Automatic Persian-Arabic Phonetic mapping

Zaid Rajih Mohammed[1], Ahmed H. Aliwy[2]

[1]Faculty of Medical Sciences, Jabir ibn Hayyan University for Medical and Pharmaceutical Sciences, Najaf, Iraq, Email: zaid.rajah@jmu.edu.iq
[2]Department of Computer Science, Faculty of Computer Science and Mathematics, University of Kufa, Najaf, Iraq, Email: ahmedh.almajidy@uokufa.edu.iq

Received: 17.07.2024 Revised: 18.08.2024 Accepted: 20.09.2024

ABSTRACT
Language is the most essential means of human communication and comes in several forms, the most significant being sound. Studying the phonetic relationship between different languages helps in building models that process and understand these interlingual connections. Phonetic mapping refers to converting the phonetic of words from one language to another. The main objective of this research is to develop a framework for phonetic mapping from Persian to Arabic. We created a bilingual Persian-Arabic phonetic dataset and applied a statistical model to identify shared phonetic elements. Additionally, we used a Hidden Markov Model (HMM) and developed a rule-based approach to refine the dataset and derive Arabic phonetic representations from Persian. The proposed model was evaluated based on accuracy, phonetic error rate (PER), and word error rate (WER). Using the rule-based approach, the accuracy of phonetic mapping from Persian to Arabic reached 85.6%.

Keywords: HMM, PER, WER, process, connections.

1. INTRODUCTION

Persian and Arabic has same scripts with little changes where Arabic is written in Arabic scripts while Persian is written in the Perso-Arabic script. In spite of they are using the similar scripts, there are huge differences in pronunciation and in hence in phonetic[1]. Phonetic mapping between Persian and Arabic is a difficult task result from the distinct speech sounds found in these two languages, the differences in phonetic structures, such as consonants, vowels, and vowels-consonant sequence.Phonetic mapping serves as a valuable tool to establish a comprehensive link between the phonological systems of Persian and Arabic. By navigating the complexities of phonetic variation, phonetic mapping contributes to effective language learning, linguistic analysis, and the development of speech recognition systems that cater to the nuances of both Persian and Arabic, fostering improved cross-cultural communication and understanding[2].

In scientific terms, the process of Phonetic mapping between different languages, in this context, is a mathematical relationship between two sets: the first set representing sounds in one language (L1), and the second set representing sounds in another language (L2). A sound segment from L1 is mapped to a comparable sound segment in L2, representing this connection as a mathematical function.It has various applications within Natural Language Processing (NLP), including Neural Machine Translation (NMT), Named Entity (NE) matching, and speech-to-speech conversion, especially for names.

The complexity of the mapping process becomes evident, particularly when dealing with languages that are different in terms of complexity, richness, and inflection, such as Arabic and Persian. Additionally, the problem becomes more challenging when the transformation is many-to-many, meaning that several phonemes in L1 may correspond to multiple phonemes in L2. The fundamental problem lies in the limited accuracy and variability of the pronunciation of English names in Arabic. For example, the Persian name "گهرچهر" in English “Geharchahar “ can be translated by some as "قهارجهر," while others may pronounce it as "غهارجهر" or "جهار جهر.". This make most of the current known systems are failing in translation of names (phonetically) between any two languages[3].

In our research, an integrated system is proposed that converts Persian phonetic representations into Arabic phonetic representations. This is accomplished by utilizing Hidden Markov Models (HMM) with an improved preprocessing technique[4], which instead of determining the phonemes, identifies the matching sounds in both languages to determine the phonetic segment[5]. The proposed system is then tested using two metrics, Phoneme Error Rate (PER) and Word Error Rate (WER), to evaluate its performance.

The main contributions of this proposed model can be summarized by: (i) Constructing Bilingual phonetic dataset with the format (Persian word + Persian phonetic + the equivalent Arabic word +Arabic phonetic) where Persian word or Arabic word are names or loanwords. (ii) Using a new method for phonetic segmentation depending on common phonetics between Persian and Arabic languages with reducing the segment size. (iii) Employing HMM with the special formatting of the used dataset. (iv) using a proposed method for selecting the best phonetic segmentation based on n-gram model.
The remainder of the paper is structured as follows: Section 2 discusses related works, and Section 3 describes the suggested model in detail. The experiment of our model and the discussion of the implementation and results are presented in Section 4, while the conclusion is provided in Section 5.

