Iraqi Colloquial Dialect Dictionary Corpus Creation, Analyzing and Translation

Zinah Abdulridha Aboodiheen1, Rand Abdulwahid Albeer1, Aliaa F. Mousa2, Nisreen M. Hamza1, Dhamyaa A. Nasrawi*,1

1Computer Science Department, College of Science, University of Kerbala, Iraq
2General Directorate of Education in Kerbala, Ministry of Education, Iraq

*dh.alnasrawy@uokerbala.edu.iq

Abstract. The Iraqi dialect is one of the most beautiful dialects in the Arabic world. It contains a wide variety of vocabulary as well as phrases spoken in Iraqi population. The aim of this paper is to create, analysis, and translate of Iraqi Colloquial dialect corpus. The hardest step is to create the Iraqi corpus that were generated by writing of Iraqi films or stories and published them in GitHub. There are two phases: analyzing, which implement pre-processing includes: tokenization, remove stop words, remove punctuation, and remove duplicated words. The purpose of analyzing process is to collect the words to make Iraqi-Arabic-English dictionary. In the second phase is: testing, which involve searching about Iraqi word, translating it to Standard Arabic and English, adding it if not exist, predicting Iraqi words that depends on similar in spell, getting Iraqi synonyms words, and display the whole dictionary. New method was suggested to cover the problem of unofficially written for spoken dialects, as well as new algorithm to find the Iraqi words synonyms.

Keywords: Colloquial, corpus, dictionary, Iraqi dialect, tokenization, translation

1. Introduction to Arabic language and dialects

Arabic language is a widely used global language that has major differences from most popular languages, e.g., English and Chinese. It has many grammatical forms, varieties of word synonyms, and different word meanings that vary depending on factors like word order. Unlike most languages, Arabic is written from right to left, with no capitalization, and with 28 alphabetical characters as well as diacritics. According to Nizar Habash et al., there are multiple forms of the Arabic language such as[1,2,3]:

- Classical Arabic – which is used in reading / reciting the holy books.
- Modern Standard Arabic (MSA) – Standard Arabic, which is commonly used in writing, speech, interviewing, broadcasting, etc.
- Spoken – oral dialects that vary significantly from region to region.
Although the official language in the Arab home is MSA, the Arab countries which both of them has own accent that is descended from MSA, it is considered a separate language. It is a collective term for the spoken languages or dialects of people throughout the Arab world.

The Arabic dialect language is a collection of spoken accents with important morphological, phonological, lexical, and syntactic differences, along with a standard written language, Modern Standard Arabic (MSA).

Since the spoken dialects are not officially written and do not have standard orthography (Lack of syntactic rules there are no identified grammar rules for colloquial dialects) it is very costly to obtain adequate corpora to use as a dataset in researches[1,4].

Speakers of some of these dialects are unable to understand some speakers of other Arabic dialects such as Algeria, Libya, Sudan. Anyway, the rate of colloquially written text increases dramatically in the last time [5,6]. Iraqi dialect corpus has been used in this paper because the authors are from Iraq.

The vast geographical area of Iraq and its different environments open the way for a wide variety of dialects and divergent methods of speech. Despite this divergence and diversity in Iraqi dialects, there are general features common among the various populations of Iraq. Therefore, the dialect of middle in Iraq was considered in this work.

Dictionary usually contains terms for more languages that help learner to understand meaning these terms by entered them in his mother language, then the dictionary converts the entered term to language that learner wants it. There are many available dictionaries for everyone whether free or not, online or not such as Google translator, golden alwafi, sakhr, and so on. All these translators did not contain dialects in its terms. So, the aim of this work is to make a dictionary that deals especially with Iraqi dialect, after create, pre-processing, and analysis of Iraqi corpus. This dictionary converts Iraqi dialect words into Modern Standard Arabic (MSA) and into the English language. Also, the user can add words that not found in this dictionary. As well as, it contains a lot of options that make it interesting and useful.

The rest of this paper, section 2 shows the related works, in section 3 steps of proposed work are explained, while section 4 contains conclusion and future works.

2. Related works

Several methods have been used to deal with Arabic dialects, in [6], new approach was introduced by Khaled Shaalan and et al. That depends on a rule-based lexical transfer to convert Egyptian colloquial words into MSA words. This process involves morphological analysis and lexical acquisition of colloquial words.

Kareem Darwish and et al used Conditional Random Fields (CRF) to automate the diacritization process of Tunisian and Moroccan dialects, they achieved word-level diacritization errors of 2.9% and 3.8% for Moroccan and Tunisian respectively[7].

Multi distance metrics approaches were used in [8] by Athrein Abu Kwaik and et al. to measure the lexical distance between MSA and Arabic dialects, such as Vector space model (VSM), Latent Semantic Indexing (LSI), and Divergence Distance algorithm.

BezouiMouaz and et al. used Hidden Markov Models to build a speech recognition system, the proposed system was applied to a Moroccan dialect speech. The accuracy of proposed method is about 90%[9].

