A Hindi to Urdu Transliteration System

Gurpreet Singh Lehal
Department of Computer Science,
Punjabi University,
Patiala.
gslehal@gmail.com

Tejinder Singh Saini
ACTDPL
Punjabi University,
Patiala
tej74i@gmail.com

Abstract

In this paper, we present a high accuracy Hindi to Urdu transliteration system. Hindi and Urdu are variants of the same language, but while Hindi is written in the Devanagari script from left to right, Urdu is written in a script derived from a Persian modification of Arabic script written from right to left. To break this script barrier a Hindi-Urdu transliteration system has been developed. We have tried to overcome the shortcomings of the existing Hindi to Urdu transliteration systems and developed a system which can transliterate any Hindi Unicode text to Urdu at 99.46% accuracy at word level.

1 Introduction

Hindi and Urdu are mutually comprehensible languages written in mutually incomprehensible scripts and spoken by more than 600 million people in India and Pakistan. Hindi is written in the Devanagari script from left to right, while Urdu is written in a script derived from a Persian modification of Arabic script written from right to left. Over the time, with the influence of Persian in Urdu and Sanskrit in Hindi, the vocabularies of the two languages have also become different though they still share more lot of common words. In addition the grammar of the two languages is still same.

To break this script barrier, we have developed a transliteration system for transliterating Hindi text written in Devanagari script to Urdu. The Hindi to Urdu transliteration is different from standard transliteration schemes in the sense that the Hindi word has to be converted to Urdu with exact spellings and word boundaries. This fact has largely been ignored by existing systems, which have treated the transliteration problem as normal transliteration and used character mappings and dependency rules to convert Hindi word to Urdu without any consideration to spellings and Urdu compound words (Bushra and Tafseer 2009; Malik et al. 2008).

In this paper we present a fairly good accuracy Hindi-Urdu transliteration system, where we have taken special care to retain the correct spellings in the transliterated Urdu text.

2 Overview of Hindi and Urdu Scripts

Hindi language is written in Devanagari script. There are thirty three basic consonants and 7 other consonants that are formed with a dot diacritic on basic consonants form. In addition to this Devanagari script has 11 vowel characters, 10 vowel symbols, 2 symbols for nasalized sound and four consonant conjuncts.

Urdu script has 35 simple consonants, 15 aspirated consonants, one character for nasal sound and 15 diacritical marks (Bushra and Tafseer 2009). Urdu characters change their shapes depending upon neighboring context. But generally they acquire one of these four shapes, namely isolated, initial, medial and final. Urdu characters can be divided into two groups, non-joiners and joiners. The non-joiners can acquire only isolated and final shape and do not join with the next character. On contrary joiners can acquire all the four shapes and get merged with the following character.

3 Hindi to Urdu Character Mapping

Transliteration is a process wherein an input string in some alphabet is converted to a string in another alphabet, usually based on the phonetics of the original word. During character mapping it is observed that in many cases corresponding to one Hindi character there are multiple Urdu

Proceedings of ICON-2010: 8th International Conference on Natural Language Processing
Macmillan Publishers, India. Also accessible from http://ltrc.iiit.ac.in/proceedings/ICON-2010

characters having the same or similar sound. As shown in Table 1 ten Hindi characters have more than one mapping into Urdu script and the maximum number of multiple mappings are four.

