Vol.2011-NL-204 No.4
2011/11/21

Keypads for Large Letter-Set Languages and Small Touch-Screen Devices

ASAD HABIB　MASAKAZU IWATATE
MASAYUKI ASAHARA　YUJI MATSUMOTO
【奈良先端科学技術大学院大学情報科学研究科自然言語処理学講座」 】

Composing Urdu is a thorny task on touch-screen devices. We proposed novel keypads optimized for accurate, easy and speedy typing. We applied automated procedures for performance comparison of contemporary keypads. Besides Urdu, our proposed keypads are applicable and extendable to other Perso-Arabic script languages e.g. Arabic, Persian, Punjabi etc.

1. Introduction

Composing Urdu on modern touch screen gadgets is a thorny job. Many modern gadgets either lack a good interface for typing Urdu or provide sluggish, inconvenient and hard to use keypads. There is no widely used keyboard or IME (Input Method Environment) for Urdu. We live in the age of touch screen gadgets. The future trends also show promising growth for them. Currently available input systems developed for standard PCs have room for improvement in efficiency, visibility and usability etc. particularly for Perso-Arabic script languages e.g., Urdu. Although it is spoken by a large population, the presence of Urdu is quite limited on the WWW. Among others, one reason is the difficulty of composing Urdu on modern computers particularly the touch screen handheld gadgets.

Furthermore small touch screen devices bring about health hazards to the user i.e. they put more strain on eyes due to poor visibility. We put forth hygiene in prime focus. Thus we designed distinct touch screen keypads that are clearly visible hence putting no strain on eyes of the user. At the same time, our designs facilitate fast, correct and easy Urdu composing. We performed automated experiments on Urdu corpus that show considerable improvement over contemporary keypads. Experiments have been discussed in section 3. Our optimization technique for arrangement of alphabets and unique interface for data input is extendable and equally applicable to other natural languages.

(1) Urdu

Urdu is the 2nd largest Arabic script language according to the number of speakers (Lewis and M. Paul, 2009; Weber 1999). Designing optimized Urdu keypads for small screen widgets is a knotty problem since Urdu uses a relatively large letter set. In addition, Ligatures and Diacritics are also borrowed from Arabic in Urdu but their unigram frequencies are very low. Ligatures are fixed blocks of letters each represented by a Unicode. Therefore we allocated them a single button on this keyboard layout. Diacritics form another set of low frequency characters. They are small macrons like characters normally used to show the correct pronunciation of letters in a word. Both the Ligatures and Diacritics are used mostly in religious texts that have become part of Urdu from Arabic and Persian.

Using the unigram and bigram frequencies in a large corpus, we developed novel Urdu touch screen keypads as shown in Figures 3, 5 and 6. The results of our experiments revealed ample significance that is explained in section 3.1.

At present, more and more data is being generated and uploaded using touch screen smart gadgets. These gadgets come in various shapes and screen sizes such as tablet PCs and mobile phones etc. Recently, there have been zero button touch screen laptop systems in the market

e.g., the Acer ICONIA. The current trends and types of new gadgets being introduced in the market suggest the growth of touch screen systems in the days to come.

Different interfaces suit different devices for users who need to input data in different natural languages. Full keyboard replica designs with base and shift versions e.g., QWERTY and Dvorak etc. cause usability as well as visibility problems hence not viable for small touch screen systems. Besides, small screen devices bring about health hazards to the user. Eyesight weakness, RSI (Repetitive Strain Injuries) and CTS (Carpal Tunnel Syndrome) etc. are only a few health hazards caused by technology/devices that we use in our daily life. For example, in case of eyesight, the closer objects put greater strain on the muscles converging the eyes retina (Ankrum, 1996). Stress on convergence system of eyes is crucial factor for strain (Jaschinski-Kruza, 1988; NASA, 1995) Thus we need to keep hygiene in prime focus during design and development of input systems, particularly for small touch screen devices.

