Urdu Hindi Machine Transliteration using SMT

M. G. Abbas Malik
Faclty of Computing and IT
NJB
King Abdulaziz University,
Saudi Arabia
mgmalik@kau.edu.sa

Christian Boitet
GETALP – LIG
University of Grenoble, France
Christian.boitet@imag.fr

Laurent Besacier
GETALP – LIG
University of Grenoble, France
Laurent.Besacier@imag.fr

Pushpak Bhattcharyya
Indian Institute of Technology
Bombay, India
pb@iitb.ac.in

Abstract

Transliteration is a process of transcribing a word of the source language into the target language such that when the native speaker of the target language pronounces it, it sounds as the native pronunciation of the source word. Statistical techniques have brought significant advances and have made real progress in various fields of Natural Language Processing (NLP). In this paper, we have analysed the application of Statistical Machine Translation (SMT) for solving the problem of Urdu Hindi transliteration using a parallel lexicon. We have designed total 24 Statistical Transliteration (ST) systems by combining different types of *alignments*, *translation models* and *target language models*. We have performed total 576 experiments and have reported significant results. From Hindi–to–Urdu transliteration, we have achieved the maximum word-level accuracy of 71.5%. From Urdu–to–Hindi transliteration, the maximum word-level accuracy is 77.8% when the input Urdu text contains all necessary diacritical marks and 77% when the input Urdu text does not contain all necessary diacritical marks. At character-level, transliteration accuracy is more than 90% in both directions.

1 Introduction

Hindi is the national language of India. Urdu is the national language of Pakistan and is also one of the official languages of India. Urdu and Hindi are also considered as two dialects of the same language, called 'Hindustani' (Platts 1909), because of their common grammatical and linguistic structure (Rai 2000; Khan 2006). In the words of (Rai 2000), "One man's Hindi is another man's Urdu". In addition to the lexical differences, the other main difference between Urdu and Hindi is their writing systems. Urdu is written in a derived *Persio-Arabic* script and Hindi is written in the *Devanagari* script. Whether Hindi and Urdu are two different languages or not, they jointly represent the 2nd largest population of the world, including 1st and 2nd language speakers, after Chinese. This is shown in Table 1 (all figures are in millions).

	Native Speakers	2nd Language Speakers	Total
Hindi	366.00	487.00	853.00
Urdu	60.29	104.00	164.29
Total	426.29	591.00	**1,017.29**

Table 1: Size of Urdu and Hindi (Rahman 2004; Lewis 2009)

Empiricism was a prominent trend in computational linguistics in the 1940s and 50s. This trend was discouraged by Chomsky's claim that statistical approaches will always suffer from data scarcity and as a result radical approaches were more dominant for more than a decade. Empiricism re-emerged with the successful use of probabilistic models and Hidden Markov

The 4th Workshop on South and Southeast Asian NLP (WSSANLP), International Joint Conference on Natural Language Processing, pages 43–57,
Nagoya, Japan, 14-18 October 2013.

Model (HMM) by the *speech recognition* group at CMU (Harpy system) and IBM in the 1970s (Rabiner and Juang 1986; Rabinder 1989). Statistical approaches have brought significant advances and have made real progress in various fields of Natural Language Processing (NLP), like word sense disambiguation, POS tagging, information retrieval and especially Statistical Machine Translation (SMT) (Brown et al. 1990; Brown et al. 1993; Koehn et al. 2007; Lopez 2008; Koehn 2010), where the sense of real progress is the most visible on Internet. Statistical approaches are data-driven, and the sparsity of data is indeed a major problem for Statistical language processing, especially for under-resourced languages or language pairs.

