AN EFFICIENT HINDI-URDU TRANSLITERATION SYSTEM

Saqib Maqbool[1], Nisar Ahmed[2], Gulshan Saleem[3], Muhammad Munwar Iqbal[4]
[1,2,4]Dpartment of CS & CE, University of Engineering and Technology, Lahore, Pakistan.
[3]National University of Science and Technology, Rawalpindi, Pakistan
[4]Dpartment of CS, University of Engineering and Technology, Taxila, Pakistan.
[1]saqib_maqbool2003@yahoo.com, [2]nisarahmedrana@yahoo.com, [3]gulshnsaleem26@gmail.com, [4]munwariq@gmail.com

ABSTRACT: *Urdu and Hindi languages share common phonology, morphology and grammar but are written in different scripts; Urdu is written in Persio-Arabic script while Hindi is written in Devnagri script. Transliteration system is used to transcribe a word written in one language/script to another language/script by conserving its articulation. Such system is required to handle out of vocabulary words in different domains of natural language processing (NLP). Such system can also be used for Inter-dialectal translation such as Urdu-Hindi. This paper presents the basic model for Urdu-Hindi transliteration system. A mapping table is made for rule-based transliteration from one language into another. This paper also addresses the importance of Urdu diacritic marks in Urdu-to-Hindi transliteration and solves an ambiguity that arises due to the multiple Urdu characters equivalent to one Hindi character. The system has undergone extensive experimentation and testing and has shown high accuracy of transliteration.*

Keywords*: Natural Language Processing (NLP), Urdu, Hindi,*

1. INTRODUCTION

Hindi and Urdu are official languages of India and Pakistan respectively. Hindi is spoken by around 853 million people and Urdu by around 164 million people [1]. Although native speakers of Urdu and Hindi can understand spoken version of these languages due to common phonology and morphology but it is difficult for each of them to read text written in another language. They also share common grammatical structure, pronouns, case markers, auxiliaries and verbs are also same. Both languages share large proportion of vocabulary inherited from Sanskrit, Arabic, Persian and other languages. The extent of vocabulary overlaps depends upon the nature of text origin. Literary work in Urdu inclines towards Arabic and Persian whereas Hindi towards Sanskrit. However, media, news, sports and politics inclines towards shared vocabulary [2].Nevertheless both languages are written in a different way, Urdu is written in Persio-Arabic script with Nastaliq calligraphy style of the right to left whereas Hindi is written in left to right Devnagri script.

For Hindi to Urdu and vice versa conversation, a language transliteration system with high accuracy can be used. There are various methods of language conversion which converts one language into another based on bilingual dictionaries. The problem arises when the system is unable to look for a word in dictionary such as noun or cultural specific words. Transliteration can serve the purpose of conversion of such words from one language script to another for proper articulation [1-7].

Transliteration is a process of transcription of a word written in one language/script into another language/script with proper articulation. These system basically maps lexicon of texts from one language to another language for facilitation[5, 6]. Such system is required to handle out of vocabulary words in different domains of natural language processing (NLP) such as Inter-dialectal translation (without lexical changes) and out of vocabulary handling.These systems are very useful to display information in multiple languages such as signboards, reservation charts and publication of proceedings in bilingual countries.

This paper focuses on the issues related to transliteration between two languages (Hindi and Urdu). Simple word to word mapping was not enough so some rules were introduced to transliterate text data. The solution of ambiguities for single Hindi character against multiple Urdu characters is also provided. The accuraccy of transliteration from Urdu-to-Hindi is also increased by the use of automatic diacritization algorithm proposed by Abbas [8].

2. URDU AND HINDI WRITING SYSTEMS

Urdu writing system contains 35 simple consonants, 15 aspirated consonants, 15 diacritical marks and 10 vowels. Vowels are written using ا, آ, و, ی , ◌ِ,◌ُ,◌ٗ characters [9]. Urdu writing system is context-sensitive. Characters written in Urdu change their shape based on context making it a complex writing system [7].In Urdu all 15 aspirated consonants are formed using simple consonants going to be aspirated and Doachasmi hay ھ. Hamza ئ is used between two vowels as in لمبائی. Zer-e-izafat ◌ِ is used to get the effect of joining two words and it shows possession of the second word for the previous one. Noon-gunna e.g. ں is used as a nasalization marker. Shadda e.g. ◌ّ is used to geminate a consonant.

