# Sindhi Language Processing: A Survey

Wazir Ali Jamro
Department of Computer Science, Shah Abdul Latif University, Khairpur, Pakistan.
wazeer.ali_ms2014@salu.edu.pk

***Abstract*— In this era of information technology, natural language processing (NLP) has become volatile field because of digital reliance of today's communities. The growth of Internet usage bringing the communities, cultures and languages online. In this regard much of the work has been done of the European and east Asian languages, in the result these languages have reached mature level in terms of computational processing. Despite the great importance of NLP science, still most of the South Asian languages are under developing phase. Sindhi language is one of them, which stands among the most ancient languages in the world. The Sindhi language has a great influence on the large community in Sindh province of Pakistan and some states of India and other countries. But unfortunately, it is at infant level in terms of computational processing, because it has not received such attention of language engineering community, due to its complex morphological structure and scarcity of language resources. Therefore, this study has been carried out in order to summarize the existing work on Sindhi Language Processing (SLP) and to explore future research opportunities, also some potential research problems. This paper will be helpful for the researchers in order to find all the information regarding SLP at one place in a unique way.**

*Keywords— Sindhi Language Processing, SLP, Morphological Analysis, Word Segmentation, Parts-of-Speech Tagging, Diacritization, technique and approaches.*

## I. INTRODUCTION

The survival of human languages on the web has become important phenomena because of increased usage of internet based technologies. Natural languages are the best way for human to human communication but communicating with machines in such way that machine can understand and response as like human is a critical task. Therefore, the NLP has become a novel field of this decade and plenty of research has been done on the most of the world languages. Sindhi has a great influence on the large population, historical background and classical writing system. Although pretty work has been initiated on the SLP, in the result important applications including morphological analyser, Part-Of-Speech (POS) tagger, Diacritic restoration system, Spell Checkers, Optical Character Recognition (OCR), Text-To-Speech (TTS) synthesis system, Machine Translation (MT) systems and recently digital Thesaurus and stemmer have been investigated and various resources are developed by individual researchers for the digitization of Sindhi language. However, it demands more efforts in order to stand among the mature languages in terms of computation. Therefore, in this paper we have presented almost all the progress made in SLP along with future research opportunities and some potential research problems. It is worth to mention that no any survey paper is previously written on SLP and its related problems. The main contributions of this research are:-

1. To highlight the importance of Sindhi language and its computational processing.
2. To discuss morphological characteristics and structure of Sindhi language.
3. Summarize the work initiated on SLP
4. Classification of techniques used in SLP.
5. Insights about the areas of SLP applications.
6. To find out future research opportunities.

In nutshell this paper presents the detailed increased interest and progress made in computational processing of Sindhi. The paper is organized in a following sequence: section two presents the overview of Sindhi language, in the third section progress made SLP applications is presented and summary of developed SLP tools, applications and datasets is summarized in fourth section. The discussion and conclusion are presented in sixth section and finally future research opportunities are suggested in sixth section.

## II. BRIEF OVERVIEW OF SINDHI LANGUAGE

Sindhi is an example of Indo-Aryan language and popular due to some unique linguistic characteristics. It has very rich morphological structure. It stands among the most ancient language of the world, having vast historical literary and cultural background. Sindhi is an official language of both countries Pakistan and India. Only in Sindh province of Pakistan there are over 40 million native speakers of Sindhi. It is also spoken in some states of India like Ulhasnagar is largest Sindhi speaking region and other regions like Rajasthan, Gujarat and Maharashtra and some other states where Sindhi native speakers have migrated like Hong Kong, Singapore, Philippines, Canada, America, British, Tanzania, Uganda, Kinya, South and East Africa. According to the number of native speakers, Sindhi stands 23rd in the most speaking languages in the world, currently spoken by 75 million people. Currently, it is frequently used on the web and on the official websites of Sindh government, literary websites, online newspapers, social-media platforms, social- blogs, discussion forum and recently added on the google translator. It is imperative to mention that currently Sindhi can be used on any type of personal computer.

### *A. Morphological Characteristics.*

Sindhi is an Indo-Aryan poly-morphemic language like Arabic, Persian and Urdu. It has very rich and complex morphological structure, inflectional and derivational. Sindhi language has the capability to borrow words from other languages. Its morphology is complex due to word formation in

978-1-5090-3310-2/17/$31.00 ©2017 IEEE

many ways, some ambiguities like homogenous word structures and vowel deletion. The richness of morphological structure makes it more complex, however inflectional and derivational features are considered the characteristics of great language of the world. Two word types primary and secondary words are found frequently. The primary words are known as the minimum free forms which are not further divisible and secondary words are further categorized as compound and complex words. Complex words are made by adding affixes to the primary or root words, while compound words are the combination of two or more simple words. The prefixes and suffixes are the free form morphemes. The variations in Sindhi morphology, create many problems for its computational processing. These complexities also include different positions of word replacements like prefixes, suffixes and stems. The inflection and derivation of words are found frequently with the usage of prefix and suffix words. These unique characteristics of Sindhi are evidence for its richness in morphology.