SN	Hindi	Urdu Mapping Character			
		Map1	Map2	Map3	Map4
1	अ [ə]	ا	ع		
2	आ [ɑ]	آ			
3	इ [ɪ]	اِ			
4	ई [i]	ای			
5	उ [ʊ]	اُ			
6	ऊ [u]	اُو			
7	ए [e]	اے			
8	ऐ [æ]	اَے			
9	ओ [o]	اَو			
10	औ [ɔ]	اَو			
11	क [k]	ک	ق		
12	ख [kʰ]	کھ			
13	ग [g]	گ			
14	घ [gʰ]	گھ			
15	ङ [ŋə]	نگ			
16	च [ʧ]	چ			
17	छ [ʧʰ]	چھ			
18	ज [ʤ]	ج			
19	झ [ʤʰ]	جھ			
20	ञ [ɲə]	نج			
21	ट [ʈ]	ٹ			
22	ठ [ʈʰ]	ٹھ			
23	ड [ɖ]	ڈ			
24	ढ [ɖʰ]	ڈھ			
25	ण [ɳ]	ن			
26	त [t̪]	ت	ط	ۃ	
27	थ [t̪ʰ]	تھ			
28	द [d̪]	د			
29	ध [d̪ʰ]	دھ			
30	न [n]	ن	ں		
31	प [p]	پ			
32	फ [pʰ]	پھ			
33	ब [b]	ب			
34	भ [bʰ]	بھ			
35	म [m]	م			
36	य [j]	ے			
37	र [r]	ر			
38	ल [l]	ل			
39	ळ [ɭ]	ل			
40	व [v]	و			
41	श [ʃ]	ش			
42	स [s]	س	ص	ث	
43	ह [h]	ہ	ح		
44	ा [ə]	ا	ٰ	ہ	
45	ि [ɪ]	ِ			
46	ी [i]	ی			
47	ु [ʊ]	ُ			
48	ू [u]	ُو			
49	े [e]	ے			
59	ै [æ]	َے			
51	ो [o]	و			
52	ौ [ɔ]	َو			
53	ख़ [x]	خ			
54	ग़ [ɣ]	غ			
55	ज़ [z]	ز	ض	ظ	ذ
56	ड़ [ɽ]	ڑ			
57	फ़ [f]	ف			
58	conjunct	ّ			
59	ं[ɲ]	ں	ن		
60	ँ[ɲ]	ں	ن		

Table 1. Hindi to Urdu Character Mapping

To establish default mapping a statistical analysis of the Hindi corpus is performed to determine the percentage of times the Urdu character was mapped to the similar sounding Urdu character and the result is shown in Table 2. The Urdu characters with highest occurrence are selected for default mapping.

Hindi Character	Equivalent Urdu	Frequency	Default
ह[h]	ہ ح	85.88% 14.12%	ہ
स[s]	س ص ث	84.66% 12.46% 2.88%	س
क[k]	ک ق	86.99% 13.01%	ک
त[t̪]	ت ط ۃ	90.10% 9.87% 0.03%	ت
ज़[z]	ز ض	58.81% 16.31%	ز

	ظ ذ	14.10% 10.78%	
न [n]	ن ں	99.53% 00.47%	ن
अ [ə]	ا ع	91.82% 08.18%	ا

Table 2. Frequency of Occurrence of Multiple Mapping Characters

4 Issues in Hindi-Urdu Transliteration

The Hindi to Urdu transliteration is different from standard transliteration schemes and there are many challenges involved to develop a system with good accuracy. The main challenges are discussed as follows:

4.1 Multiple Mappings

As shown in Table 1, corresponding to one Hindi character there are multiple mapped characters present in Urdu script having the same or similar sound. This creates ambiguity at character level. For example, the default mappings for Hindi words सफ़ा and ग़लती will produce غلتی , سفا which are wrong transliterations. The correct transliterations of these words are صفا, غلطی where स[s] and त[t̪] characters are mapped to less frequent mappings. A statistical analysis of corpus reveals that these lesser frequently occurring similar sounding Urdu characters have 3.48% frequency of occurrence, which means that on an average 3.48% of characters will always be wrongly mapped by the normal rule based systems.

4.2 No exact equivalent mappings in Hindi for some Urdu characters

There are certain Urdu characters such as ع(ain), ٰ (khadi zabar) and ء (hamza) for which there are no exact equivalent Hindi characters. Though hamza can be generated using character combination rules, but there are no specific rules for ain.

4.3 Decrease in use of *Nukta* Symbols in Hindi Text

Hindi language has borrowed many words from Arabic, Persian etc. For proper pronunciation of these borrowed words, consonants with *nukta* (ख़[x], ग़ [ɣ], ज़ [z], फ़ [f]) are used. But over the years, the usage of these characters particularly, ख़ [x], ग़ [ɣ], फ़ [f] has been on decline as many Hindi speakers do not make a distinction between ख[k^h] ख़[x], ग[g] ग़[ɣ] and फ[p^h] फ़[f]. The result is that most of the words in Hindi are now written without nukta symbol. As a result, character to character based mapping produces wrong spellings. So फकीर will be transliterated to پھکیر in Urdu, which is wrong while the actual transliteration is فقیر, which is obtained if the correct Hindi spellings फ़कीर are used.

4.4 Difference between Pronunciation and Orthography

In certain cases, the Hindi words are written with short vowels, while they are pronounced with long vowels. The equivalent words in Urdu are also written with long vowels and so the rule based mapping system which converts those short vowels in Hindi Urdu give wrong results in such cases. Some examples of such words are गुरू, खुश, ख़ुराक and पति. They are written with short vowels in Hindi, while the corresponding words in Urdu are written with long vowels گورو، خوش، خوراک، پتی respectively.