All consonants in Hindi inherit [ə] sound. All the vowels in Hindi are attached to the top or bottom of the consonant or to an [ा] vowel sign attached to the right of the consonant, with the exception of the ि vowel sign which is attached on the left [10]. Hindi has 44 consonants; 29 non-aspirated and 15 aspirated consonants and 11 vowels [10]. 11 among 15 aspirated consonants are represented by separate characters e.g. थ [thh], झ [jh]. The other four aspirated consonants are formed by simple consonant and the conjunct form of Hay e.g. म्ह [mh] = म [m] + ◌्+ह [h].

3. MAPPING OF URDU, HINDI CHARACTERS

During transliteration, characters from one language are mapped to the equivalent characters in other language. Following tables show the character to character equivalence of Urdu and Hindi non-aspirated and aspirated consonants.

3.1 Mapping of Urdu, Hindi Non-Aspirated Consonants

Urdu has 35 and Hindi has 29 non-aspirated consonants. Table 1 shows mapping of these characters.

Table 1. Mapping of Non-aspirated Urdu, Hindi letters

Urdu Letters	Unicode Value	Hindi Letters	Unicode Value
ا(vowel)	627	अ	905
آ (vowel)	622	आ	906
ب	628	ब	92C
پ	67E	प	92A
ت	62A	त	924
ٹ	679	ट	91F
ث	62B	स	938
ج	62C	ज	91C
چ	686	च	91A
ح	62D	ह	939
خ	62E	ख	959
د	62F	द	926
ڈ	688	ड	921
ذ	630	ज	95B
ر	631	र	930
ڑ	691	ड़	95C
ز	632	ज	95B
ژ	698	ज	95B
س	633	स	938
ش	634	प	937
ص	635	स	938
ض	636	ज	95B
ط	637	त	924
ظ	638	ज	95B
ع	639	--	--
غ	63A	ग	95A
ف	641	फ	95E
ق	642	व	958
ک	6A9	व	915
گ	6AF	ग	917
ل	644	ल	932
م	645	म	92E
ن	646	न	928
و	648	व	935
ہ	6C1	ह	939
ی	6CC	य	92F
ۃ	629	त	924
ھ	6BE	ह	939
ں/nasal sound	6BA	◌	902

3.2 Mapping of Urdu, Hindi Aspirated Consonants

Urdu and Hindi each has 15 aspirated consonants. In Urdu aspirated consonant is made by joining a consonant and Doachashmi-hai e.g. ھ. In Hindi 11 aspirated consonants have their own shape but the other 4 are made by simple consonants and conjunct shape of ह. Table 2 shows the mapping of these aspirated consonants.

Table 2. Mapping of Aspirated Urdu, Hindi letters

Urdu Aspirated Letters	Hindi Aspirated Letters
بھ	भ
پھ	फ
تھ	थ
ٹھ	ठ
جھ	झ
رھ	ह़
ڑھ	ढ़
کھ	ख
گھ	घ
لھ	ल्ह
چھ	छ
دھ	ध
ڈھ	ढ
مھ	म्ह
نھ	न्ह

4. MAPPING OF VOWELS

This section describes the vowel mapping rules for words of each language. In Hindi each vowel has two forms. Independent form is used if vowel comes in the start of a word and dependent form is used for its medial form.

4.1 Mapping of Vowel Sound 'Ə' (Urdu Character ا, Hindi Character अ)

This vowel in Urdu is represented by ا and comes in the start of the word. It is mapped to the Hindi character अ. Reverse is true in Hindi to Urdu mapping.

Examples:

अनसान/ meaning human → انسان

अगला/ meaning next → اگلا

अज़ान/ meaning prayer call → اذان

4.2 Mapping of Vowel Sound ' ɑ' (Urdu Character آ, Hindi Character आand ◌ा)

This vowel is represented by آ in the start of the Urdu word. It is mapped with Hindi character आ. In middle of Urdu word it is represented by ا and mapped to Hindi vowel ◌ा. Reverse is true for transliterating from Hindi to Urdu.

Examples:

آدمی/ meaning man → आदमी

جانا/ meaning to go → जाना

अलाही/ meaning God → الٰہی

कबरा/ meaning bigger → کبریٰ

4.3 Mapping of Vowel Sound 'E' (Urdu Character ی+ ا, Hindi Character ए and ◌े)

This vowel sound in Urdu is represented by ی + ا in the start of the word and mapped to Hindi vowel ए. In the middle of Urdu word it is represented by ے+ ا sound and mapped to ◌े. Reverse is true for Hindi to Urdu transliteration with additional rule of mapping that if ◌े comes in end of Hindi word; it is mapped to ے in Urdu word.