### *B. Writing System*

Historically Sindhi has many writing systems, such as Persio-Arabic, Devanagari, Khojki, Landa, Gurmukhi and Khudabadi but currently only Persio-Arabic and Devanagari scripts are available. Sindhi Persio-Arabic is standard script and most of the literature is available in this form. Only in India both scripts are adopted. Devanagari script of Sindhi resembles with Hindi or other Indian languages and very little bit literature is available in this script. Sindhi Persio-Arabic is written right to left while Devanagari is written left to right. The diacritic marks are not used in daily routine of Sindhi Persio-Arabic Script but in Sindhi-Devanagari script diacritic marks are mandatory. Table 1. Shows the list of characters, figure 1. Shows complete sequence of Sindhi alphabet and an example of Sindhi Persio-Arabic writing is given in figure 2. respectively.

TABLE I. LIST OF SINDHI PERSIO-ARABIC ALPHABET CHARACTERS.

| Arabic letters | 29 letters |
|---|---|
| Persian letters | 03 letters |
| Modified letters | 20 letters |
| Total | 52 letters |

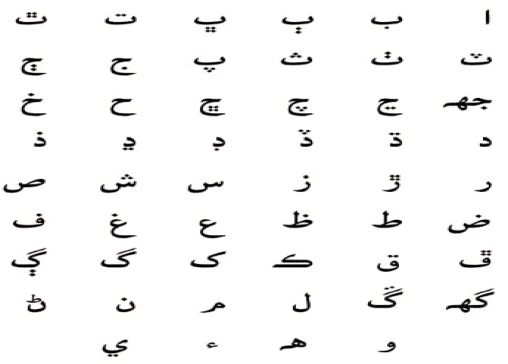

Fig.1.Sindhi Persio-Arabic script sequence of characters.

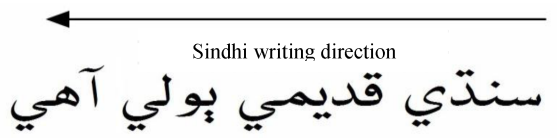


Fig. 2. Writing system of Sindhi Persio-Arabic Script *(Sindhi is ancient language).*

## III. SINDHI NATURAL LANGUAGE PROCESSING APPLICATIONS

Previously, SLP received little bit interest of the researchers due scarcity of language resources but currency many application have been developed by applying hand-engineered and statistical methods. These applications include morphological analysis, word segmentation, POS tagging, Diacritization, Spell Checker, NER, OCR, word recognition, TTS synthesis, digital thesaurus, stemmer and some language resources developed by individual researchers. The detailed progress made in these SLP applications is presented onward.

### *A. Morphological Analysis*

The morphological analyser is used to capture the inherent properties of words and morphological features of the language, which help to improve other NLP applications such as POS tagging, Parsing, MT and spell checking [2]. It deals with the processing of word forms without considering the context for two essential purposes. One is to acquire grammatical information from work forms and second is to assign word forms to its base forms. A significant work has been carried out for the morphological analysis of Sindhi. Recently Motlani [2] has developed tools and resources for SLP, in which raw and annotated data-sets have been created and POS tagger, Morphological analyser, transliteration system and a Statistical MT system in an unsupervised fashion for Sindhi language. The developed tools are applied on some different data-sets. The Finite-State Transducers (FSTs) are proposed by Rehman [5] for Sindhi morphological construction and noun inflection. In their work noun inflection rules are investigated for Sindhi language and equivalent computational rules used by FSTs are presented. Motlani [6] developed first free and open source finite-state morphological analyser for Sindhi. An Apertium's toolbox is adopted as the finite-state toolkit for the implementation of transducer. A paradigm- based approach has been adopted for the development of proposed analyser, wherein, all word forms have been defined and their morphological characteristics for given lemma (stem). The proposed analyser is evaluated on freely available large corpus on the Sindhi-Wikipedia. The system reached the reasonable coverage of 81% and 97% precision. The brief introduction of Sindhi Morphological analysis task is presented by Narejo [7], which includes important aspects of morphological structure of Sindhi language. In their work formation of words is explained according to the word types like prefixes, suffixes and their influence on other words. Furthermore, their work also covers the important areas of Sindhi morphology like structure, function, nature and categories of words like prefix, suffix, compound and prefix- suffix words in Persio-Arabic script of Sindhi language. The task of statistical language modeling proposed by Mahar [24] based on word prediction for Sindhi by adopting bi-gram, tri-gram and 4-gram language modeling

techniques are investigated in order to predict words. It is proposed that 4-gram model is most efficient for word prediction in Sindhi.