2. RELATED WORKS

Researchers have developed numerous methods to identify phonetic relationships between different languages. In this section, we present a concise overview of key contributions to the existing methodologies for phonetic mapping. Zouhar[6]developed a number of novel methods for building word embeddings with phonetic insights by using articulatory properties. Additionally, he suggested multiple methods for evaluating the inherent qualities of phonetic word embeddings, including elements such as word recall and association with phoneme similarity. Libovicky[7]introduced the idea of a neurological model for string transduction that is based on the string edit distance, known as neural string edit distance. Their empirical results including cognate recognition, and grapheme-to-phoneme conversion show that the proposed model can reach similar performance levels as standard black-box models when given contextualized input representations. Nehar[8]introduced two novel methods for the pairwise comparison of Arabic personal names. The initial approach relies on string alignment and phonetic transcription to identify similarities between Arabic personal names. The second method uses machine learning techniques to create a useful model for this purpose. The performance of the newly proposed models demonstrates favorable comparisons with top-performing similarity metrics. Cheng[9]integrated phonetic data into neural networks through two distinct methods: generating additional data via forward and back-translation with a focus on phonetics, and pre-training models on a phonetic task prior to transliteration learning. Yousef[10]presented and implemented a system for the cross-language mapping of names between Arabic and English. A recent iteration of Arabic Soundex has been employed to expedite the creation of a base dictionary from pre-existing information. Alshuwaier[11]studied offered a transliteration methodology that uses pronunciation and phonetic principles to convert English to Arabic. In order to enable the automated transcription of English names in programming systems, they devised algorithms based on phonetic criteria. Rao[12]used rules-based techniques for Phonetic matching between Hindi and Marathi or in cross-language as part of information retrieval system. he used Private dataset for Hindi and Marathi languages and applied for IR system. However, these approaches depend heavily on training data in the language of interest and their specific phonemic transcriptions. Our approach, on the other hand, abstracts away the dependency to Phonetic Mapping by identify the common phonetic and HMM.

3. The proposed model

In the realm of phonetic and language processing, the integration of Hidden Markov Models (HMMs) with phonetic mapping represents a good selection that holds tremendous potential for enhancing our understanding of spoken communication. By combining the probabilistic modeling capabilities of HMMs with the intricacies of phonetic mapping, this framework aims to unravel the complex relationships between the phonetic alphabet of one language and the phonetic alphabet ofanotherlanguage.
The synergy between Hidden Markov Models and phonetic mapping holds the key to advancing the state-of-the-art in phonetic processing, paving the way for more refined and context-aware systems in the field. The proposed system has four phases: (i) Data Collection, (ii) Segmented Phonetic Vocabulary, (iii)Reduce the length of Segments. (iv)Hidden Markov Model, (v) Optimal Segmentation. Figure 1 shows the proposed system for Persian-Arabic phonetic mapping.