In this work, the importance of this research is to create data set (documents in Iraqi colloquial dialect), and studying, analyzing, and learning, the Iraqi colloquial dialect.

3. Methodology

In this section, focus on the methodology of implement the steps of creation, analyzing and translation of Iraqi colloquial corpus. Firstly, we explain the features of our software which implemented using python3.5 programming language. The main features are:

- Display all words in dictionary.
- Looking for Iraqi words, then get the translated word in standard Arabic and English.
- Adding new words to dictionary.
- Display words in categories according to familiar words, Iraqi folk proverbs, Iraqi cuisine, professions, tools, etc.
- Display cloud word of selected file.
- Implement pre-processing to each selected file.
- Display Iraqi synonym
- Display Iraqi words Predicting

This work relies on pre-existing data called analyzing data, through which the input word is tested which is called test data. We associate the test data (input word) with analyzing data to know the result obtained from this application through a proposed algorithm that match between entered word with analyzing data. Figure 1 explain proposed work.

![Figure 1. Methodology of proposed work](image)

Now, each phase was explained in details:

3.1. **Analyzing phase**

This phase consists multiple steps as figure 2 show the steps of analyzing phase:

![Figure 2. Analyzing phase](image)
The details of this phase demonstrated in the following subsections.

3.1.1. Create Iraqi colloquial corpus
A corpus is an electronic form of collection of texts (written or spoken) language. It helps in using the language in real situations as a directory which is used by lexicographers to write meaningful and accurate dictionary entries [10].

This corpus was created from Iraqi films (movies) that converted into text was a good idea, i.e. convert speech to text by online software to get text documents to represent the corpus. There is multiple software such as Google doc, Google translation, and other online software but the results were very poor so we wrote films manually which is hardest process in this project. The total words of our corpus are approximately 50,000 words without pre-processing, or stories published on GitHub.

3.1.2. Pre-processing
Pre-processing must be implemented on this corpus for investigating the purpose of this research. Pre-processing includes tokenization, remove stop words, remove punctuations, and remove duplicated words.

- **Tokenization:**
  Tokenization is the task of chopping strings up into pieces, called tokens, perhaps at the same time throwing away certain characters, such as punctuation. In this step, tokenization process takes text files (Iraqi corpus) and run suitable instructions by NLTK library in Python. The output of this process is a list of words, see figure 3.

- **Remove stop words and punctuation marks**
  This step was implemented to get pure words for Iraqi corpus, figure 4 shows the list of Arabic stop words which removed from files.

- **Remove duplicated words**
  For all corpus may be get duplicated words, this step removes them.

- **Extract Iraqi's words**
  After we get pure words, we extract Iraqi's words manually to add them to vocabulary.

3.2. Testing phase
In this phase, more test can be implemented to show the performance of the project. The main feature of the Iraqi dictionary is searching method about Iraqi terms. The problem in colloquial dialects is that no identified grammar rules because they are not officially written and do not have standard orthography. Therefore, a new algorithm was proposed to search about Iraqi terms to check the existence of them in the dictionary.

This algorithm is depended on threshold basely when compared input word with each word in training data, and deference of length between them. We take the value of threshold is 1 that represents the difference between the two lengths because if we take more than 1 maybe it brings word from training data irrelevant. If the difference was 1, then the algorithm compares between letters (word entered and word in training data). This step came to reduce run and CPU times. The outputs of this algorithm are predicting all Iraqi words that matching completely or partially the input word.

**Algorithm (1): Matching words**

**Input:** set of words after preprocessing set_words[1..n] (training), input_word (testing).

**Output:** set of matching words or one matching word

**Step 1:** let \( L_{\text{input word}} \) is length of input_word

**Step 2:** let \( L_{\text{set word}} \) is length of set_words,

**Step 3:** let \( L_{\text{difference}} \) is absolute deference length between \( L_{\text{input word}} \), and \( L_{\text{set word}} \).

**Step 4:** if \( (L_{\text{difference}} < \text{threshold (1)}) \) then

\[
M = \text{compare (input_word, set_words)}
\]

\[
L_m = \text{abs}(M - L_{\text{input word}})
\]

if \( (L_m != 0) \) then

if \( (L_m < \text{threshold (1)}) \) then

print(set_words)

else

print(‘Match 100%’)

**Step 5:** End

Example: searching for Iraqi word "شلونكم" in Arabic, "كيف الحال" in English “how are you”?", the output of this algorithm are all words in dictionary that similar in spell "شلونكم، اشلونكم، شلونك". The main features of designed software with example are illustrated in the following GUI, see figure 5:

![Figure 5. GUI of Testing Phase](image)
• Search: First feature is search about Iraqi words, then get the translated word to standard Arabic and English. Also, we can search about standard Arabic words to get the pairs in Iraqi words, figure 6 shows the English translation after search about Iraqi word (شلونكم).

![Figure 6. Searching and Translation to English](image)

• Display: all words in dictionary can be displayed, as well as display words in categories according to familiar words, Iraqi folk proverbs, Iraqi cuisine, professions, tools, etc.

After implement analyzing, the Iraqi dictionary contains 522 words applicable to extended. see the small dictionary in figure 7.

![Figure 7. Part of Iraqi dictionary](image)

• Add: to enlargement our dictionary, if the Iraqi word not exist in proposed dictionary, add it. This step contributed to expansion of the Iraqi dictionary to contain more than 800 words.

• Synonym: Display Iraqi words synonym depends on our dictionary words. The proposed algorithm depends on the similarity in semantic between English words of dictionary, then extract the corresponding Iraqi words. The threshold value was respected to get more confidence that the similarity factor is more than half, therefore set to 0.6. For example, these Iraqi words are synonyms "ھﻮاﯾﺔ، ﻛﻮﻣﺔ" that mean "a lot of" in English. This process was explained in algorithm 2.

```
Algorithm(2): Display Iraqi words Synonyms
Input: Iraqi word
Output: set of synonyms
Step1: Get English word corresponding to input Iraqi word.
Step2: Find similarity in semantic of English Word from our dictionary.
Step3: Get Iraqi words corresponding to English word which similarity >= similarity factor 0.6.
Step4: End.
```

• Word Cloud: an electronic image that shows words used in a particular piece of text which allows you to identify the most frequently used words. The words are different sizes according to how often they are used in the text. Word cloud can be displayed after pre-processing of selected text file. Figure 8 shows the word cloud of Iraqi film "الجابي".
4. Conclusion and future work

The Iraqi dialect contains many different and beautiful words, especially the containment of sound ( 접, چ, گ ). Of course, any work faces challenges by theoretical and practical. In this paper, we face the hardest step is creation of Iraqi dialect corpus. As well as, synthesis of synonyms such as "ودي", "ارس", "اندز" that meaning in MSA "بعث" and meaning in English language "send", and convert words Iraqi dialect into MSA, and English language manually.

We managed to overcome all the difficulties and obstacle that we faced in this paper and obtained satisfactory results as we expected.

In Future, there are words in the Iraqi dialect that appear together such as "شکوماکو", "السافوية دافية "، "الصاية وصرماية" ....etc.), so we can develop this software to infer them using association rules. Another development to the proposed woke can be made to translate from MAS to more than Arabic dialects such as Egyptian, Tunisian, ...etc. Also implement more analysis tools for extracted more information about Iraqi dialect.

In fact, we must increase the Iraqi corpus to treat with Iraqi dialects widely. We can compute the degree of similarity between Iraqi dialect and MSA to show Is the Iraqi dialect is close to MSA?

Acknowledgement

I cannot express enough thanks to my department (Computer Science) in college of Science at University of Kerbala for their continued support and encouragement. I offer my sincere appreciation for the learning opportunities provided by my department.

My completion of this project could not have been accomplished without the support of my colleagues.

References

[1] Nizar Y. Habash (2010), ‘Introduction to Arabic natural language processing (Synthesis lectures on human language technologies)’ Morgan & Claypool Publishers; 1st edition, 186 pages.

[2] Maha M. Hassan et al.(2018), ‘Rule Based Method of Name Entity Recognition for Matching Allah’s Finest Names in Holy Quran’, Journal of Engineering and Applied Sciences 13 (10): 3618-3623.

[3] Dhamyaa A. AL-Nasrawi et al.(2015), ‘From Arabic Alphabets to Two Dimension Shapes in Kufic Calligraphy Style Using Grid Board Catalog’, Communications in Applied Sciences, 3(2): 42-59

[4] Owen Rambow et al.(2006) , ‘Parsing Arabic Dialects’, final report version 1, done at a Johns Hopkins http://www.mit.edu/~rplevy/papers/rambow-etal-2006-techreport-final.pdf

[5] Wael Salloum and Nizar Habash(2013), ‘Dialectal Arabic to English Machine Translation: Pivoting through Modern Standard Arabic’, Proceedings of NAACL-HLT, Atlanta, Georgia, pp. 348–358.
[6] Khaled Shaalan et al. (2007), ‘Transferring Egyptian Colloquial Dialect into Modern Standard Arabic’, International Conference on Recent Advances in Natural Language Processing.

[7] Kareem Darwish and et al. (2018), ‘Diacritization of Moroccan and Tunisian Arabic Dialects: A CRF Approach’, Conference: the 11th edition of the Language Resources and Evaluation Conference, at Miyazaki, Japan.

[8] Athrein Abu Kwaik and et al. (2018), ‘A Lexical Distance Study of Arabic Dialects’, The 4th International Conference on Arabic Computational Linguistics (ACLing2018), Dubai, United Arab Emirates.

[9] Bezoui Mouaz and et al. (2019), ‘Speech Recognition of Moroccan Dialect Using Hidden Markov Models’, International Symposium on Machine Learning and Big Data Analytics for Cybersecurity and Privacy (MLBDACP), Leuven, Belgium.

[10] Niladri Sekhar Dash (2010), ‘Corpus Linguistics: A General Introduction’, CIIL, Mysore, 25th August 2010.