to collect Arabic orthographic transcriptions.

Another problem researchers face with Arabic orthographic transcriptions is the lack of diacritics. Diacritics represent short vowels, nunation, gemination, and silent letters. In [16], this problem was overcome by using ACA for transcriptions rather than AO. This is based on the fact that it usually includes the short vowels that are omitted in AO. Furthermore, it was found that the majority of computer users type faster in ACA than AO according a survey conducted in [16]. The second goal of this work is to collect transcriptions using ACA.

Finally, the third goal of this work is to collect mapping of words written in ACA to their corresponding form in AO. Such a corpus can be used in the task of converting text from ACA to AO and vice versa.

In this work, we present a Game With A Purpose (GWAP) named Ma  ameo (pronouced as makhamekho). The aim of the game is threefold:

* Collect transcriptions of Egyptian dialectal Arabic using AO.

[1] http://www.mturk.com

* Collect transcriptions of Egyptian dialectal Arabic using ACA.
* Collect mappings of words written in AO to their corresponding form in ACA.

The rest of the paper is organized as follows: In Section 2, an overview is given on the Arabic language, crowdsourcing, Games With a Purpose, and previous work done in gathering data for ASR systems using crowdsourcing. In Section 3, the game Ma3ame6o is introduced. Section 4 presents the evaluation and results. Finally, in Section 5, conclusion and future work are provided.

2 Background

2.1 The Arabic Language

The Arabic language is one of the most popular languages in the world. It is the 6th most used language based on number of first language speakers. There are three types of the Arabic language: classical Arabic, modern standard Arabic (MSA), and dialectal Arabic. The classical Arabic is the standard and most formal type of Arabic. MSA is classical Arabic written without diacritic marks. It is the formal written standard language of education across the Arab world and is used in writing, news broadcast, formal speeches, and movies subtitling. However, MSA is not the language used in everyday life and is considered as a second language for all Arabic speakers. Dialectal Arabic is the language used in informal daily communication. Every country has its own dialect, and sometimes there exist different dialects within the same country. Dialectal Arabic is also used in folktales, songs, movies, and TV shows. Egyptian Arabic is the most widely understood dialect among Arabs, due to the interest and acknowledgment the Egyptian films and TV series gain worldwide [5, 6].

Transcriptions written in Arabic orthography lack diacritization. Diacritics represent short vowels, nunation, gemination, and silent letters. They greatly affect the pronunciation of words. The absence of diacritization can lead to ambiguity. For example, the word مدرسة has 2 valid meanings: school (madrasa) and teacher (modarresa).

It is common among computer users to use Arabic Chat Alphabet (ACA) in typing dialectal Arabic text. As shown in the example above, the ACA forms of the word مدرسة are written as they are pronounced with no ambiguity in the meaning as in AO. ACA usually includes the pronunciation effects of diacritics. In [16], a survey conducted involving more that 100 Arabic computer users. It was recorded that 86% of the users stated that they type faster using ACA, 9% do not feel a difference, and 5% type Arabic letters slightly faster than ACA. All users asserted that it is almost impossible to type a correct fully diacritized Arabic text.

2.2 Crowdsourcing and Games With a Purpose

Crowd-sourcing is the act of taking a task that is usually performed by experts and addressing it to a large, usually online, group of people. The public users then participate in solving the problems available in the open call. The term is a combination of the words: "crowd" and "outsourcing". The idea is to outsource a task to a crowd of people. Four main factors (known as the 4 Fs) were identified to foster the participation in crowdsourcing: Fame (ex. Wikipedia), Fortune (ex. MTurk), Fun (ex. ESP game [7]) and Fulfilment.

Amazon Mechanical Turk (MTurk) system provides a crowdsourcing platform that allows individuals (referred to as Requesters) to use the human intelligence to perform tasks that computers currently cannot do. The Requesters post tasks known as HITs (Human Intelligence Tasks). Workers (referred to as Turkers) can then choose to solve tasks in return of a predefined monetary value.