Transliteration is a process of transcribing a word of the source language into the target language such that when the native speaker of the target language pronounces it, it sounds as the native pronunciation of the source word. Statistical approaches, especially SMT, are getting attention of researchers for solving the problems of transliteration (Knight and Graehl 1997; Lee and Choi. 1998; Stall and Knight 1998; Al-Onaizan and Knight 2002; AbdulJaleel and Larkey 2003; Li, Zhang and Su 2004; Ekbal, Naskar and Bandyopadhyay 2006; Finch and Sumita 2009; Kirschenbaum and Wintner 2009; Malik et al. 2009; Nabende 2009; Durrani et al. 2010). Malik et al. (2009) used a hybrid transliteration model for *Urdu–to–Hindi* transliteration. They replaced the translation model of the classical SMT by a rule-based non-deterministic transducer and used the Hindi language model to select the best transliteration. They reported an accuracy of 79.1%. Durrani et al. (2010) used a conditional probability model (M1) and a joint probability model (M2) for *Hindi–to–Urdu* translation. They incorporated their character-based transliteration model learnt from Urdu – Hindi parallel word list into Models M1 and M2. They reported very low BLUE scores of 19.35 and 18.34 for models M1 and M2 respectively.

We have developed statistical models for solving the problem of Urdu ↔ Hindi transliteration (bidirectional) using an Urdu – Hindi parallel word list resource, extracted from the data of a dictionary of Urdu, classical Hindi and English (Platts 1884) [1], digitized under the project "Digital South Asian Library" at University of Chicago, USA[2] and Center for Research Libraries[3].

In this paper, we have discussed our training data, the Urdu – Hindi parallel lexicon in Section 2. Various components of Statistical Transliteration (ST) systems like data alignment techniques, transliteration models and target language models are discussed in Section 3. 24 ST systems for Urdu – Hindi transliteration are described in Section 4. In Section 5, we have described the experimental setup and discussed the results of our tests. Finally the conclusion of the paper is given.

2 Training Data

Urdu – Hindi is an under-resourced language pair. The Digital South Asian Library (DSAL) has digitized "A Dictionary of Urdu, classical Hindi and English" (Platts 1884). The originally published dictionary includes the Urdu/Persio-Arabic transcriptions of all word entries, while it contains Hindi/Devanagari transcriptions only for words that are not of Persio-Arabic origin. The original DSAL dictionary data that we received in March 2007, did not contain the Urdu word transcriptions in the Persio-Arabic script, instead it contained their Roman transcriptions[4]. Two sample entries in the original DSAL's dictionary data are shown in Figure 1. We are particularly interested by the highlighted areas of these entries.

<div2 type="article" id="ābādī"><head><hi>ābādī</hi></head> P <pa>ābādī</pa> ābādī, s.f. Inhabited spot or place; colony; population, number of inhabitants; cultivated place;cultivation; the part of a village lands brought under cultivation; increased assessment (=beshī); prosperity; stateof comfort; happiness, joy, pleasure. </div2>

<div2 type="article" id="अबार_abār"><head><hi>अबार abār</hi></head> H <pa>abār</pa> अबार abār [S. अवार], s.m. This side, the nearbank of a river. </div2>

Figure 1: Sample entries in the DSAL dictionary

The silver highlighting ('id=') contains only the Roman transcription instead of the Urdu transcription of the word entry (1st sample entry), when the word is of Persio-Arabic origin, otherwise it contains the Hindi/Devanagari transcription '_' Roman transcription for the word entry (2nd sample entry). Each word entry has 'hi' and 'pa' tags for Hindi/Devanagari and Urd/Persio-Arabic transcriptions respectively. The format of the data is not strictly XML-based, but contains certain tags like 'hi' and 'pa'.

From this raw data, we extracted the Roman transcription and its Hindi/Devanagari transcrip-

[1] http://dsal.uchicago.edu/dictionaries/platts/
[2] http://dsal.uchicago.edu/
[3] http://www.crl.edu/
[4] The current online version contains Urdu/Persio-Arabic transcription, updated in August, 2008.

tion (if present) for each word entry. After an exhaustive analysis of the extracted Roman transcriptions, we built a finite-state transducer that can convert these Roman transcriptions into the Urdu/Persio-Arabic and Hindi/Devanagari (if not existed) transcriptions. In this way, we developed an Urdu – Hindi parallel lexicon containing total 55,253 words. We used 50,000 Urdu – Hindi parallel words as training data to build our statistical transliteration models for Urdu – Hindi machine transliteration. The remaining 5,253 Urdu – Hindi parallel words were used for testing and tuning purposes.