### *B. Word Segmentation*

The segmentation technique is primary and crucial step for un-segmented languages. This work has great importance in SLP because segmentation of Sindhi text is basic pre-requisite for many applications like diacritization, text-to-speech synthesis and text recognition. The most accurate segmented text can be used for other NLP tasks. Mahar [8] proposed an algorithm for the word segmentation of Sindhi language by adopting lexicon driven approach. The experiment was initiated on 16,601 words containing prefixes, suffixes and stems. The proposed algorithm is tested on 3,984 words and reported 9.54% cumulative Segmentation Error Rate (SER). Recently, Narejo [9] has designed and implemented an algorithm for the segmentation of compound and complex words. Their system is able to segment the words into their possible morpheme. However, the system is tested on a small amount of data, for the testing purpose 109 compound words, 179 prefix, 1343 suffix, and 50 prefix-suffix words have been used and reported acceptable cumulative SER of 5.02% by applying proposed algorithm on both compound and Complex Sindhi words. The ability of proposed algorithm is to deal with all basic grammatical units of the language like root words, prefixes and suffixes. The presented work is a foundation for the segmentation of complex and compound words jointly into their minimum free forms. Moreover, it is claimed in their work, that the SER can be further reduced by increasing the context of lexicons. The work presented by Bhatti [10] is great initiative for the segmentation of Sindhi language text. In proposed work various techniques and algorithms have been proposed and implemented in order to tokenize words from given text documents. The tokenization model identifies the sentence boundaries and extract sentences in an isolated form. The separated sentences are further divided into words hard space characters by using word boundaries and soft spaces considering as part of word and thus ignored from segmentation. In the end each word is filtered for removal of special character and then word is converted and saved as token after the validation process. In the first step, words are segmented into tokens and secondly each generated token is verified. The results are validated by adopting pre-built words source. For this purpose, Sindhi word processor is adopted as primary tool for testing purpose and for display of the results. This proposed word segmentation model works in four different levels which is able to perform text segmentation into sentences, segmentation of sentences into words, creation of tokens and token matching. The reported results of the segmentation model are very encouraging and accurate than previous reported results on a large data-set. It is claimed is their presented work that the same model can also be applied on other NLP task such as MT, Grammar Checking, Spell Checking and Text to Speech synthesis. Mahar [11] proposed a model for segmentation of Sindhi text into word tokens on a corpus contained of 11,124 sentences and 108,556 words. The system consists of 3 layers. The first layer is designed to get input text, segment the input text and then search the matching of word from lexicons. Second layer of the proposed model deals the segmentation of simple and compound words into tokens, and lastly those words having typing error or wrong spelling are handled in 3rd layer. It is reported that the proposed system achieved the accuracy up-to 100% in tokenization of simple words and compound words are segmented at 83.9% accuracy. The overall accuracy of the proposed system is 91.76% respectively.

TABLE II. PRESENTS THE RESULTS OF SINDHI WORD SEGMENTATION TASK.

| Developer | Data-sets | Accuracy% | SER % |
|---|---|---|---|
| Mahar et al. | Lexicon 3,984 words | - | 9.54 |
| Narejo et al. | 1681 compound and complex Words | 94.08 | 5.02 |
| Bhatti et al. | Articles, news articles, books and dictionary total 1,57,509 words | 92.78 | - |
| Mahar et al | 108,556 Words | 91.76 | - |

### *C. Parts-of-Speech Tagging*

The application of POS tagging is used to assign grammatical classes to words in sentence. It is difficult task due to some challenges like ambiguity in parts-of-speech. However, these ambiguous POS are most common features in majority of the world languages. Therefore, the problem of POS tagging is handled by sentence-level and word-level. In sentence-level series of tags corresponding to the sequence of words are obtained whereas the word-level POS tagging is modeled as a classification problem, in which an appropriate tag for a word is found. This application is used as the fundamental process in many computational linguistics tasks such as: Information Extraction (IE), MT, Speech Recognition, Speech Synthesis, Question Answering (QA), and Information Retrieval (IR) etc. The first POS tagging system for Sindhi (Persio-Arabic) was developed by Mahar [15] by applying Rule-based approach on a corpus of Sindhi dictionary. The lexicons of 26,366 entries from the corpus of dictionary and tag-set contained 67 tags were adopted for this task. Both resources were used along- with 186 rules for disambiguation. For testing purpose 1,500 sentences are used, consists of 6,783 words, taken from Daily Kawish Sindhi newspaper. Their proposed tagger is trained on lexicons and tested on 6,783 words. The reported overall accuracy of proposed rule-based POS tagging system is 96.28%. Another Rule-based POS tagging system is developed by Mahar [16] for Sindhi by adopting WordNet technique in order to mitigate the ambiguous situations in un-diacritized text. The proposed tagger was trained on the corpus of comprehensive Sindhi dictionary and tested on lexicon contained on 26,366 tagged words. The performance of their system reached up-to 97.14%. In these reported experimental results, it is analysed that proposed tagger performed well on the present and past sentences but the accuracy of the tagger remained lower on poetry and future tense sentences.