Thus we tried to develop touch screen keypads that would be health friendly having much visibility and usability coupled with crafty arrangement of keys that is ideal for fast, correct, easy and efficient composing. Our optimization technique for arrangement of alphabets and unique interface for data input will be extendable and equally applicable to other natural languages and various sizes of touch screen devices.

(2) **Size of Urdu**

Urdu is the national language of Pakistan. It is the Lingua franca of Indo-Pak subcontinent and widely spoken across a score of countries e.g. UAE (United Arab Emirates), UK (United Kingdom), Afghanistan and Fiji etc. Phonetically, it is quite similar to Hindi. Written Urdu and Hindi use different and mutually exclusive scripts. However, in spoken they appear to be the same language. Rai and Alok (2000) stated, "One man's Hindi is another man's Urdu". Hindi is written in Devanagri script while Urdu is written in Perso-Arabic script. Ethnologue (Lewis and M. Paul, 2009) considered Urdu and Hindi as the same language and ranked it the 5th largest language of the world according to the number of speakers. The numbers of Urdu and Hindi speakers are given by table 1 (Malik et. al. 2009).

Table 1 Hindi and Urdu Speakers

	Native Speakers	**2nd Language Speakers**	**Total**
Hindi	366,000,000	487,000,000	853,000,000
Urdu	60,290,000	104,000,000	164,290,000
Total	426,290,000	591,000,000	**1,017,000,000**

(3) **Urdu Script**

Urdu is written from right to left. Arabic has 28 base letters while Persian has 32 letters. Both Arabic and Persian letter-sets are subsets of Urdu. However, the exact number of Urdu letters is not agreed upon. Various articles report different numbers of letters (Ijaz and Hussain, 2007; Malik et al. 1997; Habib et al. 2010). The largest letter set contains 58 letters (NLA Pak). It is shown in figure 1.

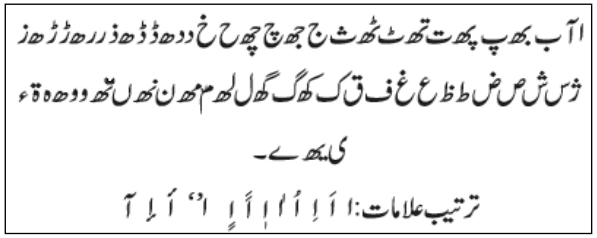

Figure 1 58 character set for Urdu alphabet

Hussain, (2004) reported 41 letters in Urdu letter-set. Ijaz and Hussain, (2007) mentioned 56 letters. Habib et al. reduced the Urdu letter set to 38 that is shown in table 2.

Table 2 Basic 38 Urdu alphabets and corresponding PEC

چ	ج	ث	ٹ	ت	پ	ب	آ	ا
C	**j**	**S**	**t**	**T**	**p**	**b**	**~**	**a,e**
ژ	ز	ڑ	ر	ذ	ڈ	د	خ	ح
J	**z**	**R**	**r**	**Z**	**d**	**D**	**K**	**H**
ف	غ	ع	ظ	ط	ض	ص	ش	س
f	**G**	**3**	**zx**	**Tx**	**Zx**	**Sx**	**sx**	**s**
ء	ہ،ھ	و	ں،ن	م	ل	گ	ک	ق
'	**h**	**o,v,w**	**n, N**	**m**	**l**	**g**	**k**	**q**
							ے	ی
							Y	**y**

Urdu has no distinct upper and lower case letter forms. However the above Romanization

scheme is case-sensitive (Roman letters only) that helps in distinguishing the correct Urdu pronunciation. Each Urdu letter is mentioned along its respective letter used for Romanization. Lower-case Roman letters represent the pronunciations exactly similar to their respective pronunciations in English. Upper-case letters represent similar but non-equal English pronunciation for the same letter. Based on the above letter-set, we introduce novel keypads for touch screen mobile phones, tablet PCs and full scale touch screen computers.