3 Statistical Transliteration (ST)

Following the classical Statistical Machine Translation (SMT) techniques, first we develop *alignments* between the parallel data. The only difference between a classical SMT system and our Statistical Transliteration (ST) system is the parallel data. In SMT, the parallel data consist of parallel sentences; In ST, it consists of parallel words. Examples of Urdu – Hindi parallel words are shown in Table 2 with their International Phonetic Alphabet (IPA) transcriptions and English glosses.

Hindi	Urdu with diacritics	Urdu without diacritics	IPA	English
अब्ब	اَبّا	ابا	əbbɑ	Father
इबलाग़	اِبلاغ	ابلاغ	ɪbəlɑx	Conveying
उबलाना	اُبلانا	ابلانا	ʊbəlɑnɑ	To boil
इबलीस	اِبلِیس	ابلیس	ɪbəlis	Devil
अभागेपन	ابھاگیپن	ابھاگیپن	əbʰɑgepən	Unfortunate
अप्रैल	اَپْرَیل	اپریل	əpræl	April
अच्छा	اَچّھا	اچھا	əʧʧɑ	Good

Table 2: Sample Urdu – Hindi parallel lexical entries

The Urdu words that we have generated from the Roman transcriptions from the DSAL dictionary data contain all required diacritical marks, clearly shown in Table 2. Diacritical marks are the back bone of the Urdu vowel system and they are mandatory for the correct pronunciation of an Urdu word, as well as Urdu computational linguistics (Zia 1999). Like in Arabic, diacritical marks are sparingly used in written Urdu (Zia 1999). To model this unfortunate situation, we developed another Urdu – Hindi parallel lexicon by removing all diacritics from the fully diacritized Urdu words, also shown in Table 2. In this way, we developed two types of Hindi – Urdu parallel data.

From these parallel Urdu – Hindi entries, we developed two types of alignments. Secondly, we developed *transliteration models* based on the alignments and *language models* based on monolingual Urdu and Hindi corpus. Finally, we joined these models to perform Urdu ↔ Hindi bidirectional transliteration.

3.1 Alignments

We developed two types of alignments that are discussed in the following two sections:

3.1.1 Character alignments

We can align the data at *character-level* by considering each Hindi – Urdu parallel word pair as a parallel sentence pair like classical SMT and each character in the parallel entry as a word. For *character alignments*, a space is introduced after each character in the Urdu (whether diacritized or not) and Hindi words in Urdu – Hindi parallel lexicons. Table 3 shows sample Urdu – Hindi parallel data for character alignments.

Hindi	Urdu with diacritics	Urdu without diacritics
अ ब ् ब	ا ب َ ا	ا ب ا
इ ब ल ा ग़	ا ِ ب ل ا غ	ا ب ل ا غ
उ ब ल ा न ा	ا ُ ب ل ا ن ا	ا ب ل ا ن ا
इ ब ल ी स	ا ِ ب ل ِ ی س	ا ب ل ی س
अ भ ा ग े प न	ا َ ب ھ ا گ ی پ ن	ا ب ھ ا گ ی پ ن
अ प ् र ै ल	ا َ پ ْ ر َ ی ل	ا پ ر ی ل
अ च ् छ ा	ا َ چ ھ َ ا	ا چ ھ ا

Table 3: Sample Urdu – Hindi parallel data for character alignment

From these two types of Urdu – Hindi parallel data, we developed two types of character alignments using GIZA++ (Och and Ney 2003) in both directions:

- Hindi and Urdu with diacritics character alignment,
- Hindi and Urdu without diacritics character alignment.

3.1.2 Cluster alignments

Alignment plays a critical role in SMT (Fraser and Marcu 2007; Kumar, Och and Macherey 2007; Huang 2009). The quality of parallel data and the word alignment have a significant impact on learning the translation model and consequently on the quality of the SMT system (Fraser and Marcu 2007; Huang 2009). It is always better do an analysis of the alignment and correct the alignment errors to reduce the Alignment Error Rate (AER).

We also analyzed the alignments produced by GIZA++. We found that we can improve our alignments to reduce the AER. The incorrect alignments are highlighted in Table 4 (below) that shows Hindi to Urdu with diacritics alignments of our sample words of Table 3.

The vowel representation in Urdu/Persio-Arabic script is highly complex and context-sensitive (Hussain 2004; Malik, Boitet and Bhattacharyya 2008; Malik et al. 2009). This highly complex and contextual representation leads to wrong character alignments, highlighted in Table 4. In the second row of Table 4, the Hindi vowel इ [ɪ] is aligned with ALEF (ا) and ZER (ِ) is aligned to NULL. The alignment is not completely incorrect, but the vowel इ [ɪ] must be aligned with both ALEF (ا) and ZER (ِ). Similarly, the Hindi vowel उ [ʊ] must be aligned with ALEF (ا) and PESH (ُ) in the third row. In these examples, one character in Hindi must be aligned with a sequence of characters in Urdu. Interestingly, we have observed that GIZA++ correctly aligns such cases for Urdu (with or without diacritics) to Hindi alignments.

1	# Sentence pair (6) source length 4 target length 5 alignment score : 1.70006e-05 अ ब ि ब ा NULL ([ ]) 5]) ا ([3]) ِ ([4 2]) ب ([1]) ا])
2	# Sentence pair (114) source length 6 target length 5 alignment score : 0.00032306 इ ब ल ा ग़ NULL ([ ])]) غ ([4]) ا ([3]) ل ([2]) ب ([ ]) ِ ([1]) ا 5])
3	# Sentence pair (115) source length 7 target length 6 alignment score : 0.000154595 उ ब ल ा न ा NULL ([ ])]) ن ([4]) ا ([3]) ل ([2]) ب ([ ]) ُ ([1]) ا 6]) ا ([5])
4	# Sentence pair (128) source length 7 target length 5 alignment score : 5.58545e-05 इ ब ल ी स NULL ([ ])]) ی ([ ]) ِ ([3]) ل ([2]) ب ([ ]) ِ ([1]) ا 5]) س ([4])
5	# Sentence pair (167) source length 9 target length 7 alignment score : 3.20243e-05 अ भ ा ग ै प न NULL ([ ])]) ی ([4]) گ ([3]) ا ([ ]) ھ ([2]) ب ([1]) ا 7]) ن ([ ]) ً ([6]) پ ([5])
6	# Sentence pair (464) source length 6 target length 6 alignment score : 7.74271e-06 अ प ् र ै ल NULL ([ ]) ل ([5]) ی ([3]) َ ([4]) ر ([2]) پ ([1]) ا 6])])
7	# Sentence pair (1183) source length 5 target length 5 alignment score : 3.13657e-05 अ च ् छ ा NULL ([ ]) 5]) ا ([3]) ً ([4]) ھ ([2]) چ ([1]) ا])

Table 4: Character alignment examples from Hindi to Urdu with diacritics

All such vowel alignments can be improved by clustering the specific sequences of characters in the Urdu side.

In the case of consonants, we also observe a few alignment problems. In Urdu, *gemination* of a consonant is marked with a SHADDA (ّ), while in Hindi, it is written as a *conjunct* form. The highlighted alignments of the first row of Table 4 align the Hindi characters ब [b] and ् with the Urdu characters BEH (ب) and SHADDA (ّ) respectively. Again this alignment is not completely wrong, but this geminated consonant alignment problem is more evident for Urdu to Hindi alignment where a Hindi consonant is aligned with NULL. An example of Urdu to Hindi alignment is shown in Table 5. This problem can be resolved by clustering the gemination representations both in Urdu and Hindi.

Sentence pair (754) source length 9 target length 10 alignment score : 8.0263e-13 ا ِ ت ّ ِ ف ا ق ا ٔ NULL ([9]) इ ([1 2]) त ([3]) ् ([4]) त ([ ]) ि ([5]) फ़ ([6]) ा ([7]) क़ ([8]) न ([10])

Table 5. Gemination alignment from Urdu to Hindi

We also observe alignment problems for aspirated consonants because they are represented by a sequence of characters in Urdu and by either a single character or a sequence of characters in Hindi. For Hindi to Urdu alignment, this problem is highlighted in row 5 and 7 of Table 4. For Urdu to Hindi alignment, an example is shown in Table 6.

Sentence pair (1183) source length 5 target length 5 alignment score : 8.57561e-05 ا ھ چ ا NULL ([ ]) अ ([1]) च ([2]) ् ([ ]) छ ([3 4]) ा ([5])

Table 6. Aspirated consonant alignment from Urdu to Hindi

All these problems increase the AER. Thus we decided to cluster certain character sequences in Urdu – Hindi parallel lexical entries to enhance the alignments and consequently transliteration between Urdu and Hindi. For clustering, we developed finite-state transducers that generate Urdu – Hindi clustered parallel lexicons. Clustered Urdu – Hindi parallel lexicons for our sample data of Table 2 with IPAs are shown in Table 7.

Hindi	Urdu with diacritics	IPA	Urdu without diacritics	IPA
अ ब्ब	اَبّ ا	ə bb ɑ	ا ب ا	ə b ɑ
इ ब ल ा ग़	اِ ب ل ا غ	ɪ b l ɑ x	ا ب ل ا غ	ə b l ɑ x
उ ब ल ा न ा	اُ ب ل ا ن ا	ʊ b l ɑ n ɑ	ا ب ل ا ن ا	ə b l ɑ n ɑ
इ ब ल ी स	اِ ب ل ی س	ɪ b l i s	ا ب ل ی س	ə b l e s
अ भ ा ग े प न	اَ بھ ا گ ی پ ن	ə bʰ ɑ g e p ə n	ا بھ ا گ ی پ ن	ə bʰ ɑ g e p ə n
अ प ् र ै ल	اَ پ ْ ر َی ل	ə p r æ l	ا پ ر ی ل	ə p r e l
अ च्छ ा	اَ چّھ ا	ə ʧʧʰ ɑ	ا چھ ا	ə ʧʰ ɑ

Table 7: Sample clustered Urdu – Hindi lexical entries

Using these clustered Urdu – Hindi parallel lexical entries, we developed two types of *cluster alignments* using GIZA++ in both directions:

- Hindi and Urdu with diacritics cluster alignments
- Hindi and Urdu without diacritics cluster alignments

All the alignments problem discussed above are solved in cluster alignments as shown in Table 8.

1	# Sentence pair (6) source length 3 target length 3 alignment score : 0.0214204 अ ब्ब ा NULL ([ ]) 3]) ا ([2]) بَ ([1]) ا])
2	# Sentence pair (114) source length 5 target length 5 alignment score : 0.0942275 इ ब ल ा ग़
	NULL ([ ]) 5]) غ ([4]) ا ([3]) ل ([2]) ب ([1]) ا])
3	# Sentence pair (115) source length 6 target length 6 alignment score : 0.0373352 उ ब ल ा न ा NULL ([ ]) 5]) ن ([4]) ا ([3]) ل ([2]) ب ([1]) اُ 6]) ا ([ ])
4	# Sentence pair (128) source length 5 target length 5 alignment score : 0.0430949 इ ब ल ी स NULL ([ ])]) س ([4]) ی ([3]) ل ([2]) ب ([1]) اِ 5])
5	# Sentence pair (167) source length 8 target length 7 alignment score : 0.000313045 अ भ ा ग े प न NULL ([ ]) 5]) ی ([4]) گ ([3]) ا ([2]) ہ ([1]) ا 7]) ن ([ ]) ٔ ([6]) پ ([ ])
6	# Sentence pair (464) source length 5 target length 6 alignment score : 1.71945e-05 अ प ् र ै ल NULL ([3])]) ل ([5]) یٔ ([4]) ر ([2]) پ ([1]) ا 6])
7	# Sentence pair (754) source length 8 target length 7 alignment score : 0.000371183 इ त्त ि फ़ ा क़ न NULL ([ ]) 5]) ا ([4]) ف ([3]) ِ ([2]) تّ ([1]) اِ 7]) ٔ ([ ]) ا ([6]) ق ([ ])
8	# Sentence pair (1183) source length 3 target length 3 alignment score : 0.0207299 अ च्छ ा NULL ([ ]) 3]) ا ([2]) چّ ([1]) ا])

Table 8: Sample Urdu - Hindi cluster alignments

These better cluster alignments will turn out to help to learn a better quality transliteration model and to enhance the quality of our Urdu ↔ Hindi transliterations.

3.2 Transliteration/Translation Models

Based on the character and cluster alignments, we developed 8 Urdu – Hindi transliteration models (or translation models of classical SMT) using the Moses toolkit (Koehn et al. 2007).

1. **M1**: learned from Hindi to Urdu with diacritics character alignment
2. **M2**: learned from Hindi to Urdu without diacritics character alignment
3. **M3**: learned from Hindi to Urdu with diacritics cluster alignment
4. **M4**: learned from Hindi to Urdu without diacritics cluster alignment
5. **M5**: learned from Urdu with diacritics to Hindi character alignment

6. **M6**: learned from Urdu without diacritics to Hindi character alignment
7. **M7**: learned from Urdu with diacritics to Hindi cluster alignment
8. **M8**: learned from Urdu without diacritics to Hindi cluster alignment

For developing these transliteration models, we used the training script 'train-factored-phrase-model.perl' with options 'grow-diag-final' and 'msd-bidirectional-fe' (default options) for alignment and re-ordering respectively, to learn these models from different type of alignments.

3.3 Target Language Models

A *target language model* $P(e)$ is a probabilistic model that scores the well-formedness of different translation solutions produced by the translation model (Koehn, Och and Marcu 2003; Zens and Ney 2003; Och and Ney 2004; Al-Onaizan and Papineni 2006). It generates a probability distribution over possible sequences of words and computes the probability of producing a given word w_1 given all the words that precede it in the sentence (Al-Onaizan and Papineni 2006). We developed multiple target language models depending on the type of alignments used in the transliteration models and the target language. We broadly categorize them into *word language models* and *sentence language models*, discussed below.

3.3.1 Word Language Models (WLM)

A *word language model* is a 6-gram statistical model that gives a probability distribution over possible sequences of characters and computes the probability of producing a given character or cluster C1, given the 5 characters or clusters that precede it in the word. We developed Hindi – Urdu parallel lexicons for learning the various Hind – Urdu alignments and transliteration models. The target side words of the parallel lexicons are used to generate word language models using the SRILM freeware [5]. For example, we developed *Urdu Word Language Models with Diacritics* from our character and clustered Urdu words with diacritics, and used them as target language models with the corresponding transliteration model. We thus developed total 6 different word language models, 2 for Hindi (character-based & cluster-based) and 4 for Urdu.

[5] http://www.speech.sri.com/projects/srilm/

3.3.2 Sentence Language Models (SLM)

Similarly to a word language model, a sentence language model is also a 6-gram statistical model that computes the probability of producing a given character or cluster C1, given the 5 characters or clusters that precede it in the sentence. The Urdu – Hindi pair is an under-resourced pair, but there exist some monolingual corpora for both Urdu and Hindi.

For Hindi, a Hindi corpus of more than 3 million words is freely available at the "Resource Center for Indian Language Technology Solutions" of the Indian Institute of Technology Bombay (IITB)[6]. We processed this Hindi corpus and extracted a Hindi sentence corpus that contains one sentence per line, for a total of 173,087 Hindi sentences. From it, we developed a character-level Hindi corpus by introducing a space after each character and a cluster-level Hindi corpus by applying our Hindi clustering finite-state transducer on the character-level Hindi corpus. These two character-level and cluster-level Hindi corpora were used to develop character-level and cluster-level Hindi Sentence Language Models using the SRILM toolkit.