Examples:

ایک / meaning one → एक

ایثار/ meaning → एसार

मेरा/ meaning mine → میرا

सारे/ meaning all → سارے

4.4 Mapping of Vowel Sound 'Æ' (Urdu Characters ے + ◌َ, Hindi Characters ऐ and ◌ै)

This is represented by ے + ◌َ in the start of the word and mapped to ऐ Hindi vowel.In the end of the word it is represented by ے. Reverse is true for Hindi to Urdu mapping.

Examples:

है/ meaning is → ہَے

میلا/meaning dirty → میلا

4.5 Mapping of Vowel Sound 'I' (Urdu Characters ◌ِ + ا, Hindi Characters इ and ◌ि)

In the start of an Urdu word it is represented by ◌ِ + ا and mapped to इ Hindi letter. In the middle of Urdu word it is

represented by ◌ि and mapped to ि. Reverse is also true for Hindi to Urdu mapping.

Examples:

اِختراع / meaning innovation ⟶ इख़तरअ

احترام/ meaning respect ⟶ इहतराम

4.6 Mapping of Vowel Sound 'I' (Urdu Characters ی + ◌ِ + ا, Hindi Characters ई and ी)

In the start of Urdu word it is represented by ی + ◌ِ + ا and mapped to ई Hindi character.In the middle of Urdu word it is represented by ی + ◌ِ and mapped to ी. ی in the end of word is mapped with ी.

Example:

अमीरी/ meaning wealthy ⟶ امیری

4.7 Mapping of Vowel Sound 'U' (Urdu Characters ◌ُ + ا, Hindi Characters उ and ◌ु)

In start of Urdu word it comes with ◌ُ + ا and mapped to उ. In all other cases, ◌ُ in Urdu is mapped with ◌ु in Hindi. Reverse is true for Hindi to Urdu mapping.

Example:

उधर/ meaning there ⟶ اُدھر

مُصیبت/ meaning anxiety ⟶ مُسیبت

4.8 Mapping of Vowel Sound 'U' (Urdu Characters و + ◌ُ + ا, Hindi Characters ऊ and ◌ू)

In start of Urdu word it comes with و + ◌ُ + ا and mapped to ऊ. In the middle of Urdu word it is represented by و + ◌ُ and mapped to Hindi vowel ◌ू. Vice versa is true from Hindi to Urdu mapping.

Examples:

مُوسیٰ/ meaning prophet Moses ⟶ मूसा

خربُوزہ/ meaning Melon ⟶ ख़रबूज़ा

4.9 Mapping of Vowel Sound 'O' (Urdu Characters و + ا, Hindi Characters ओ and ो)

Words starting with و + ا generate this vowel sound and are mapped to ओ Hindi letter otherwise و in middle of Urdu word is mapped to ◌ो Vowel symbol. Reverse mapping is performed in Hindi to Urdu transliteration.

Examples:

اوچھا/ meaning vulgar ⟶ ओछा

ہولی/ meaning ⟶ होली

4.10 Mapping of Vowel Sound 'ऑ' (Urdu Characters و + ◌ٓ +ا, Hindi Characters औ and ◌ौ)

و+ ◌ٓ +ا in the start of Urdu word generate this sound and these letters are mapped to Hindi letter औ. Similar sound is generated if و + ◌ٓ comes in the middle of the word. This is mapped to ◌ौ. A reverse process is true for Hindi to Urdu mapping.

4.11 Mapping of Nasalized Characters

In Urdu nasalization is achieved with this Noon-gunnah ں , In Hindi it is mapped to anusavar.

5. PROPOSED SOLUTION To SOME AMBIGUITIES IN TRANSLITERATION

This section deals with somewhat unresolved issues and their solutions are provided for enhanced accuracy. These issues are related to character ambuiguity, diacritic marks and vovel issue in Urdu to Hindi transliteration.

5.1 Ambiguities for Single Hindi Character Against Multiple Urdu Characters

From table (1) it can be seen that there are some characters in Urdu which has only one equivalent character in Hindi. These characters with their respective equivalents are

(1) س،ص، ث ⟶ स
(2) ت، ط، ة ⟶ त
(3) ض، ز، ظ ⟶ ज
(4) ح، ہ، ھ ⟶ ह
(5) ک،ق ⟶ व

The ambiguity for these characters can be understood in this way that if we want to transliterate this Hindi word सतह (meaning surface) into Urdu, which character will be placed for the same letter स in Hindi word ? The correct transliteration for each of these Hindi words is as follows

मसाल ⟶ مثال

सतह ⟶ سطح

The same ambiguity arises if we want to translate these Hindi words in Urdu मरेज़ (meaning patient) ज़कान (meaning charity) अनतज़ार (meaning to wait someone).

