Sindh Univ. Res. Jour. (Sci. Ser.) Vol.49 (2) 363-366 (2017)

SINDH UNIVERSITY RESEARCH JOURNAL (SCIENCE SERIES)

Statistical Approaches to Instant Diacritics Restoration for Sindhi Accent Prediction

H. SHAIKH, J. A. MAHAR[++], M. H. MAHAR

Department of Computer Science, Shah Abdul Latif University, Khairpur Mir's, Sindh, Pakistan

Received 2nd May 2016 and Revised 28th March 2017

Abstract: Sindhi script highly abounds in the homographic words which lead the reader and machine to many complexities. Due to the possibility of several meanings of one homographic structure, the interpretation and understanding of the text becomes severely difficult. Before the interpretation, pronunciation varies which is the leading cause to the complexity. Diacritics help us remove such complexities and comprehend the text easily and accurately. Due to the time saving nature of the people of current era, they don't bother to write diacritics in routine writings. Apart from the difficulties in reading for human beings, the absence of diacritics creates difficulty for machine reading as well. The text prediction systems produced the basis for the instant diacritics restoration approach. This instant system of diacritics restoration is an entirely novel and unique work in the field of natural language processing. A framework of N-Grams and Maximum Entropy is proposed in this research work. The highest attention catching point of this system using unigram, bigram, trigram and quad-gram is 98.98% accuracy on the corpus of Sindhi language. The super edge of instant diacritics restoration is to be leading initiative to the highly advancing performance of other natural language and speech processing applications.

Keywords: Instant Diacritics Restoration, Sindhi Corpus, Text Prediction, N-Grams, Maximum Entropy

1. INTRODUCTION

Many homographic words are found in Sindhi language having dissimilar meanings. The function of diacritic signs is important for homographic words. These signs assist the script of particular language for being vivid and readable. Generally, Sindhi people do not use short vowels and other diacritic signs while typing text. Therefore, it is a great need of instant diacritics restoration system. The Instant diacritics restoration is basically useful for typing systems and similar with text prediction system. Instant diacritics restoration is mandatory component for various Sindhi natural language and speech processing applications (Shah, 2004).

Generally, diacritics restoration techniques are classified into three categories: rule-based, statistical and hybrid (Zayyan, 2016). Statistical approaches are selected in this research study. Various statistical and non-statistical techniques have been used by various researchers across the world for the task of diacritics restoration such as N-Grams (Harby, 2008), Neural Networks (Sultan, 2001), Maximum Entropy (Zitouni, 2008), Memory-Based Learning (Kubler, 2008), and Weighted Finite State (Nelken, 2005). The encouraging results have been achieved using N-Gram language model (Harby, 2008) (Mahar, 2014) at word level whereas Maximum entropy approach (Zitouni, 2008) yields acceptable results for diacritics restoration of Arabic script-based languages at letter level.

The core aim of this study is to design and developed software application that automatically assigns diacritic sign instantly to the every undiacritized character of input text words. For getting the valuable results, two intelligent techniques i.e. N-grams and Maximum Entropy are jointly used in such a way to solve the problem at letter and word level because both approaches are required for instant diacritics restoration. The N-grams are well known in the literature (Jurafsky, 2000) and have already been used for the same task (Shaikh, 2017). Most of the researchers used MaxEnt classifier for the classification of the diacritic signs with every character of the text. The maximum entropy approach mainly works at letter level using the probability distribution; hence, classifier is sufficient to appoint diacritics to each character. The process of instant diacritics restoration is based on the feature vectors of strings. The unigram, bigram, trigram and quad gram letters probabilities are used as the features and associated with the MaxEnt classifier. The instances are taken from our developed corpus and already stored into memory.

2. CORPUS PREPARATION

The corpus of language is prerequisite for software application using statistical approaches. Two type of text sets i.e. diacritized and undiacritized are required for successful experiments (Mahar, 2011). Hence, both types of corpora are designed and developed and

[++]Corresponding author Email: mahar.javed@gmail.com

already been used for instant diacritics restoration (Shaikh, 2017). For training and testing the developed system, the corpus of Sindhi language consists of 2, 65,257 words are built and given in (**Table 1**). The developed corpus is categorized into three parts: fully diacritized, partially diacritized and not diacritized.