4.5 Transliteration of Proper Nouns

The transliteration of Urdu proper nouns such as names of persons and places from Hindi to Urdu poses another challenge. Many times the spellings of such words in Urdu are typical and it is not possible to formulate transliteration rules for generation of such spellings. As for example consider the Hindi words अबदुल्ला, बुशरा, रहमान and हैदराबाद which are written in Urdu are , عبداللہ بُشریٰ, رحمٰن, حیدرآباد but which will yield wrong results if transliterated using the usual rule based mapping system.

4.6 Nasalized Sound Characters

There are two frequent nasalized sound characters "chandrabindu" ँ [ɲ] and "bindu" ं [ɲ] in Hindi Script. The mapped Urdu characters corresponding to these two characters are "noon" ن [n] and "noon-gunna" ں [ɲ]. But there are certain nasalized words in Hindi where this mapping fails and to handle such cases special rules need to be formulated. For example consider the words like लंबाई and संभाल the nasalized symbol should be mapped to م [m] as لمبائی and سمبھال respectively and not as نبائی and سنبھالل.

4.7 Urdu Compound Words

Urdu language has many compound words in its vocabulary (Bushra and Tafseer 2009; Raza et al. 2009). These words are based on multiple constituent words. In Urdu, the general rule of writing compound word is to put space among the constituents however Hindi writers usually don't follow this convention and they write it by connecting the constituent words together. The same rule is also applied on words from English vocabulary also. For example consider the words बातचीत, जानकारी, रिशतेदार, वैबसाइट, telephone etc. Secondly, there are some words begin with बे or some words ending with गा, गे or गी that need to be transliterated as separate multiple constituent words in Urdu like बेसहारा, बेशक, बेमिसाल or होगा, लगेंगे, सकेगी etc. These have to be transliterated as:

بے سہارا، بے شک، بے مسال، ہو گا، لگین گے، سکے گی

and not as:

بیسہارا، بیشک، بیمسال، ہوگا، لگینگے، سکیگی

These issues cannot be resolved by the existing rule based systems and other alternative techniques have to be devised to solve them.

5 Lexical Resources

In order to perform statistical analysis during the various phases of the purposed transliteration system we have been created various useful lexical resources. These lexical resources (Table 3) have been developed by analyzing both the Urdu and Hindi Corpora and these resources are integrated in to the current system.

Resource	Count
Urdu Word Frequency List	2,31,344 words
Hindi Word Frequency List	2,42,451 words
Urdu Word Bigram List	31,82,511 bigrams
Hindi-Urdu Lexicon	12,344 terms

Table 3. Lexical Resources

6 Hindi Urdu Transliteration System

The system architecture of the Hindi Urdu transliteration developed by us is shown in Figure 1. The system has been able to take care of most of the issues raised in the previous section. The complete Hindi-Urdu transliteration system is divided into three stages: pre-processing, processing and post-processing.