(4) **Shape of Urdu letters**

Urdu letters change their shape based on their respective positions inside a word. A letter can have up to four different shapes i.e. base, initial, middle and final shape.

A letter is in its base shape when it appears alone as a disjoint letter. Initial shape refers to the shape of a letter when it appears in the beginning of a ligature. Middle shape of a letter is written when it is joined by both the preceding and the following letters inside ligature. Final shape appears when a letter marks the end of a word or ligature. Durrani and Hussain (2010) discussed this property of Urdu letters in much detail.

(5) **Character-Level NLP Applications**

NLP has numerous applications at the "characters level" as shown by Figure 2. These include Romanization, Transliteration, Script Generation, Input System and/or Interface Designs etc. This research targets on the Interface Designs. We have come up with novel keyboard and keypads for text input on various types of touch screen devices such as mobile phones, tablet PCs and completely touch screen PCs.

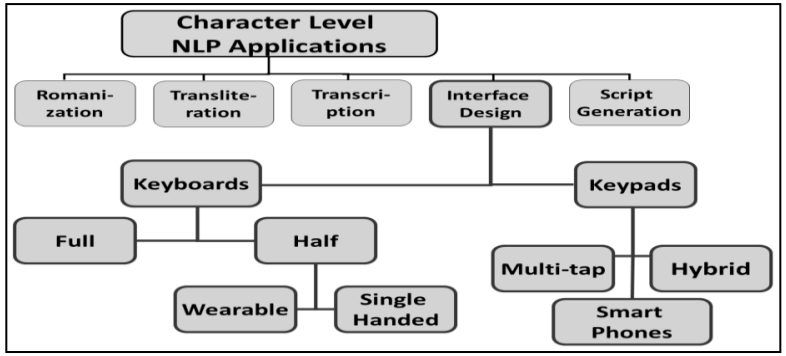

Figure 2 Character Level Applications of NLP.

2. Proposed Keypads

Apart from the conventional QWERTY and Dvorak keyboards, there are a number of key-pads used for text entry e.g., Muti-tap, odometer-like, touch-and-flick, Septambic keyer and Twiddler etc. (Wigdor, 2004).

Existing on-screen Urdu keyboard is replica of Microsoft Windows QWERTY type key-board. For Mobile phones, Multi-tap T9 replica keypads are in use. The working of existing Urdu Multi-tap keypad is explained in the Table 3. The columns show the characters that will be typed when the corresponding key (numeral in row header) is tapped/pressed a specified number of times.

Table 3 Multi-tap input table for T9 keypads

Key	I	II	III	IV	V	VI	VII
2	ب	پ	ت	ۃ	ٹ	ث	
3	ا	آ	ؤ	ۂ	ء	ئ	
4	س	ش	ص	ض			
5	د	ڈ	ز	ر	ڑ	ذ	ژ
6	ج	چ	ح	خ	ہ		
7	ن	و	ھ	ی	ے	ۓ	
8	ف	ق	ک	گ	ل	م	ں
9	ط	ظ	ع	غ			

.Full sized QWERTY like keyboards are not feasible for touch screen devices, in particular devices with small screen where limited screen area needs to be used astutely. This issue becomes more challenging when we design keypads for languages with a large number of alphabets. The trade-off issues in size and position of keyboard, editor, and buttons etc. require great care at design time. A good design must comply with the five principles of Ergonomics; safety, comfort, ease of use, productivity/performance and aesthetics (Karwowski, 2006).

Keeping the above points in view, we pro-pose the following two keypads for small size touch screen devices.

