

Romanized Urdu Corpus Development (RUCD) Model: Edit-Distance Based Most Frequent Unique Unigram Extraction Approach Using Real-Time Interactive Dataset

Faisal Baseer¹, Asad Habib², Jawad Ashraf³
Institute of Information Technology (IIT),
Kohat University of Science and Technology (KUST)
Kohat, Pakistan
faisalbaseer2000@yahoo.com
asadhabib@kust.edu.pk
jawad.ashraf@kust.edu.pk



ABSTRACT: Urdu ranks very high among languages used for communication in the Southern Asia. Even though with great following, it clearly lacks computational support that is why it is written in Romanized Urdu script. Even though, a lot of Romanized Urdu data is available online but it still lacks a refined Corpus. In our research, we have proposed a refined Romanized Urdu Corpus using tokens with the highest frequency of occurrence in the data set, which was collected from volunteer participants who used this language as a mode of communication interactively.

The raw corpus is passed through a series of steps such as Preprocessing, Tokenization and Annotation before passing it to computationally extensive subsequent steps. "Edit Distance" and "K-means Clustering" techniques are used for identification of candidate tokens and their potential selection/ inclusion in the refined lexicon. We have also identified most commonly used tokens, candidate tokens and other linguistic attributes from the data collected. Based on analysis, we have proposed a computational model for refined colloquial Romanized Urdu lexicon development.

Key words: Urdu Corpus Development, Colloquial Urdu Corpus, Romanized Urdu Corpus, Computational Lexeme Extraction, Natural Language Engineering.

Received: 15 April 2016, Revised 18 May 2016, Accepted 25 May 2016

© 2016 DLINE. All Rights Reserved

1. Introduction

With the growing popularity of the Internet in the recent past, many new trends of communication have evolved. Internet and telecommunication have paved paths for the new forms of the textual communication. Weblogs, Internet based social media, Emails, SNS feeds, Tweets, Simple Mail Services (SMS) and Multimedia Mail Services (MMS) which are now considered a very

fast and reliable source of communication, have given birth to informal languages. These Informal languages in general and their dynamics (Lexica/ Corpus development) in particular are the contemporary issues of the linguistic domains.

One of the most frequently spoken languages across Southern Asia is Urdu. With more than 50 million regular users of the language across countries like Pakistan, Afghanistan, India, Iran and Bangladesh, Urdu ranks 19th among the 7105 languages of the whole world [1]. With growing economics across the Southern Asia and introduction of global village philosophy, Urdu language's significance as a mode of communication in the defined region has grown immensely. With all the facts depicted, Urdu as a language for the computational purposes is still under-resourced. Lack of text editors supporting the Urdu Scripture writing can be termed as the main cause of the problem. Limited text editors are available for computational or text editing in Urdu. To overcome this kind of limitation, Urdu is normally written in Romanized form. In comparison with the native Urdu-script, large amount of Romanized Urdu data is available across the Internet and telecommunication modes.

Corpus development for the natural languages like Arabic, English, and French etc. has been the focus of the Natural Language Processing (NLP) researchers. Considerable work has been done in the domain but a lot is needed to be done for the Urdu language. As Romanized Urdu is broadly used for the communication for Urdu language, hence there is a great need of Refined Corpus for Romanized Urdu.

In our research, we have gathered an interactive dataset of Romanized Urdu language from assorted means like weblogs, Internet based social media, Emails, SNS feeds, Tweets, Simple Mail Services (SMS) and Multimedia Mail Services (MMS) etc. for the development of Refined Corpus for Romanized Urdu language.

Data was collected from volunteers, which was passed through initial screening for error removal and then passed for further processing through computational techniques. Edit-Distance computational techniques was used to find similarities between the strings and Subsequently, K-means clustering was applied for string grouping and extracting the most frequent string for inclusion in the Refined Reference Corpus (RRC). The corpus developed comprised of the most frequently used strings which were around 4 thousand and along with that all the candidate string were also recorded. It will also somewhat formalize the use of different vocabulary set for the Romanized Urdu. A definite set of corpus will guarantee a more formalized language and conservation with more understandability. This not only helped in the current corpus development but can also help in the auto-correct feature of any native script Urdu text editor also.

The composition of the research paper is as follows. Section 1 presents introduction of the research topic and motivation behind it. Section 1 discusses the related work and contributions in the relevant domain. Section 3 depicts methodology used, how data was collected and data source. In Section 4, a refined model has been proposed and a pseudo code has been presented for the Corpus development and description of the steps has been given. Section 5 consists of results collected by utilizing the proposed model for corpus development. Section 6 discusses the analysis of the errors encountered in the research. Section 7 and 8 is about the future work which can be done and references respectively.

2. Literature Review

The amount of Natural Language Engineering (NLE) research on Urdu language is limited [3]. However recent research trends in this domain are changing rapidly [4][10]. English to Urdu translation and transliteration system is investigated to generate Urdu text from English transcription using bi-lingual evaluation understudy [6].

Resnik and Smith used STRAND system for mining parallel text on the World Wide Web for construction of a parallel corpus for low-density language pairs [2]. Irvine et. al. Proposed a model which combines information at the word and character levels, allowing it to handle out-of-vocabulary items [5].

Conversion of informal, Romanized Urdu messages into the native Arabic script and normalization of non-standard SMS language is formulated. Ample work has been done in the field of transliteration and processing for Urdu native language as well Romanized Urdu.

Our emphasis is on the development of the Corpus for the Romanized Urdu language. In this research, edit-distance technique which is also termed as Levenshtein distance has been used for identifying how dissimilar two strings are from one another by counting the minimum number of operations required to transfer from one string to another [7][9].

K-means clustering is used afterwards for selecting most frequently used item and candidate items [8]. It is a method of vector quantization (division of large no. of points into groups same number of defined point closet to them) that is popular for the cluster analysis in data mining.

3. Methodology

In this section, we discuss our methodology for the lexicon development.

3.1 Data Source

We collected data from 643 volunteers with diverse demographic backgrounds and different walks of life, professions, educational background, religions etc. Their valuable contribution in the shape of their personal Romanized Urdu text data enabled us to study the real-life and present-day trends in language usage with particular focus on the Romanized colloquial usage of Urdu language. People with ages ranging from 17 to 54 submitted their data for our research and analysis. A great effort was made to have adequate data being collected from both genders. Gender ratio in our data collection phase was 65:35 according to males: females respectively. The raw data was passed through encoding conversion because some data provided by participants was not according to our required format. Therefore we converted all data into simple text files in order to make it suitable for computational experiments and analysis.

Our initial goal was to collect a general genre raw corpus for greater language and domains coverage. Therefore all participants were requested to submit purely natural data from the source of their own preference including Facebook posts and comments, Skype conversations, SNS feeds, Twitter tweets, Emails, SMS, MMS or simply composed Romanized Urdu text by them used anywhere for the communication. The users were requested to remove data that could reveal their privacy and guidance was provided for the rest of user data in order to manipulate data so that exact representative data could be collected.

3.2 Data Set

Our dataset is basically a general genre raw corpus of colloquial Urdu written in Romanized script. The corpus was collected from volunteer participants who used this language as a mode of communication on the Internet and text massaging through their mobile phones. The dataset was collected over a period of 14 months. It was an uphill task to find volunteers willing to share their original and colloquial data due to privacy issues. Ample efforts were made for the said purpose and many of the online and offline sources like Skype, Facebook, emails, SMS, MMS and tweets were used to collect dataset.

The basic motivation behind collected dataset from various age groups and genders was to identify, whether, writing patterns differ for different age groups and genders. This may pave path for us to identify the trends and patterns, which would be helpful in formulating the marketing strategies for E-business and other methods to reach appropriate target consumers. Our developed lexicon is expected to be useful in the development of text editors for the Romanized Urdu and development of other high level computational resources.

The raw corpus is passed through a series of steps such as Preprocessing, Tokenization and Annotation before passing it to the core and computationally extensive subsequent steps. These steps were applied in series that can be broadly categorized into multiple major categories such as removal of noisy data (Numerical data, Emoticons, Special Characters), identification and separation of English words, Netlingo e.g. lol, tc and extraction of nouns followed by Romanized Urdu and alternative or candidate words. For each category, separate reference lists were created for following experiments and further studies. Categorization also depicted usage of multi-lingual nature of the participant. Finally, Edit Distance and K-means Clustering techniques are used for identification of candidate lexemes and their potential selection for inclusion in the refined Romanized Urdu lexicon.

4. Proposed Model & pseudo Code

In order to develop a more dynamic and accurate lexicon, a formal model has been proposed with the help of which, we have developed a refined lexicon. The proposed model is illustrated in figure 1 and its functional detailed description is mentioned in the following sub-sections.

4.1 Collected Data

As mentioned earlier in the paper, data is collected from various sources offline (SMS, MMS, Text files) and Online (Emails, SNS

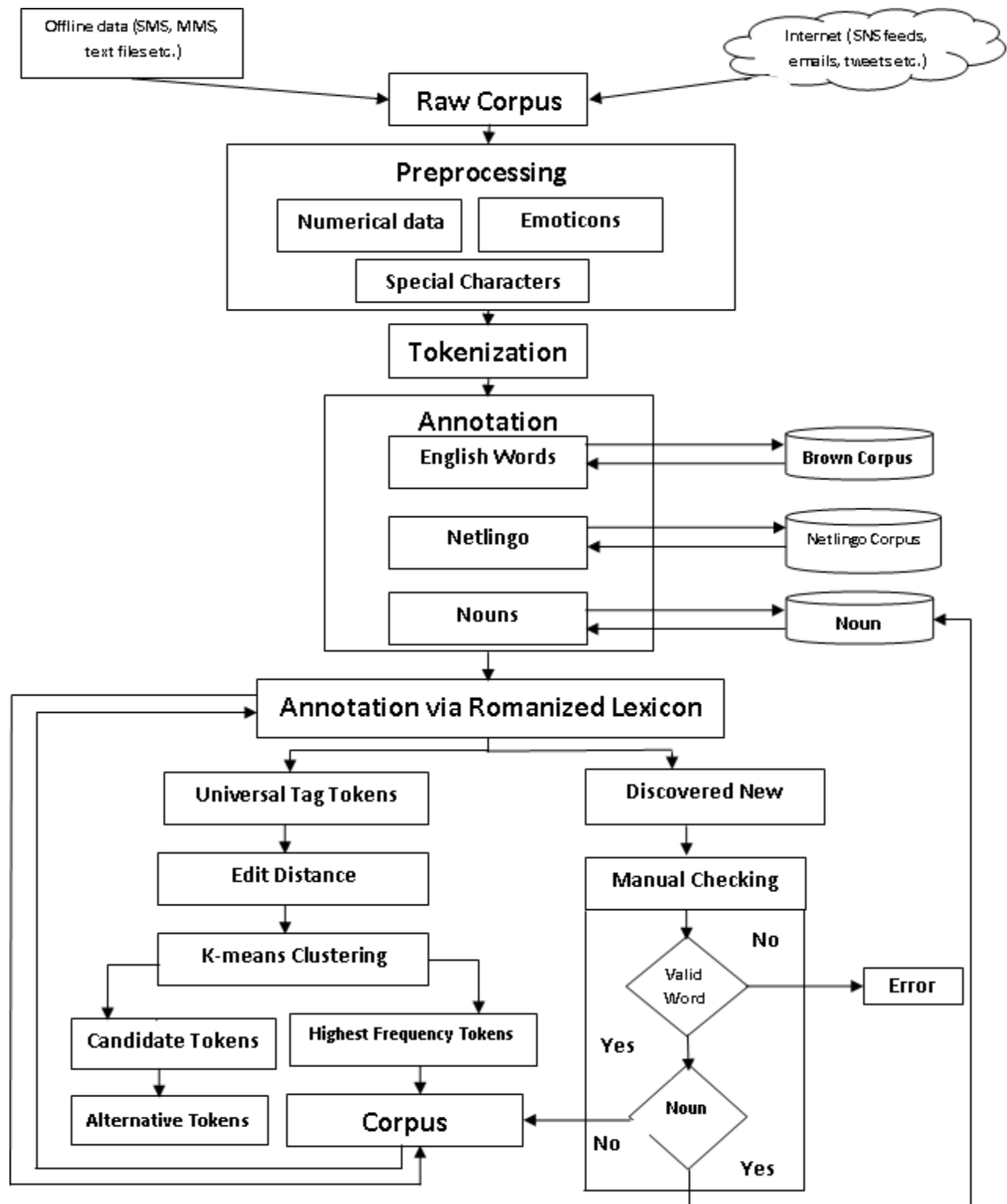


Figure 1. Proposed Model for Refined Corpus Development

MMS, Text files) and Online (Emails, SNS Feeds, Tweets etc). All of the data is saturated into single location for further processing.

PSEUDO CODE

```
i.      Extract noise in corpus using noise corpus
ii.     Tokenize all remaining words
iii.    Annotate English words in corpus using English Corpus
        If English words  $\beta$  untagged words
        SET to English complete message
        else
        SET to English Corpus
iv.     Annotate Netlingo words in corpus using Netlingo Corpus
        If NetLingo words  $\beta$  untagged words
        SET to NetLingo complete message
        else
        SET to NetLingo Corpus

v.      Annotate Known Nouns in corpus using Noun Corpus Check all remaining
        words for master_tag
vi.     If words untagged
        SET new_wordset  $\beta$  untagged words
        manually check new_wordset
        If new_wordset[n]  $\beta$  Noun Corpus
        SET to Noun Corpus
        else
        SET to Corpus
vii.    else
        COLLECT all master-tagged words into
        set roman_set
        SET roman_frequencies  $\beta$  find_frequency
        (roman_set, corpus)
        SET n  $\beta$  length(roman_set)
        CREATE ed_matrix[n][n]
        for each wc IN roman_set
        for EACH w2 IN roman_set
        If wc == w2 THEN
        ed_matrix[w1][w2]  $\beta$  0
        else
        ed_matrix[w1][w2]  $\beta$ 
        edit_distance(w1, w2)
        Get the value of k from user
        SET k_clusters  $\beta$  k_means_cluster(ed_matrix, k)
        T  $\beta$   $\emptyset$  // empty set
        for each c IN k_cluster
        SET w = get word with highest frequency in c using roman_frequencies
        T  $\beta$  t U add (w, c-w)
        END
```

Figure 2. Pseudo Code for the Proposed Model

4.2 Preprocessing

In anticipation of an error free corpus development, Preprocess of Noise removal is carried out. In this activity, Numerical data (23, 1 etc) are identified first. These are the most easy to locate and removed as they are not the part of the Urdu corpus. Emoticons such as “:-)”, “;- (“ etc are most widely used emotional symbols which are nowadays part of the natural text. They are also filtered out of the text which is forwarded for further filtering.

Special characters like “!”, “#”, “%” etc are removed as they will not be part of the corpus. Purified data is further passed on for tokenization. For example, text message:

“bat sun meri pehly jo 5 hensy hai na wahi code hai mere bhai... lol”
 will be left as
“bat sun meri pehly jo hensy hai na wahi code hai mere bhai lol “

4.3 Tokenization

All the data received after the preprocessing, is then tokenized. Tokens used and their description is defined in the following table.

Tag	Description
<EW>	English Word
<UT>	Universal Tag
<NW>	Noun Word
<NLW>	Netlingo Word

Table 1. Description of Tags

The purpose to do so is to break up the large sentences into small tokens as in the further steps these tokens will be individually identified by using different corpora. For example text message:

“bat sun meri pehly jo hensy hai na wahi code hai mere bhai lol”
 will be tokenized as

bat, sun, meri, pehly, jo, hensy, hai, na, wahi, code, hai, mere, bhai, lol

4.4 Annotation

After the tokenization, the words will be annotated against a series of reference corpora. The first will be the “Brown Corpus Lexicon” for the English words as a part of the initially identified text. From the above example, it is quite clear that the words like “bat”, “sun” & “code” will be tagged as English words. For example

bat <EW>, sun <EW>, meri, pehly, jo, hensy, hai, na, wahi, code <EW>, hai, mere, bhai, lol

Although the “bat” (بات) and “sun” (سُن) are also proper Urdu words but for our proposed model they will be tagged as English words. Next step is the annotation of the Netlingos with the help of Netlingo corpus. After so, we will have all Netlingo words tagged.

bat <EW>, sun <EW>, meri, pehly, jo, hensy, hai, na, wahi, code <EW>, hai, mere, bhai, lol <NLW>

Afterwards the remaining tokens will be annotated against the Noun corpus for the tagging of the known noun tokens.

bat <EW>, sun <EW>, meri, pehly, jo, hensy, hai, na, wahi, code <EW>, hai, mere, bhai<NLW>, lol<NLW>

In case of a complete English or Netlingo message, all the tokens will be tagged as either <EW> or <NLW>.for example

“hi there, how are you”

Hence these will be identified during annotation and will be set to separate list of “complete message” and will not be passed on to for the further processing.

5.5 Annotation via Romanized Corpus

The remaining tokens are compared against reference Romanized corpus. All the tokens already part of the corpus are Universal tagged. While the tokens, not part of the corpus are not tagged and are moved to a list of untagged new words. The untagged new words list is presented for a manual checking by a specialist, to find it out whether the new word, which is identified and added to the list, is a proper Urdu language word or not. If it is identified as an error or false word (for example “abcdef” which is by no mean a valid Urdu word), it is flushed as trash, otherwise it is again checked for its nature.




Figure 2. New Token Discovery Form

As defined in the above given figure, a new token “tamasha” has been forwarded to the specialist for the identification. Out of the three categories i.e. “Newly Discovered Token”, “Noun” or “Error”, the specialist selects the best option for the token and sends it to the appropriate corpus (if identified as Noun or valid word) or term it as an Error. The reason behind making all the words, part of the corpus (noun or general) is that all the new words occur for the first time in the text, so either they are noun or words.

Every next time, that new word identified, now part of the corpus will have either exact copy or false candidate token, with which it will be matched on further processing. In the above give example, word “hensy” is identified as an untagged word. This will be checked manually and after above mentioned process will be made part of the corpus as new word. Universal tagged token will be like:

bat <EW>, sun <EW>, meri <UT>, pehly <UT>, jo <UT>, hensy, hai<UT>, na <UT>, wahi <UT>, code <EW>, hai <UT>, mere <UT>, bhai<NW>, lol<NLW>

All the Universal tagged words will be collected into a set. Frequencies of each Universal tagged in the set is calculated and then set to another list. The total number of the Universal tagged words is calculated. By using edit distance technique, distances of each Universal tagged word from all the remaining words are calculated and listed along in a separate list. This will help in identifying, for a single token that how many different tokens are near to that specific token. For the best of the results, the max distance is taken as 4. Hence for every token, a list will be generated, with its all nearest tokens listed in that along with their distances.

Then the user is asked to provide the value for the K. The value of K defines the number of the clusters, which will be formed through K-Clustering. Every cluster will have the most frequent token as its center and all the nearest tokens marked as candidate tokens surrounding it. Here, it is quite possible to get very unrealistic results like, many of the words will have very less edit distances and will fall quite near to the token under observation but, they themselves will be valid tokens and cannot be considered as candidate tokens. Hence a human expert opinion will be required to scrutinize the tokens for correct selection of the tokens for the lexicon and for the candidate list. This will help in the identification of the most frequently used token and all of its candidate tokens. The most frequent tokens will be part of the Refined Corpus and the candidate tokens will be made part of the separate list for future processing.

5.6 Pseudo Code Description

All master-tagged words are collected into a set “roman_set”. Frequencies of each non-tagged word in the “roman_set” is calculated and then set to “roman-frequency”. The total number of non-tagged words in the “roman_set” is calculated and assigned to variable “n”. A matrix “ed_matrix” is created. A nested loop is set to calculate the “edit-distance” between each non-tagged word in “roman-list”. Every word is compared with all the entries of the roman-list to find its distance which is then submitted to “ed_matrix”. User is asked to define the number of clusters by giving a value to “k” which will be used in the “K-means algorithm”. K_clusters is a compound set referring to a plot, where different K no. of clusters is located. K_means_cluster is a function of k-means clustering which clusters ed_matrix into K no of clusters. An empty set “t” is created. For each single cluster “c” in “k” no. of cluster, “w” represents the highest frequency string, putting all the candidate strings denoted by “c-w”

Frequently used Tokens	Candidates Tokens	Frequency
Abi	Abi	121
	Abhi	116
Kya	Kya	377
	Kia	367

Table 2. Tokens with nearest frequencies

are collected in the “t”. It is necessary to narrate that for majority of the tokens, there were more than one candidate. For example the most frequently string used was “hai” which had several candidate Romanized strings like ha, haii, haye etc. The final selection of a string to be made part of the refined corpus was purely based on the string-frequency. Many to many or almost similar words were manually selected as the correct strings.

5. Results

We collected a raw corpus containing 103,566 words through different sources both online (Emails, SNS feeds, tweets etc) and offline (SMS, MMS, Text files). Out of these 71,437 words were used as the training corpus and the remaining 32,129 words were reserved as test corpus. The test corpus was again bifurcated into 4 equal groups of 8032 words. For checking, first part of the test corpus was passed through the model and 988 new words were detected. In the second pass 112 new words were identified by the system. In the third pass, 53 new words were detected and made the part of reference corpus. In the last run 11 new words were detected as new words. The average induction after first four test runs was calculated to be 293.5 words per run.

Frequently used Tokens	Candidates Tokens
hai	hai
	ha
	haii
	haiii
	haiiii
	hhai
nhi	nhi
	nhii
	naii
	nai
	nahi
	nae
	nei
	nain

Table 3. Frequently used tokens with candidates

The most widely used string in the Romanized Urdu communication is “hai” with 11,863 occurrences. 88% of the total i.e. 91138 strings had multiple candidate tokens which are not included in the most frequent strings for their respective group. Therefore, we can deduce that the words with highest frequency have more candidate tokens. Rarely there were strings on which there was complete consensus, world like Allah, Kal, Baat etc. had no candidate strings. The highest percentage of candidate strings was for the Nhi with more than 63% wrong string recorded.

6. Error Analysis

Urdu is written in Perso-Arabic script. Developing its corpus in Romanized form presents many key challenges which result in large number of errors. As discussed before in the example, words like “sun” and “bat” will always be tagged as English words instead of legitimate Urdu words. Also, it is really difficult to establish a clear pattern for the words with more vowels because they are used in many different manners even by the same user. Another major error, which was encountered frequently, was related to nouns. Manual supervision would be required to clearly identify the type of noun when a new token word is identified.

7. Future Work

All languages evolve with time. Hence no matter how comprehensive the work or research may be, there is always tendency to be done more. Our corpus is in continuously growing for broader language coverage and development of corpus and other resources required for the computational systems of Romanized Urdu. Presently, this research is limited by demographic and geographic constraints related to selection of the participants. In future, we wish to broaden its scope and size.

Ambiguity resolution between the Romanized Urdu tokens and English words is a complex problem. A possible solution of this problem may be incorporating the context of the tokens. However, context analysis is beyond the scope of this research.

References

- [1] Lewis, P. M., Simons, G. F., Fennig, C. D. (2013). *Ethnologue: Languages of the World*, Seventeenth edition. Dallas, Texas: SIL International.
- [2] Resnik P., Smith. N. (2003). The web as a parallel corpus. *Computational Linguistics* 29 (3) 349-380.
- [3] Habib, A., Iwatate, M., Asahara M., Matsumoto, Y. (2012). Keypad for large letter-set languages and small touch-screen devices (case study: Urdu), *International Journal of Computer Science* 9 (3).
- [4] Habib, A., Iwatate, M., Asahara, M., Matsumoto Y., Khalil, W. (2013). Optimized and Hygienic Touch Screen Keyboard for Large Letter Set Languages, *In: Proceedings of Seventh ACM International Conference on Ubiquitous Information Management and Communication (ICUIMC)*, 2013, Kota Kinabalu, Malaysia.
- [5] Irvine, Ann, Weese, J., Chris, Callison-Burch. (2012). Processing informal, romanized pakistani text messages, *In: Proceedings of the Second Workshop on Language in Social Media*. ACL, 2012.
- [6] Malik, A. A., Habib, A. (2013). Urdu to English Machine Translation Using Bilingual Evaluation Understudy, *International Journal of Computer Applications*, 82 (7) 5-12.
- [7] Ukkonen, Esko. (1985). Algorithms for approximate string matching. *Information and Control* 64 (1-3) 100-118.
- [8] Habib, A., M., Asahara M., Matsumoto, Y., (2011). Different input systems for different devices: Optimized touch-screen keypad designs for Urdu scripts, *In: Proceedings of Workshop on Text Input Methods WTIM2011, IJCNLP*, Chiang Mai, Thailand.
- [9] Navarro, Gonzalo. (2001). A guided tour to approximate string matching. *ACM computing surveys (CSUR)* 33 (1) 31-88.
- [10] Gul, A., Habib, A., Ashraf. J. (2016). Identification and extraction of Compose-Time Anomalies in Million Words Raw Urdu Corpus and Their Proposed Solutions, *In: Proceedings of the 3rd International Multidisciplinary Research Conference (IMRC)*, Peshawar, Pakistan.