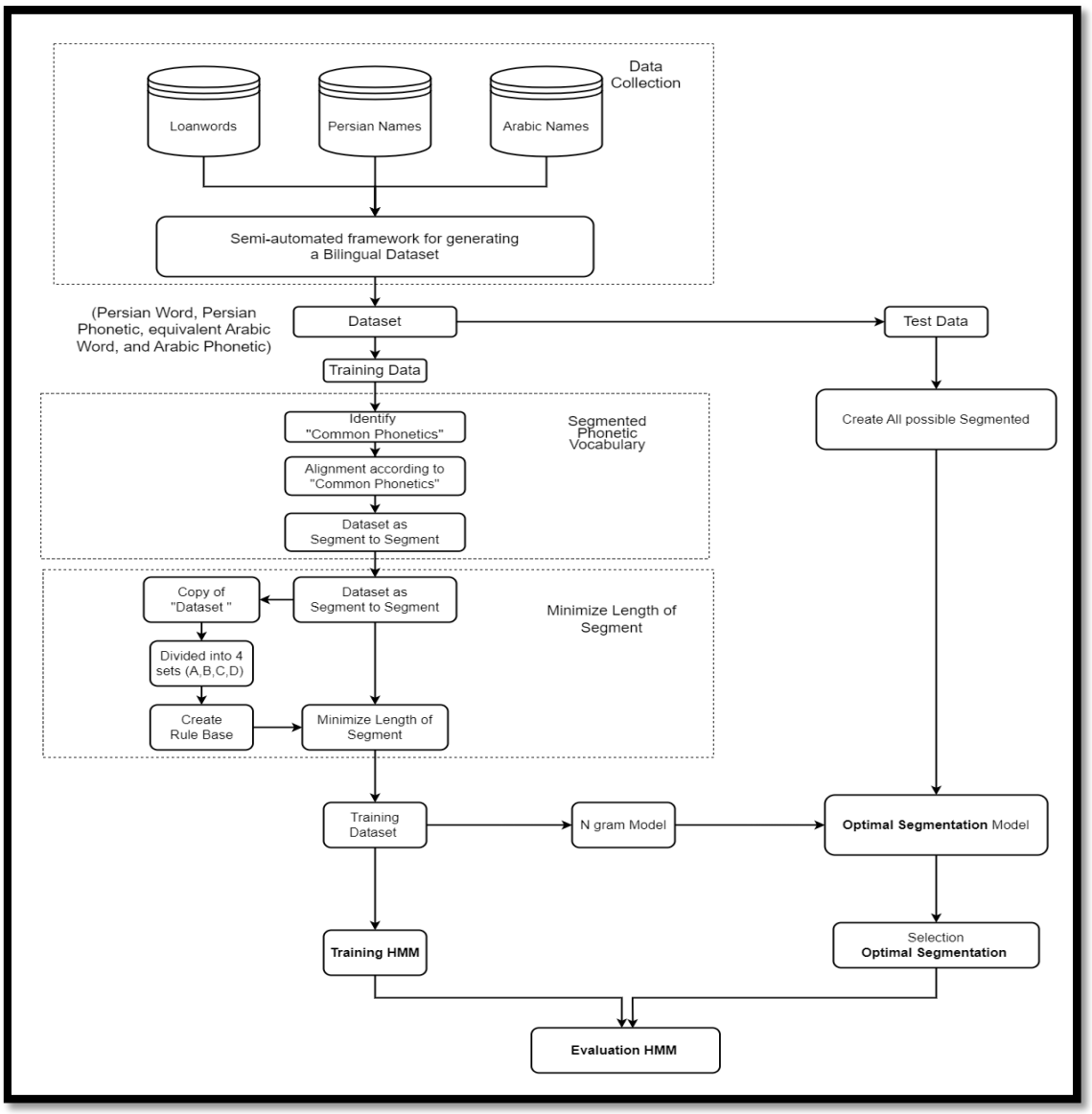

Figure 1. Main Diagram of Proposed Model

3.1. Data Collection

A semi-automated system was introducedto facilitate the creation of a Bilingual phonetic Persian-Arabic corpus to be used in the tasks of phonetic and semantic similarities. It is part of multilingual phonetic corpus as was started in previous work. Same methodology that used by Rajeh and Aliwy [13]is used in this work for data collection of Persian-Arabic phonetic dataset where it consist of four phases; (i) data gathering, (ii) preprocessing and translation, (iii) extraction of IPA representation, and (iv) manual correction. Firstly,the names in the Persian language were collected, as well as the Arabic names that can be used in the Persian language with the same Arabic pronunciation, and applying the remained steps to get the final form (Persian Word, Persian Phonetic, equivalent Arabic Word, and Arabic Phonetic). Figure 1 show an example for one name with its phonetic representation in the two languages.

Persian Word	Persian Phonetic	Arabic Phonetic	Arabic Word
سیرنگ	/sirneg/	/si:rnɣ/	سیرنغ

Figure 2: the Persian word "سیرنگ" with its phonetic representation in the two languages.

The collected dataset, which has 5850 items, used the International Phonetic Alphabet (IPA) standard for phonetic representation.

The process of preparing data involves several smaller processes, such removing forgen characters, removing gaps between letters and standardizing the accepted phonetic representation. Also we should

see that number of Persian phonetic and Arabic phonetic are 23 and 32 respectively, while the number of litters are 32 and 28 for Persian and Arabic languages respectively. The common phonemes are playing the core role for our proposed system.

3.2. Segmented Phonetic Vocabulary

In this stage, we try to produce segment to segment phonetic mapping for the dataset. This is done using the common phonetic between Persian and Arabic languages. These common phonetics is used as basic map segments and the others segmentsare taken as candidate segments. Common phonetic will be one to one segment while the candidate segments will be many-to-many, one-to-many, or many-to-one where many means more than one phonetic. These will be reduced to one or two in the next stage, i.e. many will be one or two only.There is implicit step which is known as alignment, it is responsibility make number of segments are equaled between the two phonetic representation (Persain and Arabic) depending on the common phonetic.Therefore,really there three steps in this stage; (i) identifying the common phonetic, (ii) the alignment and (iii) producing equal segments for the two equivalent representation.