Up to 150 billion hours (the equivalent of 17 million years of human effort) are spent playing games every year [17]. This gave rise to Games With A Purpose (GWAP). The idea is to design an interesting game that will get people engaged and knowingly or non-knowingly contribute in the collection of data. Some tasks are very easy, trivial for human brain, yet they're still unsolvable by algorithms such as image recognition and speech recognition. The aim of GWAP is to outsource such tasks to humans in a fun way. The collected data can then be used by machine learning techniques to improve algorithms. Examples of GWAPs are ARTigo [18], Tag-A-Tune[19] and the ESP game [7].

2.3 Automatic Speech Recognition & Games With A Purpose (GWAP)

Automatic Speech Recognition (ASR) is a technology that allows a computer to identify the words that a person speaks into a microphone. The ultimate goal is to have a system that would easily recognize the spoken Arabic alphabets and digits regardless of the environmental noise, gender, and dialect [3, 4]. In the past few years, researchers have been investigating the use of crowdsourcing for speech-related tasks. GWAP proved beneficial for 3 tasks: speech transcription, speech acquisition and speech annotation. In this section, an overview on some of the work done in speech transcription using crowdsoucring will be mentioned.

In [10, 11], MTurk was used to gather transcriptions. It was found that transcriptions entered by turkers were very accurate. In [9], Rio Akasaka conducted a study that introduced two tasks. The first task was the accent recognition and the second was the transcription. In Task 1, players were asked to identify the native language of a foreign accented speaker of English as quickly as possible out of four randomly generated choices. In Task 2, players were asked to transcribe short recordings that were randomly selected from the ones available through the CSLU-FAE corpus. Comparing the results of both techniques with

transcription already provided by other users on the server, Task 1 produced 55.26% of the accents accurately identified. As for Task 2, 1093 recordings out of the 1257 available were transcribed. In [1] a study was conducted where audio files were collected from the conversation and segmented into five second utterances. Utterances were assigned in batches of ten per HIT and played with a simple flash player with a text box for entry. The results showed that data collected with Mechanical Turk was nearly effective for training speech models, and that the main focus should be on the number of audio files used rather than focusing on quality of the audio files.In [2], Scott Novotney and Chris Callison-Burch conducted a similar study.

In 2011, a survey [20] on the existing literature was provided. It was shown that there is a growing interest in crowdsourcing for speech processing. There were only 4 publications in 2009, the number increased to 14 in 2010, and then 10 papers in early 2011. The majority of the studies were on speech labeling and transcription (59%), with speech acquisition being the second most frequent topic (27%, and only 5 studies (14%) have used crowdsourcing for assessment of speech technology. Most of the studies (57%) used the AMT for crowdsourcing. 7 of the 37 studies (19%) involved a game from which researchers could obtain the players judgments for free. Other sources of workers include volunteers (14%) and other crowdsourcing platforms (11%). Analysis of the literature also shows homogeneity in the geographical source of papers. Of the 29 papers that were indexed in the survey: 22 are from the United States, 6 are from Europe and only 2 from Asia. As can be seen from the statistics, the work done mainly covered English, with no work covering the Arabic language. Moreover, most of the studies used MTurk while GWAPs were less used.

3 The Game

The aim of this project as mentioned earlier, is to collect Arabic mappings from speech to text written in AO and ACA, as well as mappings from dialectal Arabic words written in traditional Arabic orthography to their corresponding form in ACA. This paper presents a GWAP named Ma$\dot{\smile}$ame$\dot{\smile}$o (pronounced as makhamekho) that helps transcribe as many audio files as possible to Franco-Arabic and Arabic. In this section, a description of the game will be given.