The Conditional Random Field (CRF) based technique is proposed by Motlani [17] for POS tagging task on Sindhi-

Devanagari script on the tag-set of Bureau of Indian Standards (BIS). The proposed POS tagger obtained the accuracy up-to 92% respectively. Additionally, some key steps are also suggested in detail in order to develop POS tagging system. The Sindhi POS tagging systems based on Rule-based syntactic and semantic are compared by Mahar [18] in which both approaches have been applied on the lexicons contained on 26,366 word tokens. The syntactic approach reached the accuracy up-to 96.28%, while semantic approach achieved the 97.14% accuracy. These supervised approaches are applied in order to reduce ambiguous situations on the un-diacritized lexicons. It is analysed in their study that both the taggers do not perform well on poetry and the future tense sentences. However, proposed tagger is efficient on the simple present and simple past sentences. Table III. shows the reported results of Sindhi POS tagging task.

TABLE III. REPORTED RESULTS OF SINDHI POS TAGGING.

| Developer | Approach | Data-sets | Script | Accuracy % |
|---|---|---|---|---|
| Mahar et al. | Rule-based | Lexicon | Persio-Arabic | 92.28 |
| Mahar et al. | WordNet | CSML and. SJR | Persio-Arabic | 97.14 |
| Motlani et al. | CRF | BIS | Devanagari | 92.6 |

*D. Diacritic Restoration*

The diacritization is known in the literature as diacritic restoration, vowelization and accent restoration. Using diacritic marks is the important characteristics of many languages like Arabic, Urdu and Sindhi. Although currently Sindhi Persio-Arabic is written without diacritic marks in daily routine work but while performing computational operations, many ambiguous situations will create. Therefore, in order to tackle these ambiguities diacritization marks are assigned in order to properly pronounce the words. The missing of these diacritic marks creates problems like semantic and syntactic ambiguities. The application of diacritization of Sindhi language text was initially was investigated by Mahar [25] by using WordNet technique which is based on the word analogy. Their proposed automatic diacritic restoration system rely on the WordNet technique for the identification of analogical relation of words. The reported Diacritization Error Rate (DER) is 3.39% and Word Error Rate (WER) is 0.71% respectively. It is suggested that the proposed method is also applicable for other languages like Arabic, Urdu and Persian. Another automatic diacritic restoration system developed by Mahar [26], in which an innovative diacritization system was applied on the non- diacritized text by multiplying 3 N-Gram based probabilities and well known Viterbi algorithm is adopted. The experiment is performed on the text of Shah-Jo-Risalso (SJR) contained on 27,360 words for the training and testing. The proposed system compares the non-diacritized word with diacritized, then select the highest count of diacritized word. After implementation phase a tokenization scheme is used in order to reduce the ambiguities. The accuracy of their proposed system reached up-to 81.4%, and DER of 3.21% and WER of 0.71% respectively. Furthermore, Mahar [27] has proposed and implemented and algorithm on Lexicons obtained from Shah-Jo-Risalo (SJR) and Corpus of Modern Sindhi Language (CMSL) for letter level diacritic restoration. The k- nearest neighbor is implemented for the classification of instances and at last the nearest instance is taken for the replacement of non-diacritized letter. The CMSL dataset contain on 146,118 sentences and the lexicons of SJR consist of 27,360 words. Both datasets have been adopted in order to implement letter level diacritization approach which is much better than previous work of Mahar [25,26] without any other supporting tool or technique which is great sign towards the development of accurate diacritization system for Sindhi language. Another benefit of their proposed model is its ability to deal with out of vocabulary words. Therefore, their proposed system has a capability of the implantation on other languages without and with grammatical or linguistics rules. The reported accuracy on CSML dataset is 98.95%, and SJR reached up-to 97.32% respectively. The reported DER of CMSL is 1.04% and Corpus of SJR is 2.68%, cumulative DER of 1.9% is reported.

TABLE IV. REPORTED RESULTS OF SINDHI DIACRITIZATION TASK.

| Developer | Approach | Data-sets | DER% | WER% |
|---|---|---|---|---|
| Mahar et al. | WordNet | Lexicon of SJR | 3.39 | 0.71 |
| Mahar et al. | N-Grams | 27360 words | 3.21 | 0.71 |
| Mahar et al. | letter level | CSML and SJR | 1.9 | 0.71 |

*E. Named Entity Recognition*

Named Entities are the atomic elements in the text and NER is said to be entity identification, entity chunking and entity extraction. It is used to find Named Entities and classify them into pre-defined categories, such as names of person, organization, location, designation, title of person and object, brand, measurement, abbreviation, date and time within the text. This application has vital importance and used in almost all NLP applications. Although NER tool is not developed, only Jamro [23] has discussed the applications, challenges and future research opportunities concerned with the development of Sindhi NER system. Along with this a Maximum Entropy (ME) based algorithm is also proposed for the development of NER system. This presented work is beneficial in order to initiate the work on Sindhi NER task.