(1) **Keypad for small size touch screen devices (Smart phones)**

Figure 3 shows the base image of proposed frequency-based keypad for touch screen mobile phones. The individual characters are selected based on their unigram frequencies in

55-million characters corpus. The arrangement of characters is done on the basis of their corresponding bigram neighborhood frequencies. The letters in the base version, as shown in Figure 3, are not arranged in alphabetical order in Urdu. For the sake of easy understanding, easy memorizing and better visibility, all the remaining Urdu letters are shown in small font on the corresponding edges of each button. The leftmost button on lower row will be used for changing the input language, writing Ligatures, numeric characters, special characters and Diacritics etc.

Comparison statistics of various keypads have been tabulated in section 3. The base form of keypad shows the most frequently used Urdu letters. The bigram neighborhood statistics reveal that this non-alphabetic arrangement of Urdu letters alone gives additional 17% improvement in composing Urdu text. Other statistics related to comparison of different keypads are explained in section 3.1.

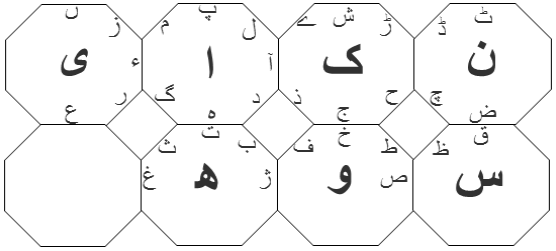

Figure 3 Proposed keypad for touch screen mobile phone.

In the event of a “button press” a single button can expand up to 8 neighbors showing the 8 new letters on a separate layer. These 8 letters consist of 4 horizontal neighbors and 4 diagonal neighbors. Beginners will need to look at the screen to select the correct neighboring letter. However experienced users can “touch type” in order to type their desired letter(s). The term “touch type” is sometimes referred to as “blind touch” also. The individual button sizes are big enough for blind touch and/or thumb typing. The size of buttons and their dimensions are flexible and can be adjusted according to the device on which the keypad is required to be deployed. A technique called onion skinning is used to show the new layer on top of the base layer. The diagonal and horizontal neighbors appear on a new layer on top of the base layer. For better visibility and aesthetics, only horizontal neighbors will be displayed on screen. However the diagonal neighbors can be used by a user just like the horizontal neighbors with practice.

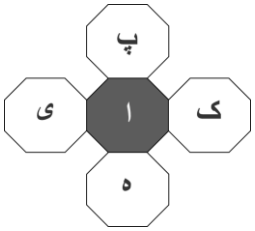

Figure 4 Illustration of a button press event (Horizontal Neighbors)

(2) **Keypad for middle size touch screen devices (Tablet PCs)**

Urdu letters can be grouped based on their shapes and their alphabetical order can still be preserved. The similar shaped letters have been grouped on a single button in our proposed keypad for Tablet PCs as shown in Figure 5.

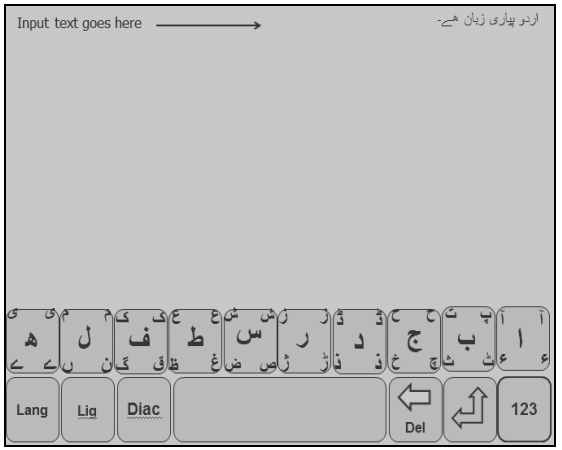

Figure 5 Proposed keypad for Tablet PCs

There are 10 buttons for typing Urdu that show the corresponding letters in native alphabetical order with some letters shown on the edges of buttons. All the letter typing buttons are shown on a single row called the home row. Unlike hardware keyboards, it is very difficult to return fingers to exactly the same position on a touch screen keypad. Thus we arranged all the letters on a single row so that the user doesn’t need to lift the entire hand in

order to type a letter. The user will keep both hands all the time above the single/home row. The user just needs to touch and flick in order to type a certain letter. The little finger of right hand will type the rightmost button on the keypad while the little finger of the left hand will be used to type the leftmost button on the keypad. The four middle buttons will be typed using the index fingers of both hands. The reason for this is that the index fingers are the strongest typing fingers (Krestensson, 2009).
The lower row includes some special buttons such as Lig and Diac that stand for "Ligatures and Diacritics respectively.