The correct transliteration of these words is as follows

मरेज़ ⟶ مریض

ज़कान ⟶ زکٰوۃ

अनतज़ार ⟶ انتظار

Similar examples can be given for Hindi words containing ह [h] and त [t].

5.2 Solution for Ambiguities for Single Hindi Character Against Multiple Urdu Characters

Different solutions are used to resolve this character based ambiguity.

1) One solution can be mapping the ambiguous Hindi character to an Urdu character, which has maximum occurrence in a given corpus. To test our Hindi to Urdu transliteration, a corpus of 25291 words was made by collecting Urdu text from BBC and Urdu digest websites. Frequency of multiple mapped Urdu characters was found in this corpus. The data is shown in table 3. The Hindi characters that have one-to-multiple mapping characters were mapped to their respective most frequent Urdu characters. But this was not accurate solution. Urdu words that use less frequent same voiced characters were transliterated wrongly by this technique.

Table 3. Frequency and count of multiple mapping Urdu characters

Hindi	Urdu	Frequency	Count	Default
स	س ص ث	84.42 % 13.82 % 1.76 %	3104 508 65	س
ज	ز ظ ض	68.11 % 17.34 % 14.55 %	660 168 141	ز
त	ت ط ة	88.26 % 11.73 % 0 %	3512 467 0	ت

व	ک ق	86.13 % 13.87 %	6303 1015	ک
ह	ہ ح	83.97 % 15.84 %	4181 798	ہ

2) Another solution is that we map ambiguous Hindi character in a word to its all possible equivalent Urdu characters and check the correctness of each transliterated word by looking it up in a lexicon. We finally get the correct word after looking up lexicon for each of these possible words. Table 4 shows the frequency based wrong transliterated Hindi word with its all possible words. These all possible words are looked up in a lexicon to find correct word.

Table 4.Frequency based wrong transliterated words with their correct words after looking up lexicon.

Hindi Word	Urdu Word Based on High Frequency Ambiguous Urdu Character	All possible Urdu Variants	Corrected Word
मसलहत	مسلہت	مسلہت, مسلہط, مسلحط,مسلحت مثلہت,مثلہط, مثلحط,مثلحت, مصلہت,مصلہط مصلحط,مصلحت	مصلحت
सतह	ستہ	ستہ,ستح, سطہ,سطح ثتہ,ثتح,ثطہ صتہ,صتح,ثطح صطہ,صطح	سطح
महबूब	مہبُوب	مہبوب محبوب	محبوب

3) Although previous two techniques remove ambiguities for many words but this cannot cater real word errors for example Hindi word हवा can be mapped to Urdu ہوا/ meaning *air* and حوا/ meaning *first woman made by God*. Both are valid words in Urdu. In this situation we decide for the correct word based on the context. To resolve this ambiguity we use N-gram technique. This technique uses the likelihood of the ambiguous words based on its context in a sentence.

In this system first and second technique has been applied to solve problem of multi-Urdu-letters for a single Hindi letter.

Figure 1 shows block diagram for Hindi-to-Urdu transliteration.

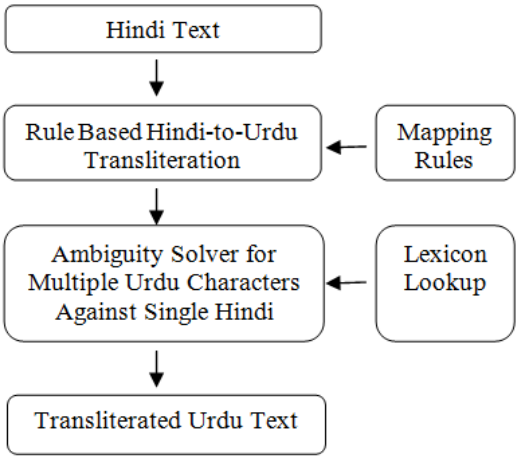

Figure 1. Hindi-to-Urdu Transliteration module

5.3 Resolving Vowel Issues in Urdu to Hindi Transliteration

People use vowels in speaking but they do not write vowels in writing Urdu. In Urdu vowels are represented by diacritic marks i.e zer (ِ), zabar (َ), pesh (ُ). But Hindi vowels are also written while writing Hindi text. So if we do not write vowels in Urdu, we cannot get an accurate transliteration of Urdu words into Hindi. To solve this problem we used automatic diacritization algorithm presented by Abbas [8] in pre-processing of Urdu Text. This algorithm takes plane Urdu text and produces Urdu text with diacritized marks. After this preprocessing step, discritized Urdu text is processed by Urdu-to-Hindi transliteration system. Conclusively we can say that if we do not write diacritic marks, we cannot correctly transliterate the given Urdu word into Hindi word. Table 5 shows Hindi transliteration of some Urdu words. Incorrect transliteration shows that the given word was not diacritized before transliteration process.