Ma$\dot{\smile}$ame$\dot{\smile}$o is a single-player game where the player hears an audio-file and is required to give the corresponding Franco and Arabic transcription. Audio-files used are of 5-7 seconds and there is no limit to the number of times the player can hear the audio-file. The audio files where chosen to have a few seconds' duration to avoid short-term memory saturation. The game page shows the audio-file, text-areas to enter player's transcriptions and top 2 transcriptions to choose from are displayed. It also shows the label for the player's current skill. The player listens to the audio and needs to transcribe what he heard either by entering his own transcriptions in the text-areas or choosing one of the top transcriptions displayed as shown in Figure 1. If the user chooses to enter his

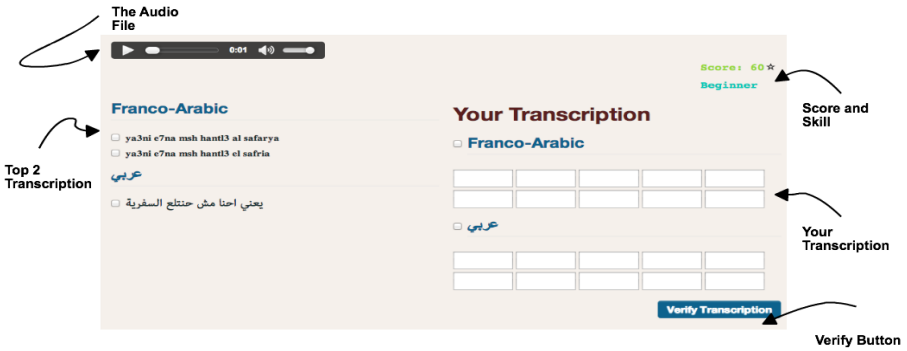

Fig. 1. Game Page

own transcription, he/she must make sure that the number of words in Franco is equal to that in Arabic. The Top 2 transcriptions are the highest transcriptions entered by other users for this audio.

3.1 Incentives

In order to make the game more creative & challenging, the following incentives were added to the game :

Categories

In order to accomadate for different users' interest, the game includes different categories to increase the variety of the game and make it more entertaining. The categories were obtained by conducting a survey in which a pool of people were asked to choose their favorite categories from a pool of predefined categories. The provided choices were: Arabic-Movies, Arabic-Series, Arabic-Cartoon, Famous Arabic-Quotes, Arabic-Songs, Arabic-Proverbs and Arabic-TalkShows. As shown in Figure 2, Out of fifty five people who took the survey, twenty three chose Arabic-Movies, thirteen chose Arabic-proverbs, while nine chose Arabic-songs, six chose Arabic-cartoons and finally only 2 chose Famous Arabic-Quotes and Arabic-series .The top 4 categories were chosen for the game (Arabic-Movies, Arabic-Proverbs, Arabic-Songs and Arabic-Cartoon).

Score

To make the game more engaging and motivate the user to enter correct transcriptions, a score is given to the players. Score is calculated as follows:

* If the user chooses the transcription with highest verification 15 points are added to the score .
* If the user chooses the second highest transcription 10 points are added to the score .
* If the user enters a new transcription in the text area 5 points are added. Furthermore, 5 points are added every time a player chooses this transcription from the list of the top 2 transcriptions.

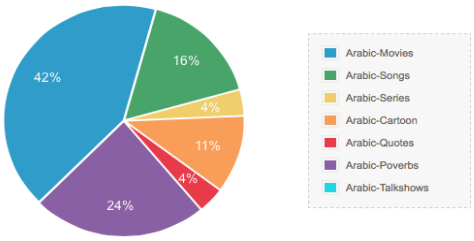

Fig. 2. Survey results

Highest Verification is calculated by counting the number of repeated Arabic and Franco transcription for a given audio-file . Accordingly the most repeated Arabic and Franco transcription is considered to be the transcription with the first **highest Verfication**. The second most repeated Arabic and Franco transcription is considered to be the transcription with the second **highest Verfication**.

Skill

The player skill is levelled up according to his/her score.

* Beginner: At the start of the game.
* Amateur: Score more than 100 points.
* Semi-pro: Score more than 200 points.
* Expert: Score more than 500 points.
* Pro: Score more than 1000 points.

Hall Of Fame

The Hall of Fame page will include the player of the day, top 6 monthly ranking and top 3 scores in the game. Finally levelling the player's skill after scoring some points will make the players eager to score as many points as possible to level up.