(3) **Keypad for large size touch screen devices (PCs)**
Figure 6 shows our frequency based full keyboard layout. The current layout for Urdu keyboard is a replica of Microsoft Windows OSK (On-Screen Keyboard). This keyboard has room for improvement in that some high frequency letters are typed in combination of Shift-key, the last thing a user will need. Similarly, the buttons arrangement is not frequency based. We propose the frequency based full keyboard layout as shown in Figure 6. Its detailed performance examination will be done in near future. However the new layout has eliminated the Shift version of Microsoft Windows replica. We also re-arranged the position of keys based on the frequencies of individual letters such that the most frequent letters should be typed by the strongest typing finger i.e. the index finger.
Additional issues related only to the touch screen keyboards such as the inter-keys distance will also be investigated and proper accommodating solutions would be put forward. Similarly the neighborhood of some standard keys might also be required to change. One such example is the neighboring keys of the "Backspace/Delete" and the "Enter/Return" keys.
In practice it is faster to type on touch screen than on multi-tap systems. Research that studies comparing the performance of touch screen and multi-tap systems could not be found. Thus for this study, we assumed "a touch" equal to the "a tap".

Figure 6 Urdu letters with their corresponding positions on QWERTY keyboard

3. Experiment

We carried out experiments on a general genre corpus of size 15,594,403 words. Existing Touch screen systems start word prediction as soon as the user types the first letter. For words with length up to two letters, this seems to bring hardly any improvement to the typing speed. On the contrary, it makes the system more complex and larger in size putting more load on CPU. We recommend that word prediction should start after the second letter has been typed by the user. Out of 15,594,403 words, 4,784,234 words are less than or equal to two letters in length. Hence for the experiments of this study, we used a reduced corpus of size 10,810,169 words.

3.1 Comparison

We compared the performance of proposed keypads with the existing counterparts. The reduced corpus size and assumption of "touch=tap" put the bias in favor of the existing systems. However, we still achieved results that show substantial improvement over the existing systems. The comparison of time required to type the corpus using existing Multi-tap and our proposed keypads are tabulated in Table 4.

Table 4 Comparison of time required to type the corpus

	Multi-tap (existing)	Touch Screen	Tablet PC
Seconds	263,380,598	135,249,436	120,096,926
Days	**3048.4**	**1565.4**	**1390**

The comparison of the number of taps/touches has been summarized in Table 5.

Table 5 Comparison of number of taps/touches required to type the corpus

	Multi-tap (existing)	Touch Screen	Tablet PC
	170,580,560	80,818,830	73,242,564
Improve-ment		**52.62%**	**57.06%**

Table 6 shows the comparison on keypad sizes between the existing and proposed keypad layouts.

Table 6 Comparison of Keypad sizes

	Multi-tap (existing)	Touch Screen	Tablet PC
	154	83	80
Improve-ment		**46.10%**	**48.05%**

4. Conclusion

We proposed different types of keyboard and keypads for different types of touch screen devices. The comparison analysis shows promising results. In addition to great amount of improvement over existing keypads, our proposed designs are flexible because the size and dimensions of keypads, buttons, and editors can be adjusted according to the device on which the keypad is deployed. Similarly our keypads offer greater usability because Urdu letters include all the letters of Arabic and Persian. Hence these layouts are equally usable by the Arabic and Persian users. The keypads are optimized for Urdu though. With minor additions, our input systems are extendible to other Perso-Arabic languages as well.