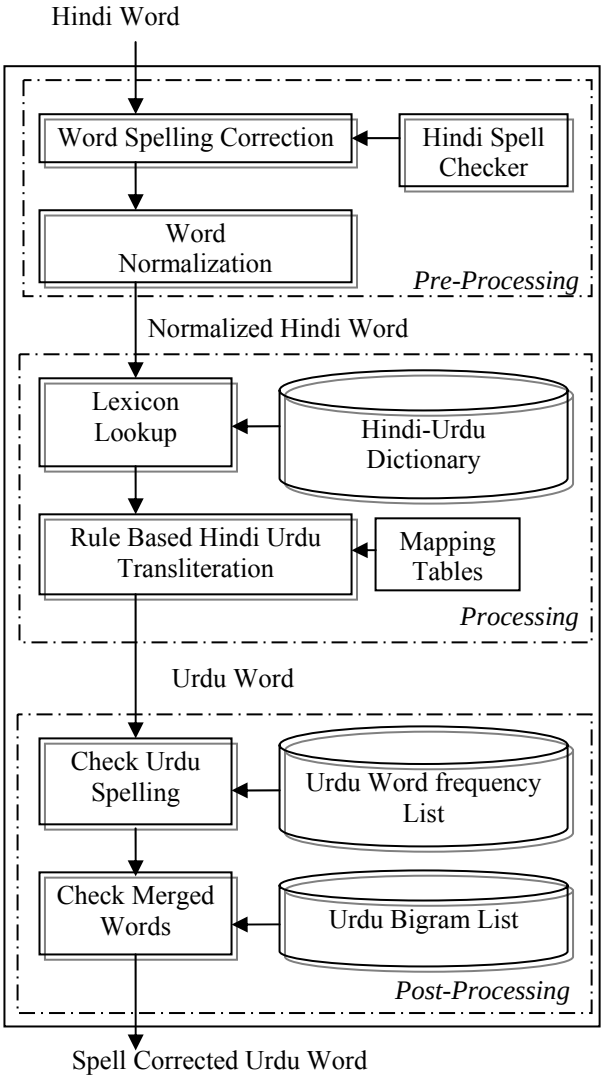

Figure1. System Architecture

In the pre-processing stage, the Hindi word is cleaned and prepared for transliteration by normalizing the Hindi word according to the Urdu spellings and pronunciation. In the processing stage, the normalized Hindi word is converted to Urdu using a hybrid rule based and lexicon based system. In the post-processing stage, the spellings of the transliterated Urdu word and word boundaries are corrected by using an Urdu corpus. The three stages are discussed in detail in the following sections:

6.1 Pre-Processing

In the pre-processing stage, the Hindi word is cleaned and prepared for transliteration by normalizing the Hindi word according to the Urdu spellings and pronunciation. The input to this stage is a Hindi word in Unicode format. As already discussed above, the Hindi word may be wrongly spelled because of the *nukta* related characters. These spelling errors may not appear

so serious for Hindi readers, but when transliterated as such they generate wrong Urdu spellings. To solve this problem, the Hindi word is sent to a Hindi spell checker, which automatically corrects the *nukta* related errors. Thus the word गजल will get converted to ग़ज़ल after being fed to the Hindi spell checker.

Besides correcting the Hindi *nukta* related spelling errors, some other word normalization operations are also performed. To handle the problem of variation in pronunciation and orthography of some Hindi words, we have created a database of all such words along with their inflections. The Hindi word is checked in this database and if found gets converted to the appropriate form.

For handling the mapping problem of Hindi nasalized sound characters in some cases, special rules have been formulated.

Another issue, which is handled in this stage, is the absence of half characters in Urdu. It is observed that many half characters in Hindi text can not be transliterated into Urdu as Urdu does not have half character while Hindi frequently uses half characters. Instead Urdu has "Shadda" ّ symbol to double the sound of a consonant after which it is placed. But it was found, that in contemporary Urdu writing "Shadda" is very rarely used. We have used the following rule for handling the Hindi half characters:

If the half character in Hindi is followed by its full form or its aspirated form then the half form is deleted from the word. For example, ढक्कन and मक्खन will be normalized as ढकन and मखन. Otherwise the half form is converted to full form. Thus the words क्या and वक्त will be normalized as कया and वकत.

At the end of this stage the Hindi word corrected for the spellings and normalised for the pronunciation based errors is generated.

6.2 Processing

In this stage the normalized word generated in pre-processing stage is converted to Urdu by using letter to letter conversion mapping rules using the mapping Table 1 and Table 2, as well some special rules. For multiple mapping, the default character, with highest frequency of occurrence as mentioned in Table 2 is selected. Though the above rules work fine for most of the common words, but they do not give proper results in case of zero or multiple mappings of a Hindi character to Urdu characters. These rules many times fail to transliterate Urdu proper nouns for example बुशरा (بشرٰی) and रहमान (رحمٰن). Sometimes the transliterated word may not adhere to the conventional Urdu word boundary as discussed in section 4.7 and may need to be split up.

The multiple mapping problem and word segmentation problems are handled in the post-processing stage, while for handling the zero mappings and transliterating proper nouns and other typical spellings, we create a separate database of Hindi-Urdu words.

The Hindi word to be transliterated is first searched in this database. If the word is found, then it is directly converted to Urdu else it is converted using the above rule based mapping. To fasten up the transliteration process, we performed statistical analysis on a Hindi corpus and the first most frequently occurring 5,000 words were extracted. These words were manually transliterated to Urdu and stored in the above Hindi-Urdu database.