3.2 Collected Data

In this game, two types of data are collected:

- Transcriptions using AO and ACA. These are gathered from users by either entering them directly into the text areas or choosing from the top 2 transcriptions previously given to the audio file. This is done for each of AO and ACA.
- Mappings of AO-ACA words. This is done by making sure that the number of words used in transcription are equal for both cases, AO and ACA. This gives a one-to-one mapping of words. For example, the word الكتاب will be written in one text area in ACA as "Al Kitab".

3.3 Game Framework

The Framework used is **Play Framework 2.2.0**. Play framework is a framework that is inspired from Ruby on rails and Django and follows the model-view-controller (MVC) architectural pattern. It uses both Java to design the backend part and HTML to design the frontend. Play Framework provides an object-relational mapping product written in Java called EBean. This Object-relational mapping along with RawSQL were used to design relationship between the models and create the game's database.

4 Evaluation & Results

The evaluation for Ma3ame3o was held in a duration of two weeks. In this period, 118 users played the game, with a total of 1121 game rounds. 120 audio files were transcribed with 1121 Arabic orthographic transcriptions and 1121 transcriptions in the Arabic Chat Alphabet.

It was observed that in the case of AO, each audio file was transcribed with 1-2 unique transcriptions. However, in the case of ACA, the number was much higher. Out of the 1121 transcriptions, there are 602 unique transcriptions for the 120 audio files. This gives an average of nearly 5 unique transcriptions for each audio file. This observation could have two possible interpretations:

- There is more variation in typing in the case of ACA than AO.
- Users might be more used to typing in ACA than AO. Therefore, they tend to choose from the top 2 transcriptions in the case of AO. This assumption could be valid for the age group of 20-25. However, in the age group of 40-50, users are usually familiar with typing in AO. Therefore, this interpretation is less likely.

It was also observed that the chosen category depends on the age group. For the age group 20-25, users mostly chose songs and films. For the age group 40-50, the proverbs category was the most popular.

Validating the correctness of the gathered transcriptions
The ability of Ma3ame3o to collect the correct transcriptions was evaluated on a small sample of the audio files. Some of the participants were asked to provide transcriptions for 5 audio files. The audio files were selected for each participant to be in different categories than those chosen in the game. The transcription provided by each participant for an audio file was checked whether it was gathered from the game or not. All the transcriptions were found to be collected through the game. This shows that the game is capable of collecting correct transcriptions. Moreover, the data collected includes the different variations in typing, which is one of the advantages of using the crowd rather than experts to collect transcriptions.

Validating the correctness of the gathered mappings
The 8 audio files receiving the highest number of different ACA transcriptions were investigated in this evaluation. For each file, the mapping with the most popular ACA transcription was validated. Fifty users were asked to rate each of the 8 mappings on a four-scale. They were asked to rate the mapping of the provided ACA with the corresponding AO. On average, 53.825% of the ratings were selected to be *strongly agree*, 44.625% were *agree*, and only 1.5% of the ratings were given *fair*. These figures show that the game is able to gather mappings of ACA to AO transcriptions. This is only a preliminary evaluation. Further testing should be done on the rest of the mappings gathered. Testing should also be done on word-to-word mappings.

Questionnaire
In order to asses the players' satisfaction with the game, 50 participants filled in a questionnaire. Participants were asked to use a four-scale (strongly agree, agree, disagree and strongly disagree) to answer the following 2 questions:

1. Did you find the game interesting?
2. The game is easy to play?

The participants' feedback on the these 2 questions is shown in Figure 3. All participants stated they would play the game again.

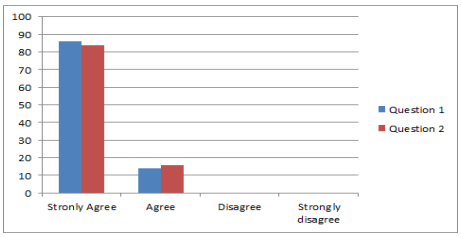

Fig. 3. Participants' feedback on Questions 1 (Did you find the game interesting?) and Question 2 (The game is easy to play?)

5 Conclusion

Automatic Speech Recognition systems for Egyptian Dialectal Arabic suffer from the lack of existing speech corpora needed for training. Moreover, most available transcriptions do not include short vowels and diacritics that reflect differences in pronunciation which leads to ambiguity. Transcriptions in Arabic Chat Alphabet (ACA) include the pronunciation effects given by diacritics. Ma3ame3o was proposed as a Game With a Purpose that aims at collecting: 1) transcriptions in AO, 2) transcriptions in ACA and 3) Mappings of words from AO to ACA. The game allows people to transcribe many audio files by playing an interesting and challenging game. Results show that the game succeeded in collecting many correct transcriptions as well as mappings from AO to ACA. Participants found the game interesting and easy to play and confirmed that they would play it

again. Further evaluations should be done to accurately assess the correctness of the gathered data.

References

1. D. Suendermann, J. Liscombe, R. Pieraccini, How to Drink from a Fire Hose: One Person Can Annoscribe 693 Thousand Utterances in One Month(2010)
2. S. Novotney and C. Callison-Burch, Cheap, Fast and Good Enough: Automatic Speech Recognition with Non-Expert Transcription (2009)
3. Y. Delendik, What is Automatic Speech Recognition? (June 2009)
4. S. Furui, AUTOMATIC SPEECH RECOGNITION AND ITS APPLICATION TO INFORMATION EXTRAC.TION (2001)
5. D. Verguria and K. Kirchhoff, Automatic Diacritization of Arabic for Acoustic Modeling in Speech Recognition (2014)
6. P. Macmillan , Sacred Language, Ordinary People: Dilemmas of Culture and Politics in Egypt. Palgrave Macmillan(2003)
7. L. von Ahn and L. Dabbish , Designing Games With A Purpose(2008)
8. G. Parent, M. Eskenazi, TOWARD BETTER CROWDSOURCED TRANSCRIPTION: TRANSCRIPTION OF A YEAR OF THE LET'S GO BUS INFORMATION SYSTEM DATA(2010)
9. R. Akasaka , Foreign accented speech transcription and accent recognition using a game-based approach(2009)
10. M. Marge Satanjeev ,B. Alexander ,I. Rudnicky , Using the Amazon Mechanical Turk to Transcribe and Annotate Meeting Speech for Extractive Summarization (2010)
11. M. Marge Satanjeev ,B. Alexander ,I. Rudnicky , Using the Amazon Mechanical Turk for Transcription of Spoken Language(2010)
12. K. Evanini, D. Higgins, and K. Zechner , Using Amazon Mechanical Turk for Transcription of Non-Native Speech(2010)
13. R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng, Cheap and fast - but is it good? Evaluating non-expert annotations for natural language tasks, in Proc. EMNLP, 2008, vol. 1, pp. 254263.
14. M. Denkowski, H. Al-Haj, and A. Lavie, Turker-assisted paraphrasing for English-Arabic machine translation, in Proc. NAACL-HLT, 2010, pp. 6670.
15. V. Ambati and S. Vogel, Can crowds build parallel corpora for machine translation systems?, in Proc. NAACL-HLT, 2010, pp. 6265.
16. M. Elmahdy, R. Gruhn, S. Abdennadher, W. Minker, "Rapid phonetic transcription using everyday life natural Chat Alphabet orthography for dialectal Arabic speech recognition," Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on , vol., no., pp.4936,4939, 22-27 May 2011.
17. J. McGonigal , Reality is broken : why games make us better and how they can change the world. New York: Penguin Press; 2011.
18. C. Wieser, et al. "ARTigo: Building an Artwork Search Engine With Games and Higher-Order Latent Semantic Analysis." First AAAI Conference on Human Computation and Crowdsourcing. 2013.
19. L. M. Law, "TagATune: A Game for Music and Sound Annotation." ISMIR. Vol. 3. 2007.
20. G. Parent and M. Eskenazi. "Speaking to the Crowd: Looking at Past Achievements in Using Crowdsourcing for Speech and Predicting Future Challenges." INTERSPEECH. 2011.