*F. Spell Checker*

Spell checker is vital component used in word processors, MT, web and mobile applications. Bhatti [30] has explored the spelling error trends and patterns in Sindhi Language. Among the various error trends, the most frequent found the similar shape of letters and similar pronunciation. The more error trends discussed in the study are omission of space characters at word boundaries. It is also found that most of the error trends are common to all languages which also encountered Sindhi but also some error exists specifically to Sindhi Language. The first spell

checker for Sindhi was developed by Bhatti [31] by using Phonetic based Sindhi SoundEx and ShapeEx algorithm. The combinational approach is adopted in order to generate list of similar words mis-spelled Sindhi string. The proposed system is able to identify mis-spelled words properly and then suggest the list of accurate words. Furthermore, Mughal [32] has also developed comprehensive spell checker for Sindhi by adopting hybrid approach, utilized dictionary and statistical analysis on the corpus. The spell checker suggested 79% of the words out of 3,336 errors. Umair [33] analysed the Sindhi spelling error patterns in order to detect errors and their correction. The proposed statistical analysis is helpful for the development of robust Sindhi spell checker. Another study carried out by Rehman [34] suggest that W3C XML schema can be useful for the sentence validation of Sindhi and Urdu languages.

### *G. Optical Character Recognition*

The application of OCR is used to transform text image into the digital format. In Sindhi language, initially, a study regarding issues and challenges for the development of Sindhi OCR system was led by Hakro [43], in which various scripts including old scripts and other distinct features of the standard Sindhi Persio-Arabic have been discussed in detail. The Sindhi OCR system Hakro [44] is designed in order to recognize isolated characters and the detailed process is also presented for the development of Sindhi OCR system. Their proposed system is tested on the various images of Sindhi alphabet, which is able to recognize isolated characters into the audible text. The reported accuracy of their system is 93% for machine printed isolated-characters of Sindhi language. An algorithm is proposed by Hakro [45] for the iterative and interactive thinning, for Sindhi language which is useful for segmentation-based and segmentation-free Sindhi OCR and it is also suggested that the same algorithm is also adoptable for the other language, who have related script with Sindhi. Another algorithm is proposed by Shaikh [46] for Sindhi character recognition by using high profile vector, which is tested on the many types of printed text of Sindhi language. It is demonstrated in their work that the proposed algorithm works in both situations namely under and over segmentation. Moreover, a survey based study Hakro [47] regarding techniques, methods and process of Sindhi OCR are presented and also discussed various techniques adopted by the researchers. The Back-Propagation ANN model is also proposed by Nizamani [50] for addressing the problem of Sindhi OCR by adopting un-supervised learning approach.

### *H. Word Recognition*

An online hand-writing recognition system Chandio [48], is developed by using Artificial Neural Network (ANN). Unsupervised strategy is adopted for the training purpose of the model. The dataset of 1200 words is used for the experiment, which were received from native speakers of Sindhi. The reported overall accuracy of their system is 83%, on the android based application, after testing from the 25 users. Nizamani [49] has also applied ANN for hand-writing recognition in Sindhi language also some complexities regarding had-writing recognition in Sindhi are also presented. The accuracy of the system of native writers is 91% and the writing of non-native speakers is recognized at 79%. The overall accuracy of 87.75 is reported.

### *I. Text to Speech Synthesis*

The TTS synthesis is used to convert arbitrary text into the audible speech in order to provide the facility to people via voice messages rather than text. A Bi-lingual TTS System was developed by Shah [20], for Urdu and Sindhi languages by adopting a knowledge-based with rule-based and concatenative acoustic method. It is shown in the presented work that the system achieved high accuracy and it is also adoptable for other applications, enough versatile and also useful for the speech recognition task. The phonology for the Letter-To-Sound (LTS) conversion is formulated Mahar [41] for the development of rule-based TTS system. Current Sindhi writing system, groups of letters shapes, various aspects of phonology, such as phonemes, syllables, structure and stress of the pronunciation in Sindhi language are discussed comprehensively.

### *J. Digital Thesaurus*

Thesaurus for Sindhi language is designed and developed by Bhatti [51] by adopting the hash table structure as a database for storing of word repository. The user interface is specifically developed for Sindhi users. The proposed thesaurus consists of 16,000 words, having synonyms an antonyms.

### *K. Stemmer*

Recently, Shah [52] proposed first rule-based stemmer for Sindhi by using stripping approach. The stemmer is tested on the dataset of 5327 words out of them 2142 prefix/suffix words. The performance is calculated separately of prefix, suffix words and overall accuracy of 84.85% and SER of 15.15 is reported. Although the performance of the proposed stemmer is not good but it is good initiative for future research.

## IV. TOOLS AND APPLICATIONNS DEVELOPED FOR SINDHI LANGUAGE

The language resources and tools have the vital importance in order to ensure the digital survival of any human language. Although such resources, tools and applications have been developed by the individual researchers but these are not available publicly. Therefore, the computational processing of Sindhi becomes more challenging and demands more efforts. In this section summary of all developed datasets, tools and SLP applications is presented in table. 5, for both Sindhi language scripts Persio-Arabic and Devanagari.