We call the sounds x, in language L1, and y, in language L2, are common if x and y have the same pronunciation. We should see that if IPA is used as unified phonetic representation, then x and y have the same symbol. For example, the phonetics d, b, l, n, m, s, z, ʤ, k, f, r, ʃ, and x are examples for common phonetic between Persian and Arabic languages.

Figure 3 show the input and the output of this stage.

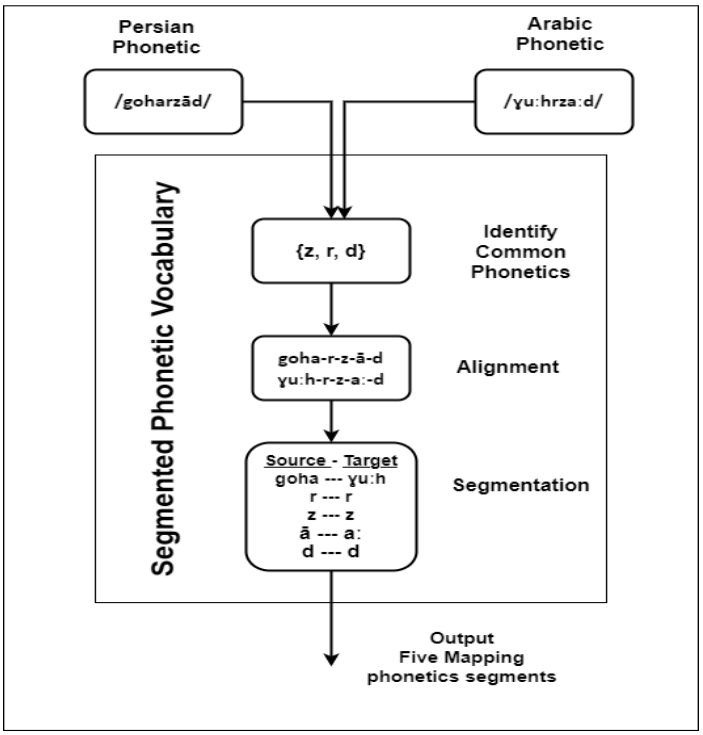

Figure 3. Show the steps of Segmentation

According to figure 3, the goha segment in the Persian phonetic representation appears as four phonetic symbols, matching the three phonetic symbols in the Arabic phonetic representation ɣu:h. Additionally, a common phonetic could occur in Persian phonetic representation of specific word; but it doesn't occur in Arabic phonetic representation of this word, then it is disregarded, regarded as a non-common phonetic, combined with another sound, or forced with the sign #.

3.3. Reduce the length of Segments

If a segment has many sounds, it cause a problem because there are errors in mapping process therefore this segment should minimized or segmented into sub-segments and hence decreasing the length of phonetic segments. As a result, when the size of segment will be reduced then the size of vocabulary of all segments will be reduced and hence the calculation of probabilities will be simplified.

All the segments are grouped into four group according to number of the phonetics in each segment as following:

Group A: contain all the segments that have one phonetic in Persian and Arabic (one-to-one segment)

Group B: contain all segment in the form (one-to-two, two-to-two and two-to-one segments).
Group C: contain all segment in the form (one-to-three, two-to-three, three-to-three, three-to-two, three-to-one segments).
Group D: contain all segments that not exist in previous groups will be in this group such as (one-to-four, one-to-five, … six-to-one, and so on).
Groups A and B are not need to a further processing and they are normal cases but Groups C and D should be repartitioned to be in the form of groups A or B. In almost all cases, the new sub-segments will be existed in groups A or B. This is done using transformation-based technique (rule-based where the rules automatically generated from the dataset). The idea is that if a sub-segment (between source and target) in a group X exist in the previous smallest group then it will be partitioned into two new segments for the source and target. This will be repeated until no any further possible segmentation. It done using a created rules from these Groups as shown in algorithm 1. Then these rules will be applied for any large segments of phonetic representation as shown in algorithm 2.
The algorithm 1 is applied for more one time for D and C groups, D and B groups, C and B groups. The algorithm's result are 3250 Rules, as shown in the example in figure 4:

Algorithm (1) Create Rule Base

```
Algorithm CreateRuleBase
Input: set_1, set_2,// each line has source and target segments
Output: RuleBase
Begin
  N = size of set_1
  M = size of set_2
RuleBase = empty set
  For i = 1 to N do
    For j = 1 to M do
      If (source_2[j] in source_1[i]) and (target_2[j] in target_1[i]) then
        Add the following rule to RuleBase:
Rule.Source=source_1[i]
Rule.Target =target _1[i]
        Rule.Source1 =source_2[j]
Rule.Source2=source_1[i]-source_2[j]
        Rule.Target1 =target_2[j]
        Rule.Target2 =target _1[i]-target _2[j]
Add Rule to RuleBase
Return RuleBase
End
```

Algorithm (2) Apply Rule Base

```
Algorithm Apply RuleBase
Input: RuleBase, Source_segment, Target_segment
Output: Source_segments,Target _segments
Begin
  For each rule in RuleBase:
    If rule is applicable on (Source_segment and Target_segment)
        Source1 = rule.source1
        Source2 =rule.source2
        Target1 = rule.target1
        Target2 = rule.target2
        Add to Source _segments (Source1, Source2)
        Add to Target _segments (Target1,Target2)
Repeat step1 to step3 for each new source and target segment
Return (Source _segments,Target _segments)
End
```

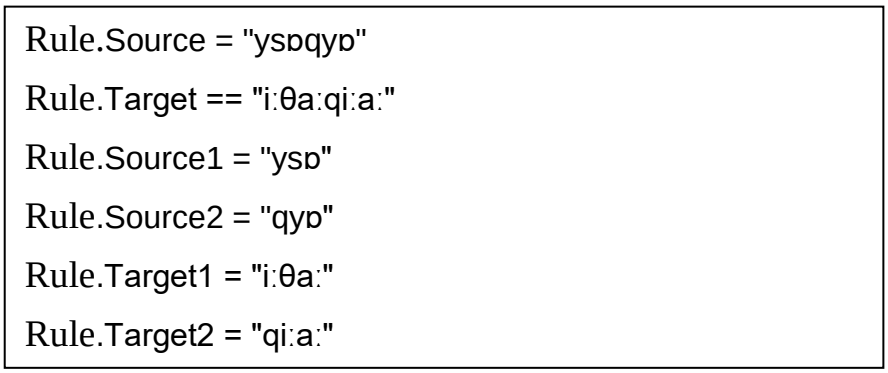

Figure 4. Show Example of Rule Base

In this and previous sections, a segment-to-segment bilingual phonetic representation dataset were collected and preprocessed. Where P_{PR} is Persian phonetic representation, A_{PR} is Arabic phonetic representation, PS_i and AS_i segment number i in Persian and Arabic respectively.

$$P_{PR} = [PS_1 \quad PS_2 \ldots . \quad PS_n] \leftrightarrow A_{PR} = [AS_1 \quad AS_2 \ldots . \quad AS_n]$$

This dataset is ready to be used in any statistical, probabilistic or machine learning algorithm. We select HMM model because it good choice for sequence prediction.

3.4. HMM(Hidden Markov Model)

Hidden Markov Models (HMMs) are powerful statistical models widely employed in various fields, ranging from speech recognition and natural language processing to bioinformatics and finance. Fundamentally rooted in probability theory, HMMs are designed to capture and model sequences of observations, where each observation represents an outcome associated with an underlying hidden state. The term "hidden" in HMMs stems from the fact that these underlying states are not directly observable; instead, they generate the observed data through a stochastic process. The strength of HMMs lies in their ability to model dynamic systems where the current state depends only on the previous state, making them particularly adept at representing temporal dependencies in sequential data. By leveraging the principles of Markov chains and probability distributions, HMMs provide a versatile framework for understanding, predicting, and analyzing sequential patterns in diverse domains. This introductory paragraph merely scratches the surface of the rich and intricate workings of Hidden Markov Models, which play a pivotal role in unraveling complex relationships within sequential data.in our methodology, the 235 Arabic phonetics are represented by Hidden States. The Arabic phonetic Si's likelihood of occurring if the Persianphonetic $O_j$does is known as the "emission probability." The likelihood that the Arabic phonetic Si will come after the Arabic phonetic $S_j$is known as the transition probability.

3.5. Optimal Segmentation

In many cases the source phonetics representation has more than one segmentation especially if it is not existed in the vocabulary. Therefore, we should add a process to select the optimal segmentation before applying the HMM. This stage's implementation is carried out during model testing rather than the HMM model creation and training phase since system testing necessitates segmenting the phonetics sequence to be evaluated. As seen in the sample in figure 5, the phonetics representation is divided into many phonetic segments.

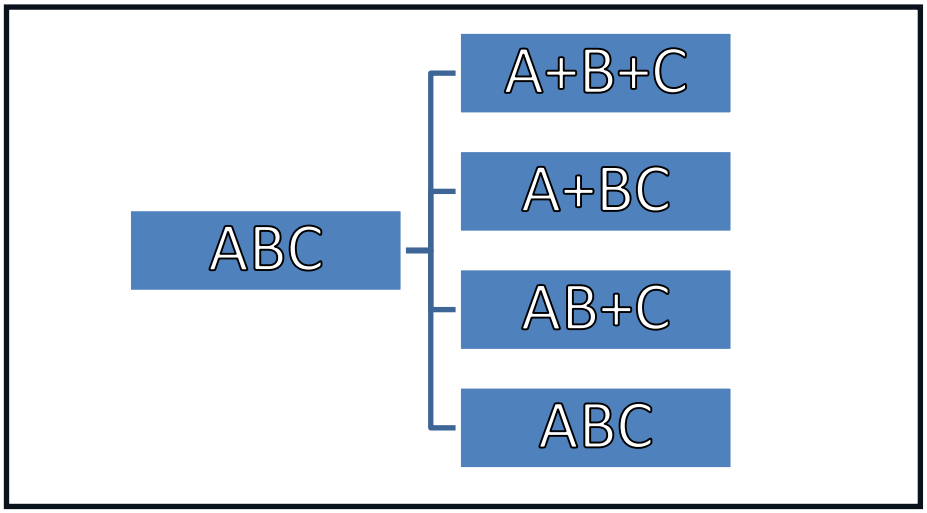

Figure 5. Show the possibility segmentation of "ABC" word

The optimal segmenting procedure is found using the following formulas.

$\omega_{Sj} = p(S_{j+1}|S_j)$ ……..(1)

where $(p(S_{j+1}|S_j))$ indicates the likelihood that the sound segment (S_{j+1}) would emerge after the sound segment (S_j), and (ω_{Sj}) indicates the weight of the sound segment (S_j).

$W_{Pi} = \prod_{j=1}^{m} \omega_{Sj}$ …….. (2)

Where (W_{Pi}) is the weight of the branch that represents one of the proposed solutions

$Ps = \max(W_{P1}, W_{P2}, \dots, W_{Pn})$ …….. (3)

Since the sequence that was selected is determined by weight, it is regarded as the best option. This sequence is represented by (Ps). The best phonetic string segmentation is selected at this point in order to be employed in the HMM to determine the phonetic representation of the other language.

4. Implementation and Results

The proposed model was tested using Python 3.11 with different libraries. The collected dataset contains5,232 words, gathered from Persian and Arabic names, as well as loanwords between the two languages. Thisdataset was divided into 80% training data and 20% testing data. In the "Segmented Phonetic Vocabulary" stage, the vocabulary was converted from phonetics into phonetic segments by identifying common phonetics such as{d, b, l, n, m, s, z, ʤ, k, f, r, ʃ, x}. After completing the "Alignment" and "Segmentation" processes, two important types of data were obtained. The first type is phonetic representation of words matching between Persian and Arabic thatsegmented based on the shared phonetics between the two languages. The second is aset of matching phonetic segments that identified which consist of 27,097 segments. As a result, the size of the Persian vocabulary was expanded from 26 phonetic units to 816 phonetic segments, and the Arabic vocabulary from 32 phonetic units to 550 segments.

Table 1. Show the effect of Rule Base on the size of sets

No	Sets	Before Rule Base	After Rule Base
1	A	66	79
2	B	381	516
3	C	478	10
4	D	313	0

In the "Reduce the length of Segments" stage, a rule-base was created and applied.The Rule-Base algorithm produced 895 rules, and after applying of it, the number of phonetic segments increased to 29,182. Additionally, no phonetic segment was longer than 3 phonetic symbols. Consequently, the size of the Persian vocabulary was reduced from 816 to 233 phonetic segments, and the Arabic vocabulary from 550 to 182 segments. The updated data were utilized in the Hidden Markov Model (HMM)usingViterbialgorithm on the test data, which contained 1,046 cognates between Persian and Arabic. The results were as shown in table 2.