Table-1 Words Information of Developed Sindhi Corpus

Type of Corpus	No. of Sentences	No. of Words
Fully Diacritized	**8326**	**49,462**
Partially Diacritized	**10190**	**93,188**
Not-Diacritized	**14869**	**1,22, 607**
Total	**33385**	**2, 65,257**

3. PROPOSED MODEL

The core aim of proposed mechanism is to correctly locate and insert vowel signs including other diacritics into all characters in the given text. The function of every diacritic sign works independently and modeling all diacritics is difficult task for all characters of text, therefore, arrangement is brought out individually (Zitouni, 2006). The problem is considered as series arrangement in which characters are formed, $c_1,c_2,....,c_L$, the diacritization task needs for every character of the alphabet, from the complete list of diacritics $d_1,d_2,....,d_L$ respectively.

The proposed maximum entropy based instant diacritics restoration model is based on 14 components. The proposed model for Sindhi instant diacritics restoration system is depicted in (Fig.1). All the diacritic signs used in Sindhi text are assigned to each letter for the training of data from the developed corpora and stored into feature space. The next step is to calculate probabilities of each diacritic sign with each letter trained from corpora. The classification of each letter and sign starts with the help of probability distribution. After classification, training of all classified letters and signs takes place. Now the actual process of diacritization begins here when the text is input. System selects each letter of input text, one by one, respectively, then calculation of N-grams on selected strings is performed; the features of Unigram, Bigram, Trigram and Quad gram do their job in the meanwhile because the context of selected letter is analyzed with N-grams based features.

Afterwards, various weights for every instance in the training data are calculated with a worthwhile judgment. When judgment is over, the integration of features using MaxEnt framework is processed and association of calculated weights with MaxEnt classifier is performed. Then the eventual phase of process for diacritization starts by estimating the optimal value of each letter. Based on the calculated optimal value, MaxEnt puts classification to each letter with its appropriate sign. The probability distribution is used for such association. By then, the justification phase comes to make the final decision. The system examines through forward and backward track from the selected letter up to 4 surrounding letters on each left and right side. Finally, a justified decision is made over the aggregated information given from classification and the undiacritized letter is replaced with a diacritized one.

4. IMPLEMENTATION AND RESULTS

During the literature review, it has been observed that MaxEnt is responsible for integrating the information and generate a decision for classification. The task of instant diacritics restoration is based on the prepared features of the strings that are already stored into memory in shape of vectors. Four different character vectors are generated from training data using Unigram (UG), Bigram (BG), Trigram (TG) and Quad gram (QG).

The developed system automatically selects one letter from the given text and takes the UG probability from the database, after that BG, TG and QG probabilities are selected for further process. Finally, an instance value is taken that is previously allocated to the selected letter. Then, feature values of every instance will be combined and computes the weight of each letters. The calculated weights are lastly linked with the MaxEnt and on the basis of probability distribution of calculated weights the diacritic symbol is coupled with selected character.

The system continuously matches the concatenated letters with the stored word. When the input word is completed then system finally shows the all possible word with correct diacritics and user is now able to select any word from the popup menu.

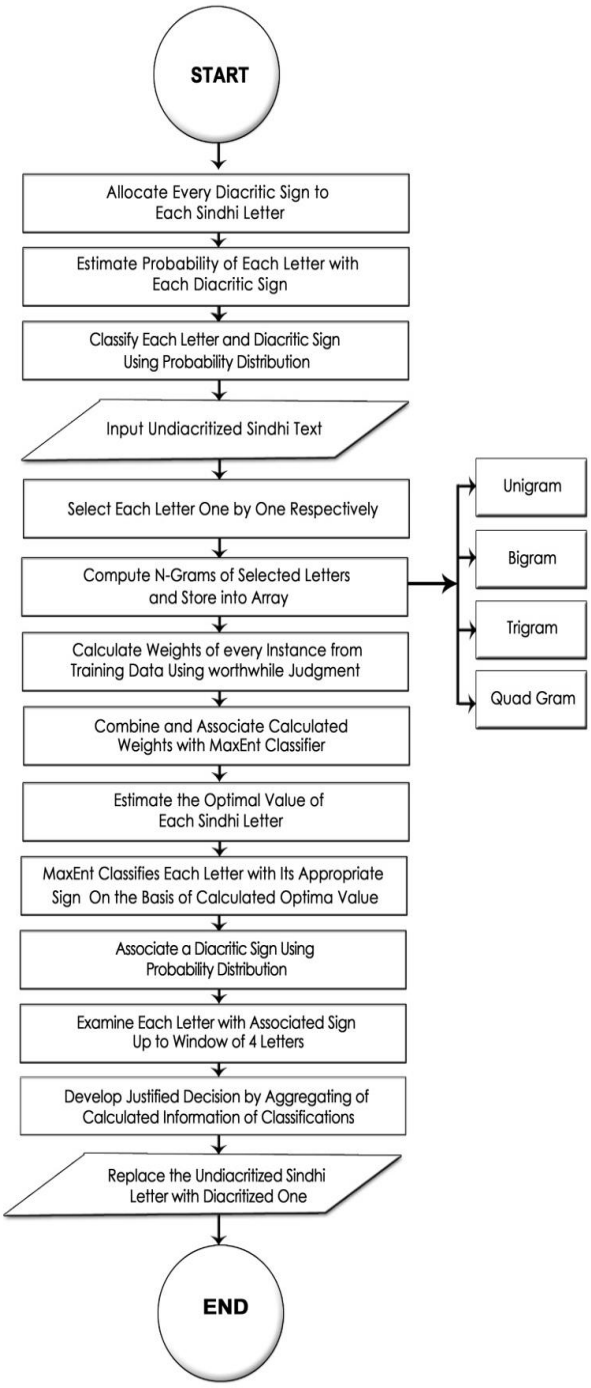

Fig. 1. Proposed Model for Sindhi Instant Diacritics Restoration

The performance of the developed application is evaluated on the developed corpus that is divided into two sets: Development Data (DS), and Test Data (TS). The main function of DS is to evaluate the selected features and alteration the training parameters. In all situations, the actual data will be same with the TS.

Moreover, the performance of the system is reported in terms of Precision (P) according to the conditions: (1) system access to unigram, (2) system access to unigram and bigram, (3) system access to unigram, bigram and trigram, (4) system access to unigram, bigram, trigram and quad gram features. The calculated results are presented in (**Table 2**) to (**Table 5**) using UG, BG, TR and QG respectively.

Table-2 Calculated Results using UG

Data Set	Diacritic Signs	No. of Characters	P
DS	Zabar	75,362	94.81
	Zair	63,851	93.37
	Pesho	52,454	93.98
	Jazam	16,027	93.78
TS	Zabar	71,489	92.05
	Zair	59,112	92.11
	Pesho	49,734	92.34
	Jazam	15,681	92.55

Table-3 Calculated Results using UG + BG

Data Set	Diacritic Signs	No. of Characters	P
DS	Zabar	75,362	95.27
	Zair	63,851	94.53
	Pesho	52,454	94.19
	Jazam	16,027	94.91
TS	Zabar	71,489	93.84
	Zair	59,112	93.77
	Pesho	49,734	93.13
	Jazam	15,681	93.75

Table-4 Calculated Results using UG + BG + TG

Data Set	Diacritic Signs	No. of Characters	P
DS	Zabar	75,362	97.41
	Zair	63,851	96.91
	Pesho	52,454	97.80
	Jazam	16,027	97.91
TS	Zabar	71,489	95.28
	Zair	59,112	95.33
	Pesho	49,734	96.42
	Jazam	15,681	95.17

Table-5 Calculated Results using UG + BG + TG + QG

Data Set	Diacritic Signs	No. of Characters	P
DS	Zabar	75,362	99.08
	Zair	63,851	98.26
	Pesho	52,454	99.05
	Jazam	16,027	99.32
TS	Zabar	71,489	97.39
	Zair	59,112	97.43
	Pesho	49,734	98.12
	Jazam	15,681	97.44

Four diacritic signs are selected for experiments, when UG feature is used then 93.99% accuracy is calculated when BG features are used with the UG then accuracy of 94.73% is achieved. The precision of 97.51% is calculated with the concatenation of UG, BG and TG. When we employed all the four N-gram features it reaches 98.98%. It is expected that better results can be achieved with backward and forward chain of all four selected features of N-grams. The calculated cumulative precision with different N-gram features is shown in **(Fig.2).**