Table 6. Urdu to Hindi transliteration with diacritic marks.

Urdu word with diacritics	Transliterated Hindi word	Transliterated Hindi word without diacritics
مُلتان	मुलतान	मलतान
پاکِستان	पाकिसतान	पाकसतान
بِل	बिल	बल
بُل	बुल	बल
محبُوب	महबुब	महबोब
شُمالی	शुमाली	शमाली
مِتر	मितर	मतर
حُسن	हुसन	हसन
جھِیل	झील	झेल

Figure 2 shows block diagram for Urdu-to-Hindi transliteration.

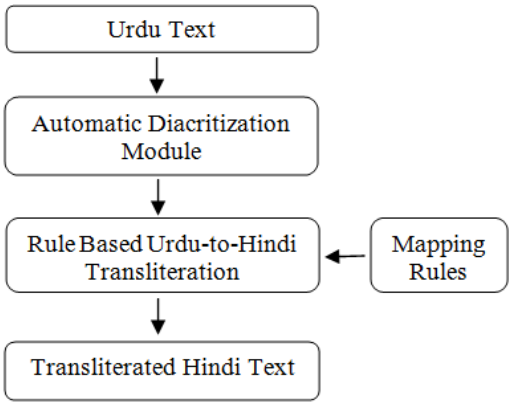

Figure 2. Urdu-to-Hindi Transliteration Module

6. RESULTS

The performance of the software is checked by taking several samples of Hindi text from BBC Hindi website. The transliterated text is checked for compliance with standard Urdu text. The accuracy of the results was around 95% which is also compared with already existing transliteration systems developed for this purpose. The same was done for transliteration of Urdu text to Hindi.

7. CONCLUSION AND FUTURE WORK

The research has tried to address two main issues in Hindi-Urdu transliteration systems (missing diacritic marks in Urdu and multiple character ambiguity for Hindi). Issues with simple rule based transliteration have been highlighted and their existing solutions are discussed. Enhancements in these solutions have been provided where needed to increase the accuracy of transliteration. The solution to multiple word ambiguity between cross language transliteration is handled successfully. Owning to the issue that diacritical marks are necessary for accurate Urdu to Hindi Transliteration, automatic diacretization algorithm is used before transliteration of Urdu to Hindi. Post-processing of transliterated words is carried out to alleviate the issues caused by differences in writing conventions. The system has gone through extensive testing and enhancements have been made to cater vowel and multiple character issues in Hindi-Urdu transliteration. In future bidirection work may be done with enhanced effiency of transliteration.

ACKNOWLEDGMENT

The authors would like to acknowledge Center for Research in Urdu Language Processing (CRUPL) and Dr Sarmad Hussain for providing needed data and guidelines to conduct this research.

REFERENCES

[1]. 1. Malik, M.G., C. Boitet, and P. Bhattacharyya. *Hindi Urdu machine transliteration using finite-state transducers*. in *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*. 2008. Association for Computational Linguistics.

[2]. 2. Durrani, N., et al. *Hindi-to-Urdu machine translation through transliteration*. in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. 2010. Association for Computational Linguistics.

[3]. 3. Ahmed, T. *Roman to Urdu transliteration using wordlist*. in *Proceedings of the Conference on Language and Technology*. 2009.

[4]. 4. Malik, A., et al. *A hybrid model for Urdu Hindi transliteration*. in *Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration*. 2009. Association for Computational Linguistics.

[5]. 5. Lehal, G.S. and T.S. Saini. *Development of a Complete Urdu-Hindi Transliteration System*. in *COLING (Posters)*. 2012.

[6]. 6. Lehal, G.S. and T.S. Saini. *A Hindi to Urdu transliteration system*. in *Proceedings of ICON-2010: 8th International Conference on Natural Language Processing, Kharagpur*. 2010.

[7]. 7. Jawaid, B. and T. Ahmed. *Hindi to Urdu conversion: beyond simple transliteration*. in *Conference on Language and Technology*. 2009.

[8]. 8. Raza, A. and S. Hussain. *Automatic diacritization for urdu*. in *Proceedings of the Conference on Language and Technology*. 2010.

[9]. 9. Butt, M., *The structure of complex predicates in Urdu*1995: Center for the Study of Language (CSLI).

[10]. 10. Bögel, T., et al., *Developing a finite-state morphological analyzer for Urdu and Hindi.* Finite State Methods and Natural Language Processing, 2007: p. 86.