Table 2. Show the Result of using HMM

	Number of sounds in Original representation	Number of sounds after applying HMM	MED	Accuracy	PER	WER
Without Rule Base	5.408	4.66	1.15	0.78578	0.21421	0.395
Rule Base	5.408	5.124	0.787	0.85656	0.14343	0.484

Table 2 shows effect of using rule-base on number of sounds with applying HMM. It is clear that using rule-based is better because after applying HMM it give number of sounds close to the original numbers. Also, four evaluation metrics were used which are Minimum Edit Distance (MED),Accuracy, Phoneme Error Rate (PER), and Word Error Rate (WER).MED is the minimum number of operations needed to change the candidate phonetic representation “A” into the correct representation for the selected word. Accuracy is the ratio of correct phonemes to the number of actual phonemes, while PER measures the ratio of insertions, deletions, and substitutions relative to the number of actual phonemes. WER represents the proportion of word errors relative to the total number of words. These metrics help assess the system's accuracy in converting phonetic representations between languages. Figure 6 shows the input and output of HMM for sample from test data.

no	Farsi Word	Arab Word	Farsi Phontic	Arabic Phontic	Result HMM
180	بُرزویه	بُرزویه	['b', 'e', 'r', 'z', 'u', 'y', 'e']	['b', 'r', 'z', 'u:', 'i:', 'h']	['b', 'r', 'z', 'a:', 'h']
181	بهمرد	بهمرد	['b', 'a', 'h', 'm', 'o', 'r', 'd']	['b', 'ɪ', 'h', 'm', 'r', 'd']	['b', 'h', 'm', 'r', 'd']
182	بابوي	بابوي	['b', 'ā', 'b', 'a', 'v', 'i']	['b', 'a:', 'b', 'u:', 'i:']	['b', 'a:', 'b', 'u:', 'i:']
183	باتيس	باتيس	['b', 'ā', 't', 'i', 's']	['b', 'a:', 't', 'i:', 's']	['b', 'a:', 't', 'i:', 's']
184	بادر	بادر	['b', 'ā', 'd', 'e', 'r']	['b', 'a:', 'd', 'r']	['b', 'a:', 'd', 'r']
185	باران	باران	['b', 'ā', 'r', 'ā', 'n']	['b', 'a:', 'r', 'a:', 'n']	['b', 'a:', 'r', 'a:', 'n']
186	بارلي	بارلي	['b', 'ā', 'r', 'l', 'i']	['b', 'a:', 'r', 'l', 'i:']	['b', 'a:', 'r', 'l', 'i:']
187	باريزان	باريزان	['b', 'ā', 'r', 'i', 'z', 'ā', 'n']	['b', 'a:', 'r', 'i:', 'z', 'a:', 'n']	['b', 'a:', 'r', 'i:', 'z', 'a:', 'n']
188	بازنه	بازنه	['b', 'ā', 'z', 'a', 'n', 'e']	['b', 'a:', 'z', 'n', 'h']	['b', 'a:', 'z', 'n']
189	بازه	بازه	['b', 'ā', 'ž', 'e']	['b', 'a:', 'z', 'h']	['b', 'a:', 'n']
190	باسكار	باسكار	['b', 'ā', 's', 'k', 'ā', 'r']	['b', 'a:', 's', 'k', 'a:', 'r']	['b', 'a:', 's', 'k', 'a:', 'r']
191	باشور	باشور	['b', 'ā', 'ʃ', 'u', 'r']	['b', 'a:', 'ʃ', 'u:', 'r']	['b', 'a:', 'ʃ', 'u:', 'r']
192	باقر	باقر	['b', 'ā', 'q', 'e', 'r']	['b', 'a:', 'q', 'r']	['b', 'r', 'r']
193	بالبان	بالبان	['b', 'ā', 'l', 'e', 'b', 'ā', 'n']	['b', 'a:', 'l', 'b', 'a:', 'n']	['b', 'a:', 'l', 'b', 'a:', 'n']
194	بالين	بالين	['b', 'ā', 'l', 'i', 'n']	['b', 'a:', 'l', 'i:', 'n']	['b', 'a:', 'l', 'i:', 'n']
195	بامين	بامين	['b', 'ā', 'm', 'i', 'n']	['b', 'a:', 'm', 'i:', 'n']	['b', 'a:', 'm', 'i:', 'n']
196	بانيبال	بانيبال	['b', 'ā', 'n', 'y', 'ā', 'p', 'ā', 'l']	['b', 'a:', 'n', 'i:', 'b', 'a:', 'l']	['b', 'a:', 'n', 'i:', 'k', 'l']
197	باوين	باوين	['b', 'ā', 'v', 'i', 'n']	['b', 'a:', 'u:', 'i:', 'n']	['b', 'a:', 'u:', 'i:', 'n']
198	بابا	بابا	['b', 'ā', 'y', 'ā']	['b', 'a:', 'i:', 'a:']	['b', 'a:', 'i:', 'a:']
199	بايغان	بايغان	['b', 'ā', 'y', 'g', 'ā', 'n']	['b', 'a:', 'i:', 'ɣ', 'a:', 'n']	['b', 'a:', 'i:', 'ɣ', 'a:', 'n']
200	بتول	بتول	['b', 'a', 't', 'a', 'v', 'a', 'l']	['b', 't', 'u:', 'l']	['b', 't', 'u:', 'l']
201	بخت آفرين	بخت آفرين	['b', 'a', 'x', 't', 'ā', 'f', 'a', 'r', 'i', 'n']	['b', 'x', 't', 'ʔ', 'a:', 'f', 'r', 'i:', 'n']	['b', 'x', 't', 'a:', 'f', 'r', 'i:', 'n']
202	بخشان	بخشان	['b', 'a', 'x', 'ʃ', 'ā', 'n']	['b', 'x', 'ʃ', 'a:', 'n']	['b', 'x', 'ʃ', 'a:', 'n']
203	بدن گل	بدن غل	['b', 'a', 'd', 'a', 'n', 'g', 'o', 'l']	['b', 'd', 'n', 'ɣ', 'l']	['b', 'd', 'n', 'ɣ', 'l']

Figure 6. Show the samples of results

5. CONCLUSIONS

We introduced a Persian-Arabic phonetic mapping model, which is based on Hidden Markov Models (HMM). Our approach of phonetic mapping critically depends on identifying common phonetic elements between Persian and Arabic words that are phonetically equivalent. Persian and Arabic are two distinct languages from different language families but use the same script. Some Persian sounds do not exist in Arabic, leading to different phonetic representations of these sounds in Arabic. Many loanwords have been exchanged between the languages due to geographical proximity and religious influence.

Our experimental results on phonetic similarity and Persian-to-Arabic transliteration show that the proposed model performs comparably to other models. We also concluded that phonetic maps offer a higher level of specialization, as they focus primarily on the phonetic representations of words. However, these maps have proven useful in various practical applications, such as cognate detection and multilingual named entity recognition. We hope that our approach will inspire further research on this type of interpretable model and that our framework will be useful for future work in this area.

REFERENCES

[1] Mojgan S. Morphosyntactic Corpora and Tools for Persian: Ph. D. thesis, Uppsala University; 2012.
[2] Al-Sharkawi M. History and development of the Arabic language: Routledge; 2016.
[3] Freeman A, Condon S, Ackerman C, editors. Cross linguistic name matching in English and Arabic. Proceedings of the Human Language Technology Conference of the NAACL, Main Conference; 2006.
[4] Kantor A, Hasegawa-Johnson M. Hmm-based pronunciation dictionary generation. New Tools and Methods for Very Large Scale Phonetics Research, University of Pennsylvania. 2011.
[5] Jakobi DN. Grapheme-to-Phoneme Mapping in Text [Master's thesis]: University of Zurich; 2022.
[6] Zouhar Ve, Chang K, Cui C, Carlson N, Robinson N, Sachan M, et al. Pwesuite: Phonetic word embeddings and tasks they facilitate. arXiv preprint arXiv:230402541. 2023.
[7] Libovick`y J, Fraser A. Neural string edit distance. arXiv preprint arXiv:210408388. 2022.
[8] Nehar A, Bellaouar S, Ziadi D, Omar KM. Arabic Personal Name Matching: Names Written using Latin Alphabet. Journal of Computer Science. 2021;17:776-788.
[9] Cheng S, Ding Z, Yan S. English-to-chinese transliteration with phonetic back-transliteration. arXiv preprint arXiv:211210321. 2021.
[10] Yousef AH. Cross-language personal name mapping. arXiv preprint arXiv:14056293. 2014.
[11] Alshuwaier F, Areshey A, editors. Translating English names to Arabic using phonotactic rules. Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation; 2011.
[12] Rao SMCaS. Rule-Based Phonetic Matching Approach for Hindi and Marathi. International journal of research in social sciences. 2011;1:26-41.
[13] Mohammed ZR, Aliwy AH, editors. English-Arabic Phonetic Dataset construction. BIO Web of Conferences; 2024.