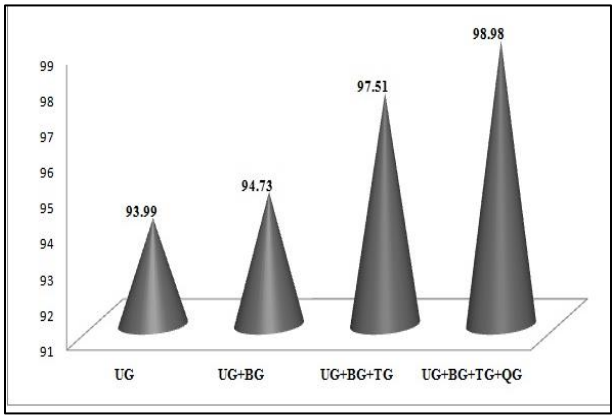

Fig.2 Cumulative Precisions Using Different N-Gram Features

5. CONCLUSION

The orthography of Sindhi language is complex due to the use of diacritic signs. The diacritics are important for correct understanding and pronunciation of words. Only four diacritic signs i.e. Zabar, Zair, Pesho and Jazam are selected for experiments. Instant diacritics restoration system is useful while typing text of Sindhi. Most intelligent approaches; N-grams and Maximum Entropy are used to achieve the acceptable results for Sindhi language. The proposed mechanism is experimented on our developed corpus and achieved 98.98% accuracy with the combination of unigram, bigram, trigram and quad gram features. After little variation, proposed approach can also be used for other Arabic script-based languages.

REFERENCES:

Harby, A. A., M. A. Shehawey, R. S. Barogy, (2008). A Statistical Approach for Quran Vowel Restoration. ICGST International Journal on Artificial Intelligence and Machine Learning, 8(3), 9-16.

Jurafsky, D., Martin, J. H. (2000). Speech and Language Processing: An Introduction to Natural Language Processing. Computational Linguistic and Speech Recognition, Prentice-Hall, 300-307.

Kubler, S., E. Mohamed, (2008). Memory-based vocalization of Arabic. In Proceedings of the LREC Workshop on HLT and NLP within the Arabic World, Morroco, 58-62.

Mahar, J. A., G. Q. Memon, (2011). Automatic Diacritics Restoration for Sindhi", Sindh University Research Journal (Science Series), 43(1), 43-50.

Mahar, S. A. (2014). Comparative Analysis of Vowel Restoration for Arabic Script Based Languages Using N-Gram Models. MS Thesis, Shah Abdul Latif University, Khairpur, Pakistan.

Nelken, R., S. M. Shieber, (2005). Arabic Diacritization using Weighted Finite-State Transducers. ACL Workshop on Computational Approaches to Semitic Languages, Association for Computational Linguistic, Michigan, 79-86.

Shah, A. A., A. W. Ansari, L. Das, (2004). Bi-Lingual Text to Speech Synthesis System for Urdu and Sindhi. National Conference on Emerging Technology, Karachi, Pakistan, 126-130.

Shaikh, H., J. A. Mahar, M. H. Mahar, (2017). Instant Diacritics Restoration System for Sindhi Accent Prediction Using N-Grams and Memory-Based Learning Approaches. International Journal of Advanced Computer Science and Applications, 8(4),149-157.

Sultan, H. (2001). Automatic Arabic Diacritization using Neural Network. Scientific Bulletin of Faculty of Engineering Ain-Shams University: Electrical Engineering, 36(4), 501-510.

Zayyan, A. A., M. Elmahdy, H. B. Husni, , J. M. Jaam (2016). Automatic Diacritics Restoration for Dialectal Arabic Text. International Journal of Computing & Information Science, 12(2), 159-165.

Zitouni, I., R. Sarikaya, (2008). Arabic Diacritic Restoration Based on Maximum Entropy Models. Computer Speech and Language, 23, 257-276.

Zitouni, I., S. Jeffrey, S. Ruhi, (2006). Maximum Entropy Based Restoration of Arabic Diacritics. 21st International Conference on Computational Linguistic and 44th Annual Meeting of the ACL, Sydney, Australia, 577-584.