TABLE. V. SUMMARY OF DATASETS, TOOLS AND APPLICATIONS DEVELOPED FOR BOTH SCRIPTS OF SINDHI LANGUAGE.

| Datasets/Tools and Applications | Devanagri | Persio-Arabic |
|---|---|---|
| POS Annotated Data | Yes | Yes |
| Urdu-Sindhi parallel data | No | Yes |
| Morphological Analyser | No | Yes |
| POS Tagger | Yes | Yes |

| | | |
|---|---|---|
| Spell Checker | No | Yes |
| Transliteration | Yes | Yes |
| Diacritization | No | Yes |
| OCR system | No | Yes |
| TTS system | No | Yes |
| Digital Thesaurus | No | Yes |
| Stemmer | No | Yes |
| MT (Urdu-Sindhi) | No | Yes |

Table. V. The Datasets, Tools and techniques for both scripts of Sindhi language.

## V. DISCUSSION AND CONCLUSION

Although, SLP is getting attention of the researchers but still lot of efforts are required in order to assign maturity level. No doubt Sindhi language has a historical literary background and great influence on large population but unfortunately there is scarcity of resources for language processing and still no language processing tools have been made publically available except a Finite-State morphological analyzer. The sharing of language resource can boost the SLP task. However, dictionaries, word processors and large amount of raw corpus is available on the web, news channels and social blogs of Sindhi Persio-Arabic. Although pretty work has been initiated on SLP, but there is dire need to investigate and develop more sophisticated techniques and tools for its digital survival. Except this, sharing of the developed language resources and tools will highly benefit the SLP for fast and collaborative computational development.

We have presented a brief overview of Sindhi language and comprehensive survey regarding progress made in Sindhi language processing. The importance of SLP is highlighted and reported results and techniques are presented. The overall progress made in SLP is systematically presented in this survey paper along with reported results and applied techniques. The resources and tools developed by the individual researchers are also presented in this work. Except this future research directions are also suggested. This work will help the researchers at present and upcoming era of Sindhi language processing.

## VI. FUTURE RESEARCH

It is very encouraging that SLP has taken the attention of researchers but still some research areas are not investigated yet. Therefore, the existing work on SLP can be further extended by resource sharing and applying more sophisticated statistical techniques on large datasets. The development of language technology tools and resources ensure the digital survival of the language. The sharing of language resource not only motivate the researchers but also Sindhi will become mature language and reach state-of-the-art performance in the field of NLP. In this regard the machine learning (ML) techniques have been successfully experienced in many NLP tasks, which have shown state-of-the-art performance because of their trainable nature, cost effectiveness and less time consuming features. Therefore, by adopting statistical approaches alone and combining both statistical and hand-engineered techniques will bring a revolutional development in the field of SLP. In order to initiate research, large amount of raw content of Persio-Arabic script is available on the web such as on literary web-sites, Sindhi news channels, social blogs and online newspapers etc. There are many open and encouraging research areas like development of language resources i-e. annotation of large datasets, development of digital lexicons, language technology tools like automatic morphological analyser, word sense disambiguation systems, Parsing, Language Modeling, pure statistical and efficient POS tagging system, efficient diacritic restoration system, NER system, Machine Translation and cross-lingual technologies.

## REFERENCES


[1] Hussain, S. (2003, April). Computational Linguistics (CL) in Pakistan: Issues and Proposals. In EACL 2003 (pp. 31-33).

[2] Motlani, R. (2016, June). Developing language technology tools and resources for a resource-poor language: Sindhi. In Proceedings of NAACL-HLT (pp. 51-58).

[3] Vikas, O. (2001). Language Technology Development in India. Ministry of Information Technology, New Delhi, India.

[4] Leghari, M., & Rahman, M. U. (2010). Towards Transliteration between Sindhi Scripts by using Roman Script. In the Conference on Language and Technology, National Language Authority Islamabad, Pakistan.

[5] Rahman, M. U., & Bhatti, M. I. (2011). Finite State Morphology and Sindhi Noun Inflections.

[6] Motlani, R., Tyers, F. M., & Sharma, D. M. (2016). A finite-state morphological analyser for sindhi. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16).

[7] Narejo, W. A., & Mahar, J. A. (2016, April). Morphology: Sindhi morphological analysis for natural language processing applications. In 2016 International Conference on Computing, Electronic and Electrical Engineering (ICE Cube) (pp. 27-31). IEEE.

[8] Mahar, J. A., Memon, G. Q., & Danwar, S. H. (2011). Algorithms For Sindhi Word Segmentation Using Lexicon-Driven Approach.International journal of academic research, 3(3).