6.3 Post-Processing

This stage is primarily used to correct the spellings and insert extra spaces in the Urdu words generated in the processing stage. The major source of spellings errors in the transliterated Urdu words is the multiple character mapping in Urdu. As for example the words सलाह, मज़बूत, किस्म and ख़ुसूसन will be transliterated as سلاہ ، مزبوت، کِسم and خسوسن while the actual spellings are صلاح, مضبوط, قسم and خصوصاً respectively. To automatically correct the spellings, we have used the Urdu corpus. We have generated unigram and bigram frequency lists from the corpus.

To solve the problem of multiple mapping, our system generates all possible words combinations formed by taking all the equivalent forms of similar sounding characters. These words are then searched in the Urdu corpus and the word with highest frequency of occurrence is finally selected. For example consider the Hindi word, सतह. The word gets transliterated to ستہ. Now the characters ط ,س and ح have 3, 2 and 2 equivalent forms, which results in 12 different word combinations:

سطح، صطح، ثطح، ستح، صتح، ثتح، سطہ، صطہ، ثطہ، صتہ، ستہ، ثتہ

On searching the corpus, the word سطح is found to be having highest occurrence and thus the word सतह is transliterated to سطح.

The last operation performed is to convert the Urdu words to the conventional way they are written in Urdu. As for example, in Table 4, the first column contains the Hindi words to be transliterated, the second column contains the words after transliteration and the third column represents the conventional way the word is written in Urdu.

Hindi Word	After Transliteration	Conventional way
ब्लैकमेल	بلیکمیل	بلیک میل
बिजलीघर	بجلیگھر	بجلی گھر
बातचीत	باتچیت	بات چیت
तुझमें	تجھمیں	تجھ میں
सबको	سبکو	سب کو
शानदार	شاندار	شان دار
रखकर	رکھکر	رکھ کر

Table 4. Hindi Word Transliteration into Urdu conventional way

We have developed the following algorithm for word boundary detection:

- Starting from the third character, we generate all possible bigrams from the Urdu word by splitting it into two parts, taking care that each part contains at least two characters. There are certain characters such as (ا ّ ً ؤ ئ ء and ں ۃ), which cannot come at the beginning of a word and if we encounter any such character in the word, we do not split at that point and move to the next character.
- Next we find the frequency of occurrence of each of bigrams obtained in the above step from the bigram frequency list. The bigram with highest frequency is selected and its frequency is compared with the frequency of original word. In case, the bigram frequency is higher, then the word is split into two, else it is retained as such.

7 Experimental Results

We have tested our system on 50 pages of Hindi Unicode text compiled from two news websites http://www.bbc.co.uk/hindi/ and http://www.webdunia.com. The transliterated test has been manually evaluated and found that the overall transliteration accuracy of our system is 99.46%. The word level transliteration accuracy of the system on BBC text was 99.30% and for WebDunia, it was 99.61%.

8 Conclusion

In this research paper we have presented a Hindi to Urdu transliteration system with high accuracy of 99.46% at word level. We have tried to overcome the shortcomings of the existing rule based Hindi to Urdu Transliteration systems. The various challenges such as multiple/zero character mappings, variations in pronunciations and orthography, transliteration of proper nouns, Urdu word boundary etc. have been handled by generating special rules and using various lexical resources such as Hindi spell checker, Urdu and Hindi word frequency lists, Urdu word bigram list, Hindi-Urdu lookup table etc.

Acknowledgement

The authors would like to acknowledge the support provided by ISIF grants for carrying out this research. The authors will also like to acknowledge the linguistic support provided by Mohammad Sadiq and Nadeem Ahmed.

References

Agha Ali Raza, Awais Athar and Sajid Nadeem. 2009. N-Gram Based Authorship Attribution in Urdu Poetry, *Proceedings of the Conference on Language & Technology*, 88-93.

Bushra J. Tafseer A. 2009. Hindi to Urdu Conversion: Beyond Simple Transliteration. *Proceedings of the Conference on Language & Technology*, Lahore 24-31.

Durrani N. 2007. Typology of Word and Automatic Word Segmentation in Urdu Text Corpus. *National University of Computer and Emerging Sciences,* Lahore, Pakistan.

Malik, M. G. Abbas, Boitet Christian, Bhattcharyya, Pushpak. 2008. Hindi Urdu Machine Transliteration using Finite-state Transducers, *Proceedings of COLING 2008*, Manchester, UK, 537-544.

http://www.crulp.org/software/langproc/h2utranslitera tor.html

http://www.puran.info/HUMT/index.html (Accessed on 8th Aug 2009)