5. Future directions

We intend to carry out thorough testing of our keypads by human subjects. Additionally, we want to extend our keypads to include other Perso-Arabic languages such as Punjabi, Pashto, Dari and Potohari etc. Another possibility to exploit this study can be in the design of single finger operated keypad and single hand operated keyboard for touch screen devices.

References

1) Asad Habib, Masayuki Asahara, Yuji Matsumoto and Kohei Ozaki. 2010. JaPak IEOU: Japan-Pakistan`s Input English Output Urdu (A Case Sensitive Proposed Standard Input System for Perso-Arabic Script clients). In prodeedings of ICIET. Karachi.
2) Mark D. Dunlop and Finbarr Taylor. 2009. Tactile Feedback for Predictive Text Entry. In proceedings of Conference on Human Factors in Computing Systems Boston. MA. USA.
3) Malik, L. Besacier, C. Boitet and P. Bhattacharyya. 2009. A hybrid Model for Urdu Hindi Tranliteration", In proceedings of ACL-IJCNLP, Suntec, Singapore.
4) M.Ijaz and S.Hussain. 2007. Corpus Based Urdu Lexicon Development. In proceedings of CLT07. Pakistan.
5) Ankrum, D.R., (1996). Viewing Distance at Computer Workstations, Workplace Ergonomics, September. pp. 10-13.
6) Jaschinski-Kruza, W. 1988. Visual strain during VDU work: the effect of viewing distance and dark focus. Ergonomics, 31, 10, 1449 - 1465
7) M.Amir K., M.A. Khan. A.Habib and M.N. Ali. 2009. Corps based mapping of Urdu characters for cell phones. CLT 09. Lahore.
8) Lewis, M. Paul (ed.), 2009. Ethnologue: Languages of the World, Sixteenth edition. Dallas, Tex.: SIL International. Online version: http://www.ethnologue.org/ethno_docs/distribution.asp?by=size
9) NASA, 1995, NASA-STD-3000, Man Systems Integration Standards. National Aeronautics and Space Administration: Houston
10) Leonard J. West. 1998. The Standard and Dvorak Keyboards Revisited: Direct Measures of Speed. Technical report, Santa Fe Institute
11) Daniel J. Wigdor. 2004. Chording and Tilting For Rapid, Unambiguous Text Entry to Mobile Phones. University of Toronto.
12) W Karwowski. 2006. Handbook of human factors and ergonomics, Chapter-1. Wiley Online Library.
13) Government of Pakistan, National Language Authority (Cabinet Division), Islamabad, Pakistan http://www.nla.gov.pk
14) Unicode. 1991-2001. Unicode Standard version. Online version: http://unicode.org/charts/PDF/U0600.pdf
15) P. O. Krestensson. Winter 2009. Five Challenges for Intelligent Text Entry Methods. In proceedings of Association for the Advancement of Artificial Intelligence.
16) S. Hussain, "Letter-to-Sound Conversion for Urdu Text-to-Speech System", CRULP Lahore, Pakistan.
17) Lewis, M. Paul (ed.), 2009. Ethnologue: Languages of the World, Sixteenth edition. Dallas, Tex.: SIL International. Online version: http://www.ethnologue.org/ethno_docs/distribution.asp?by=size
18) http://www.andaman.org/BOOK/reprints/weber/rep-weber.htm.
19) Rai, Alok. 2000. Hindi Nationalism. Orient Longman Private Limited, New Delhi.

20) K. Knight and J. Graehl. 1998. Machine Transliteration. University of Southern California.
21) George Weber. 1999. The World's 10 most influential Languages. American Association of Teachers of French (ATTF), National Bulletin, vol. 24, 3:22-28
22) N. Durrani and S. Hussain, 2010. Urdu Word Segmentation, Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics.