ISSN 1870-4069

Proposed Novel Algorithm for Transliterating Arabic Terms into Arabizi

Nabeela Altrabsheh, Mazen El-Masri, Handay Mansour

University of Qatar,
College of Business and Economics
and Arabic Department,
Qatar

{nabeela, mazen.elmasri, hanadyma}@qu.edu.qa

Abstract. Arabizi is a new trend in social media where the person uses Latin characters to represent Arabic words. Arabic letters can be replaced with different symbols according to the dialects and preference. This creates a wide range of new vocabulary in sentiment lexicons. All the Arabizi literature focus has been on transliteration of Arabizi terms into Arabic terms. In this paper, we propose a new algorithm to transliterate Arabic terms into Arabizi. This allows a larger coverage of the different ways words are written in Arabizi, which will allow a more accurate analysis of the sentiment.

Keywords: Arabizi, arabic translator, arabizi literature.

1 Introduction

Arabizi is to write Arabic text using Latin characters. Arabizi can be used to write both Modern Standard Arabic (MSA) or Arabic dialects. Arabizi can be used for many reasons including technical limitations such as the keyboard not containing Arabic letters [2]. Another reason people like to use Arabizi is because it is free of errors and there are no typos as people can write words as they like [2]. People use Arabizi to write the letters as they pronounce them in real-life [1]. Arabic letters are pronounced differently according to the different dialects. These sounds could be expressed in Arabizi. Arabizi is dependent on the local dialect and differs from one country to another [9]. One example is the letter "ذ thal" which sometimes can be pronounced and written as 'z' in Egyptian dialect and 'th' in other dialects [3].

Another example is the letter "ق Qaf" which is pronounced 'Ga': Jordanian, 'Qa': Iraqi, 'Ka': Palestinian, and 'A': Lebanese [3, 9]. Another reason users type in Arabizi is that users are not familiar with typing in Arabic. Many studies show that people tend to write more in English these days [3, 9]. Therefore users find it difficult to switch between keyboards and find it easier to use Arabizi. Arabizi may also include some English abbreviations such as 'LOL' meaning laugh out loud. They may also include Arabic abbreviations such as 'ISA' meaning "In Sha Allah" and 'MSA' meaning "Ma Sha Allah".

Table 1. Arabizi letters.

Arabic	Arabizi				
ا	2	a	‘		
ؤ ئ أ ء	2				
ث	th	s			
ح	7	h			
ج	j	g			
خ	5	7	kh	x	
ذ	z	th	dh		
ش	sh	ch	$		
ص	s	9			
ض	d	9	dh		
ط	t	6			
ظ	th	6	dh	z	
ع	3	a	o		
غ	3	gh			
ق	8	q	g	2	9
و	w	o	ou		
ي	y	i	e		

Short vowels in Arabic consist of: fatha; a diagonal stroke written below the consonant which represents ‘a’, kasra; a diagonal stroke written below the consonant which represents ‘y’, ‘i’ or ‘e’, and damma; an comma shape written above the consonant which represents ‘u, ‘ou’, or ‘o’. When Arabic is typed online, short vowels are not usually written in the text. On the other hand, when writing Arabizi, users usually include the vowels. This makes it more easier for foreigners to read Arabic script [2]. Not writing the vowels in Arabic makes it difficult to transliterate text into Arabizi. Often a vowel changes the meaning of the word in Arabic, therefore, they are essential for transliteration. One example for this is the word حب which can be “Hubb” with the damma meaning love or “Habb” with the fatha meaning seeds.

Another issue with transliterating Arabizi to Arabic is that people tend to use both Arabizi and English in their sentences making it difficult to distinguish between Arabic and English words [3, 4, 2]. One example for this is the word ‘men’ which means ‘from’ in Arabic and is also a English word. There are only a few studies on Arabizi in the literature [3, 5, 1, 9]. Only one of the studies studied sentiment analysis on Arabizi [5]. All of research in Arabizi data was focused on transliterating Arabizi terms to Arabic and distinguishing between Arabizi and non Arabizi.

Table 2. Lexicons.

Database Name	Research	Size
ArabicSentimentLexicon	Mahyoub et al.	15110
Harvard Lexicon	Stone	1662
Arabic translation of Bing Liu's Lexicon	Mohammed et al.	6789
Arabic translation of MPQA Subjectivity Lexicon	Mohammed et al.	7619
Arabic translation of NRC Emoticon Lexicon	Mohammed et al.	26740

As there are many varieties of Arabizi letters that can be used, it is important for us to capture the different ways a single Arabic letter could be written in Arabizi. In this research, we propose an algorithm to transliterate Arabic terms into Arabizi. This will allow a larger coverage of Arabizi terms for a single Arabic term. Additionally, this algorithm will facilitate Arabizi sentiment analysis in social media and will lead to a more accurate analysis. One contribution of this paper is to create a Arabizi sentiment lexicon. Duwairi et al. [6] highlighted that there does not exist any Arabizi sentiment lexicon available. Related research is outlined in section 2. The research methodology including Arabizi letters, the data collected for this research is highlighted in section 3. In section 4, we present our proposed algorithm solution. Section 5 outlines the results and discussion, followed by the conclusions and future work in section 6.

2 Literature Review

Some researchers identified Arabizi from other languages such as [8, 4]. Others researched Arabizi transliteration such as [3, 1]. We will summarise these studies in the following paragraphs. Tobaili [8] research identifies Arabizi from multi-lingual data in Twitter. They collect two datasets from Twitter; one from Lebanon and the other from Egypt using geographic information. The data contained Arabic and non Arabic Tweets. Langdetect which is a library that detects one or more languages for a given sentence was used to detect the language.

From a sample of non Arabic Tweets, they extracted 3,707 Lebanese Arabizi and English tweets and 3,823 Egyptian Arabizi. Their results show that in Lebanon, the usage of Arabic and Non-Arabic tweets are nearly equal. On the other hand, in Egyptian tweets, Arabic is dominant. Arabizi tweets from Lebanon tend to be mixed with English and French, which is different from Egypt. In addition, Egyptians sometimes abbreviate Arabizi words in most cases by avoiding to write the vowels.

For example "2na w enta"- meaning you and I, which is written as "na w nta" in the Egyptian Arabizi text. They use SVM as a classifier and perform two experiments to distinguish between English and Arabizi tweets. The features included the languages detected by Langdetect, the language detected by Twitter API, and the count of word occurrences per tweet. The first experiment was with unbalanced data with the Arabizi being only 12% for Lebanon and 25% for Egypt.

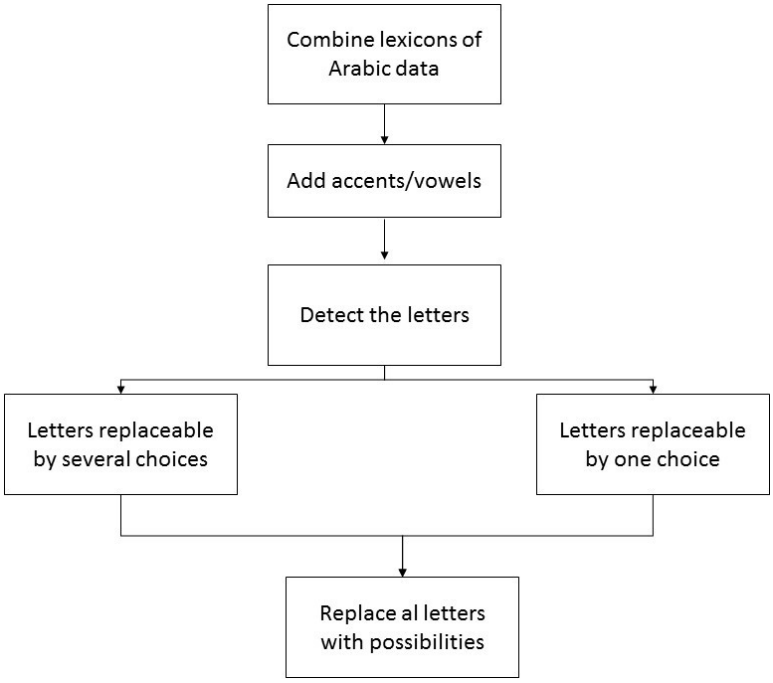

Fig. 1. Arabizi methodology.

This experiment resulted in a 93% accuracy. For the second experiment, undersampling was done and an accuracy of 97% was achieved. Darwish [4] presented methods to detect Arabizi from English text and transliterated Arabizi text into Arabic. For Arabizi detection, they used a sequence labeller which included word and character level features. They collected tweets using three keyword queries 'e7na' (we), '3shan' (because), and 'la2a' (no). They built a word list from tweets including 112 million Arabic tweets. An overall accuracy of 98.5% was achieved for Arabizi detection.

The second part of their research was to convert Arabizi to Arabic. A transliteration miner was trained to find the most likely Arabic word for the Arabizi word. Character transliteration probabilities and language modeling were used. For transliteration, they achieved a 88.7% accuracy. Bies et al. [3] research explored Arabizi-Arabic script transliteration. They collected SMS/Chat Corpus from 26 Egyptian participants. A number of 2,140 conversations and 475K words were collected. They also collected SMS from Broad Operational Language Translation program.

They automatically transliterated Arabizi into Arabic form. Annotators transliterated token by token, sentence by sentence and reviewed the corrected transliteration in full context. The second task was to correct spelling in the Arabic script transliteration to CODA (Conventional Orthography for Dialectal Arabic) standards. CODA minimizes differences between dialects and uses similarities and rules to combine them [7].

Table 3. Arabizi letter occurrences.

Letter	Occurrences
أ	25897
ث	1456
ح	5541
ج	4901
خ	2682
ذ	1027
ش	3592
ص	3249
ض	1932
ط	3332
ظ	709
ع	7587
غ	2500
ق	6122
ء	2542
و	11231
ي	16901

To correct spelling, different steps were taken. Firstly, the Egyptian words that had undergone sound changes such as the word 'mgfwl مقفول' which is sometimes written as 'm2fwl مأفول'. Some long vowels in words were written as short verbs such as 'qalt قالت' written as 'qlt قلت' . Some consonant letters are often mixed up with other consonants. For example the letters 'س and ص'.

Annotators corrected any mistakes of this type. They were also asked to correct errors from morphological ambiguities such as 'byemla بعملوا' and 'byemlh بيعمله'. Any word that was incorrectly segmented was corrected, such as 'fAl byt فال بيت' which should be 'fAlbyt فالبيت'.

Abbreviations such as 'INA' meaning 'In Sha Allah' were also changed to the full word form and typos in words were also dealt with. Lastly, words which were transliterated from English into Arabic such as "Have fun هف فن" were corrected. Al-Badrashiny et al. [1] presented a method to transliterate dialectal Egyptian Arabizi to Arabic by following CODA. The process they use is as follows: Arabizi sentences are input in their created system (3ARRIB).

Table 4. Letter occurrences in words.

Letter	Words
0	1410
1	8060
2	13986
3	10298
4	4685
5	2003
6	658
7	184
8	25
9	6
10	3

Preprocessing steps such lowercasing (de-capitalization), speech effects handling, and punctuation splitting are utilised. 3ARRIB creates a list of all possible transliterations for each word in the input sentence. This is by utilising a finite-state transducer that is trained on character-level alignment from Arabizi to Arabic text. They experiment with a morphological analyser for Egyptian Language (CALIMA) and a Language Model (LM). Several experiments were implemented. Consequently, they found that 3ARRIB with a 5-gram tokenized LM performed best, leading to an accuracy of 77.5% without normalisation and 79.1% with it.

3 Research Methodology

Yaghan [9], Attwa [2] Al-Badrashiny et al. [1] research covered the Arabic letters and their Arabizi alternatives. In this research, we combine these Arabizi letters and create a list with all the Arabizi replacements possible for each Arabic letter. An Arabic letter can be written in different English letters, numbers and characters. Most of the letters are easy to replace. However, there are some letters that are replaced differently according to the dialect, region and personal choice. These letters are illustrated in Table 1. There are 14929920 ($2\times2\times2\times2\times4\times3\times3\times2\times3\times2\times4\times3\times2\times5\times2\times3\times3$) possible combinations if all letters existed in the same word or phrase.

3.1 Data Collection

Data was collected from different online lexicons that are publicly available The total amount of words combined is 41318 Modern standard Arabic words. The database information is shown in Table 2. An amount of 16602 of these words were repeated, so we took the majority of polarity labels and dismissed the repeated instances. For example, if the word was negative twice and positive the once. We would keep one instance of the world with the negative polarity.

Some of the lexicons used scores; if the score was more than 0 it would be positive and less than 0 it would be negative. To combine all the tables we used only labels (i.e. positive, negative and neutral) and dismissed the scores.

3.2 Proposed Algorithm

In the beginning of this research, we attempted to use Darwish et al. [4] algorithm to convert the words from Arabizi to Arabic. Arabic vowels are fatha which is similar to 'a', kasra which is similar to 'e', and damma which is similar to 'o' [2]. However, we found when writing in Arabizi the vowels exist in the word. However in Arabic, the vowels are usually not typed into the writing and people read it intuitively [9]. We propose a new algorithm and follow the process illustrated in Figure 1.

1. Add diacritics to the words using an online tool. Miskhal tool uses a kaleidoscope linguistic approach based on phases of morphological analysis, grammar and semantic access[1].
2. Create a program that replaces each of the Arabic letters to all possible English letters, numbers, or characters.

One example to illustrate our method is the word: اتفاق. This can be written numerous ways as: etefaq, etefag, etefak, etefa8, etefa2, or etefa9.

4 Results and Discussion

We performed several analysis on our lexicon and found the following observations. The total number of words from the lexicon, taking into consideration all the possibilities of letters were 1426010. In the lexicon the most occurring letter was ا alef and the varieties of ء hamzas. The amount of letters occurrences (from Table 1) in the lexicon are presented in Table 3. We found that in most words only have two letters from Table 1. 1410 of the words in the lexicon did not contain any of the letters from Table 1. The amount of occurrences of letters in a single words are presented in Table 4. We found that the largest number of letters occurring together in the same word/phrase were 10 letters. The largest amount of possibilities for a single word/phrase was 38880=2×4×3×2×4×3×2×5×3×3 with the phrase:

السيكوباتي شخص مضطرب العقل "Alsekweyati is someone with mental problems".

This phrase included 10 letters ا، خ، ش، ص، ض، ط، ع، ق، و، ي.

Our proposed method is novel and can not be compared to any previous research. This research can facilitate in the creation of many new sentiment lexicons. It can also aid future systems for Arabic-Arabizi transliteration.

[1] tahadz.com/mishkal/index

Although the research has provided us with many possibilities for the same word. It is a weak approach as some of the words may not be used or written in real life. People tend to use certain combinations of letters, therefore, in future work we aim to find these combinations and narrow down the possibilities based on them. Also, another approach that can be used is to compare the lexicon with online words through social media websites (i.e. Twitter).

5 Conclusion and Future Work

In this research we explored transliterating Arabic terms into Arabizi. One of our main contributions of this research was creating an Arabizi lexicon which to the best of our knowledge has not been done before. In Arabic writing, mainly the vowels fatha, kasra and damma are not written, which makes it difficult to transliterate into Arabizi. Also, there are many alternative English letters, numbers and characters that can be written in Arabizi. We proposed a new method and algorithm for transliterating Arabic text into Arabizi. A lexicon was collected from different sources and transliterated into Arabizi. The number of words originally in the lexicon were 41318 Arabic words, the Arabizi transliteration for this lexicon was 1426010. We found that 17 letters had more than one possible replacement. We found that most words contained at two of these letters. For future work, we plan to create a search algorithm to find if there are any of the transliterated words from our system had previously been used online. This will allow us to discard Arabizi words that are not used or used less frequently.

Acknowledgments. This publication was made possible by the NPRP award [NPRP 7-1334-6-039 PR3] from the Qatar National Research Fund (a member of The Qatar Foundation). The statements made herein are solely the responsibility of the author[s].

References

1. Al-Badrashiny, M., Eskander, R., Habash, N., Rambow, O.: Automatic transliteration of romanized dialectal arabic. In: Proceedings of the 18th Conference on Computational Natural Language Learning. pp. 30–38 (2014)
2. Attwa, M.: Arabizi: A writing variety worth learning. Ph.D. thesis, The American University in Cairo, School of Humanities and Social Science (2012)
3. Bies, A., Song, Z., Maamouri, M., Grimes, S., Lee, H., Wright, J., Strassel, S., Habash, N., Eskander, R., Rambow, O.: Transliteration of arabizi into arabic orthography: Developing a parallel annotated arabizi-arabic script SMS/chat corpus. In: Proceedings of the Empirical Methods in Natural Language Processing Workshop on Arabic Natural Langauge Processing. pp. 93–103 (2014)
4. Darwish, K.: Arabizi detection and conversion to arabic. In: Proceedings of the Empirical Methods in Natural Language Processing Workshop on Arabic Natural Language Processing. pp. 217–224 (2014)
5. Darwish, K., Magdy, W., Mourad, A.: Language processing for arabic microblog retrieval. In: Proceedings of the 21st Association for Computing Machinery International Conference on Information and Knowledge Management. pp. 2427–2430 (2012)

6. Duwairi, R.M., Alfaqeh, M., Wardat, M., Alrabadi, A.: Sentiment analysis for arabizi text. In: International Conference on Information and Communication Systems. pp. 127–132 (2016)
7. Habash, N., Diab, M.T., Rambow, O.: Conventional orthography for dialectal arabic. In: Proceedings of the 18th International Conference on Language Resources and Evaluation. pp. 711–718 (2012)
8. Tobaili, T.: Arabizi identification in twitter data. In: Proceedings of the Association for Computational Linguistics Student Research Workshop. pp. 51–57 (2016)
9. Yaghan, M.A.: Arabizi: A contemporary style of arabic slang. Design Issues 24(2), 39–52 (2008)