[9] Waqar Ali Narejo, Javed Ahmed Mahar, Shahid Ali Mahar, Farhan Ali Surahio, Awais Khan Jumani. Morphological analysis: Sindhi Morphological Analysis: An Algorithm for Sindhi Word Segmentation into Morphemes. International Journal of Computer Science and Information Security (IJCSIS), Vol. 14, No. 6, June 2016,

[10] Bhatti, Z., Ismaili, I. A., & Soomro, W. J. (2014). Word segmentation model for Sindhi text. American Journal of Computing Research Repository, 2(1), 1-7.

[11] Mahar, J. A., Shaikh, H., & Memon, G. Q. (2012). A Model for Sindhi Text Segmentation into Word Tokens. Sindh University Research Journal-SURJ (Science Series), 44(1).

[12] Rahman, M. U. (2009). Sindhi Morphology and Noun Inflections. In Proceedings of the Conference on Language & Technology (pp.74-81).

[13] Rahman, M. U. (2010). Towards Sindhi corpus construction. In Conference on Language and Technology, Lahore, Pakistan.

[14] Nainwani, P. (2012, May). Blurring the demarcation between Machine Assisted Translation (MAT) and Machine Translation (MT): the case of English and Sindhi. In Workshop on Indian Language and Data: Resources and Evaluation Workshop Programme (p. 139)

[15] Mahar, J. A., & Memon, G. Q. (2010, February). Rule based part of speech tagging of sindhi language. In Signal Acquisition and Processing, 2010. ICSAP'10. International Conference on (pp. 101-106). IEEE.

[16] Mahar, J. A., & Memon, G. Q. (2010). Sindhi Part of Speech Tagging System using WordNet. International Journal of Computer Theory and Engineering, 2(4), 538.

[17] Motlani, R., Lalwanil, H., Shrivastava, M., & Sharma, D. M. Developing Part-of-Speech Tagger for a Resource Poor Language: Sindhi.

[18] Mahar, J. A., Shaikh, H., & Sangi, A. R. (2011). COMPARATIVE ANALYSIS OF RULE BASED SYNTACTIC AND SEMANTIC SINDHI PARTS OF SPEECH TAGGING SYSTEMS. International Journal of Academic Research, 3(5).

[19] Mahar, J. A., Memon, G. Q., & Shah, S. H. A. (2010, May). WordNet based Sindhi text to speech synthesis system. In Computer Research and Development, 2010 Second International Conference on (pp. 20-24). IEEE.

[20] Shah, A. A., Ansari, A. W., & Das, L. (2004). Bi-Lingual Text to Speech Synthesis System for Urdu and Sindhi. In National Conf. on Emerging Technologies (pp. 20126-130).

[21] Raza, S., Agha, F. Z., & Usman, R. (2004). Phonemic inventory of Sindhi and acoustic analysis of voiced implosives. Center for Research in Urdu language Processing (CRULP).

[22] Ismaili, I. A., Bhatti, Z., & Shah, A. A. (2014). Design & Development of the Graphical User Interface for Sindhi Language. arXiv preprint arXiv:1401.1486.

[23] Jamro, W.A. Kehar, A., Shaikh, H. (2015) "Towards Sindhi Named Entity Recognition: Challenges and Opportunites", TIIT-16 conference at QUEST, Nawabshah, Pakistan.

[24] Mahar, J. A., & Memon, G. Q. (2011). Probabilistic Analysis of Sindhi Word Prediction using N-Grams. Australian Journal of Basic and Applied Sciences, 5(5), 1137-1143.

[25] Mahar, J., & Memon, G. (2011). Automatic Diacritics Restoration System for Sindhi. Sindh University Research Journal-SURJ (Science Series), 43(1).

[26] Mahar, J. A., Memon, G. Q., & Shaikh, H. (2011). Sindhi diacritics restoration by letter level learning approach. Sindh University Research Journal-SURJ (Science Series), 43(2).

[27] Mahar, J. A., & Memon, G. Q. (2011). Lexicon Based Diacritic Restorations Using Wordnet For Sindhi. International Journal of Academic Research, 3(2).

[28] Shaikh, H., Mahar, J. A., & Malah, G. A. (2013). Digital Investigation of Accent Variation in Sindhi Dialects. Indian Journal of Science and Technology, 6(10), 5429-5433.

[29] Saeed, S., Zaman, S., Qurat-ul-Ain. (2004). Syntactical Translation System for English to Sindhi Translation. National Conference on Emerging Technologies. (pp.112-115)

[30] Bhatti, Z., Ismaili, I. A., Shaikh, A. A., & Javaid, W. (2014). Spelling error trends and patterns in Sindhi. arXiv preprint arXiv:1403.4759.

[31] Bhatti, Z., Waqas, A., Ismaili, I. A., Hakro, D. N., & Soomro, W. J. (2014). Phonetic based SoundEx & ShapeEx algorithm for Sindhi Spell Checker System. arXiv preprint arXiv:1405.3033.

[32] Mughal, M. U. (2016). Sindhi Spelling Error Detection And Correction- A Hybrid Approach (Doctoral dissertation).

[33] Umair, M., & Rahman, M. U. Analysis of Sindhi Spelling Error Patterns for Spelling Error Detection and Correction.

[34] Rahman, M., & Shah, A. (2004). Grammar Checking of Urdu and Sindhi Sentences by Using W3C XML Schema. In National Conference on Emerging Technologies (p. 120).

[35] Bhatti, Z., & Jarwar, A. A. (2013). Sindhi Academic Informatic Portal. American Journal of Information Systems, 1(1), 21-25.

[36] Abbasi, A. M., & Hussain, S. (2015). Phonetic Analysis of Lexical Stress in Sindhi. Sindh University Research Journal-SURJ (Science Series), 47(4).

[37] Mahar, J. A., & Memon, G. Q. (2009). Phonology for Sindhi letter to sound conversion. Journal of Information and Communication Technology, 3(1), 11-20.

[38] Keerio, A., Channa, N., Mitra, B., Young, R., & Chatwin, C. (2014). Acoustics of isolated vowel sounds of Sindhi. Sindh University Research Journal-SURJ (Science Series), 46(2).

[39] Raza, S., Agha, F. Z., & Usman, R. (2004). Phonemic inventory of Sindhi and acoustic analysis of voiced implosives. Center for Research in Urdu language Processing (CRULP).

[40] KEERIO, A., CHANNA, N., MALKANI, Y., QURESHI, B., & CHANDIO, J. (2014). Acoustic Analysis of the Liquid Class of Consonant Sounds of Sindhi. Sindh University Research Journal-SURJ (Science Series), 46(4).

[41] Mahar, J. A., & Memon, G. Q. (2009). Phonology for Sindhi letter to sound conversion. Journal of Information and Communication Technology, 3(1), 11-20.

[42] Abbas, Q., Ahmed, M. S., & Niazi, S. (2010). Language Identifier For Languages Of Pakistan Including Arabic And Persian. International Journal of Computational Linguistics (IJCL), 1(03), 27-35.

[43] Hakro, D. N., Ismaili, I. A., Talib, A. Z., Bhatti, Z., & Mojai, G. N. (2014). Issues and challenges in Sindhi OCR. Sindh University Research Journal-SURJ (Science Series), 46(2).

[44] HAKRO, D., MEMON, M., AWAN, S., BHUTTO, Z., & HAMEED, M. (2016). Isolated Optical Character Recognition. Sindh University Research Journal-SURJ (Science Series), 48(4).

[45] HAKRO, D., AWAN, S., MEMON, M., AAMUR, A., & MOJAI, G. (2015). Interactive Thinning for Segmentation-based and Segmentation-free Sindhi OCR. Sindh University Research Journal-SURJ (Science Series), 47(3).

[46] Shaikh, N. A., Mallah, G. A., & Shaikh, Z. A. (2009). Character segmentation of Sindhi, an Arabic style scripting language, using height profile vector. Australian Journal of Basic and Applied Sciences, 3(4), 4160-4169.

[47] Hakro, D. N., Talib, A. Z., Bhatti, Z., & Moja, G. N. (2014). A Study of Sindhi Related and Arabic Script Adapted languages Recognition. arXiv preprint arXiv:1412.4217.

[48] CHANDIO, A., LEGHARI, M., HAKRO, D., AWAN, S., & JALBANI, A. (2016). A Novel Approach for Online Sindhi Handwritten Word Recognition using Neural Network. Sindh University Research Journal-SURJ (Science Series), 48(1).

[49] Nizamani, A. M., & Janjua, N. U. H. (2012). Isolated Handwritten Character Recognition in Sindhi Language using Artificial Neural Network. Journal of Independent Studies and Research, 10(1), 17.

[50] Nizamani, A. M., & Janjua, N. U. H. (2013). Sindhi OCR using Back propagation Neural Network. International Journal of Advanced Computer Science, 3(3).

[51] BHATTI, Z., ISMAILI, I., ZARDARI, S., & SOOMRO, W. (2016). Design and Development of Unicode based Sindhi Language Thesaurus. Sindh University Research Journal-SURJ (Science Series), 48(4).

[52] SHAH, M., SHAIKH, H., MAHAR, J., & MAHAR, S. (2016). Sindhi Stemmer for Information Retrieval System Using Rule-Based Stripping Approach. Sindh University Research Journal-SURJ (Science Series), 48(4).

[53] Bhatti, Z., Ismaili, I. A., Khan, W. I., & Nizamani, A. S. (2013). Development of Unicode based Sindhi typing system. Journal of Emerging Trends in Computing and Information Sciences, 4(3), 1-21.

[54] Hristea, F. T. (2011). Statistical Natural Language Processing. In International Encyclopedia of Statistical Science (pp. 1452-1453). Springer Berlin Heidelberg.

[55] Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. Journal of Machine Learning Research, 12(Aug), 2493-2537.