A monolingual Urdu corpus (Reference # ELRA-W0037) of more than 2 million words is also available from the "Evaluations and Language Resources Distribution Agency" (ELRA/ELDA) [7]. This corpus was developed under the EMILLE [8] project of Lancaster University, UK. Like the Hindi corpus, we processed this Urdu corpus and extracted from it an Urdu sentence corpus. It contains a total of 127,685 sentences. We developed a character-level and cluster-level Urdu corpus by introducing a space after each character and then by applying clustering. Finally, we developed character-level and cluster-level Urdu Sentence Language Models using the SRILM toolkit.

Similar to *word language model*, We have developed total 6 sentence language models, 2 for Hindi and 4 for Urdu. Another set of 6 target language models are developed by combining the corresponding word and sentence language models.

4 Statistical Transliteration Systems

We have developed total 8 transliteration models, 4 for Hindi–to–Urdu and 4 for Urdu–to–Hindi transliteration. We have also developed 18

[6] http://www.cfilt.iitb.ac.in/
[7] http://www.elda.org/
[8] http://www.emille.lancs.ac.uk/index.php

target side language models. By combining these transliteration and target language models, we have developed total 24 Urdu↔Hindi statistical transliteration systems, 12 for Hindi–to–Urdu and 12 for Urdu–to–Hindi transliteration.

For Hindi–to–Urdu transliteration, we have built 4 translation models based on different Hindi–Urdu alignments and 8 Urdu target language models, discussed above. In the Moses toolkit, we can direct the SMT system to use multiple target language models. Thus we built 4 other target language models by combining our *Urdu (with or without diacritics) word language models (character and cluster level)* and *Urdu sentence language models (character and cluster level)*. Hindi–to–Urdu statistical transliteration systems are shown in Figure 2.

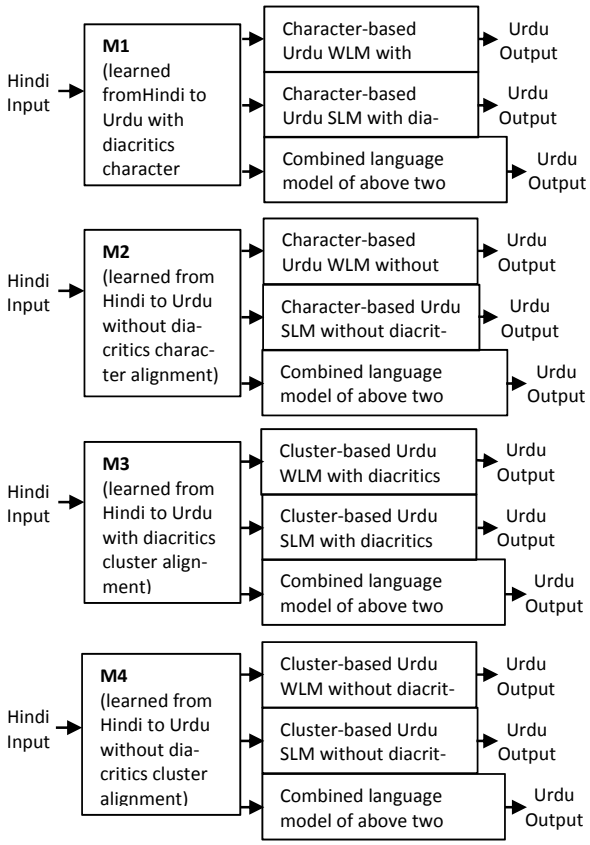

Figure 2: Hindi–to–Urdu Statistical Transliteration (ST) systems

We developed an Urdu – Hindi parallel lexicon containing total 55,253 words. 50,000 words were used to develop transliteration models and 2,500 for tuning each transliteration model. The remaining 2,753 Urdu – Hindi parallel words were used as Test Set 1 for testing purposes.

We also built 12 statistical transliteration systems for Urdu–to–Hindi transliteration. We developed 4 transliteration models based on different Urdu–Hindi alignments and 4 Hindi target language models. As for Hindi–to–Urdu transliteration, we combined *Hindi word language model* and *Hindi sentence language model*. Figure 3 shows different Urdu–to–Hindi statistical transliteration models.

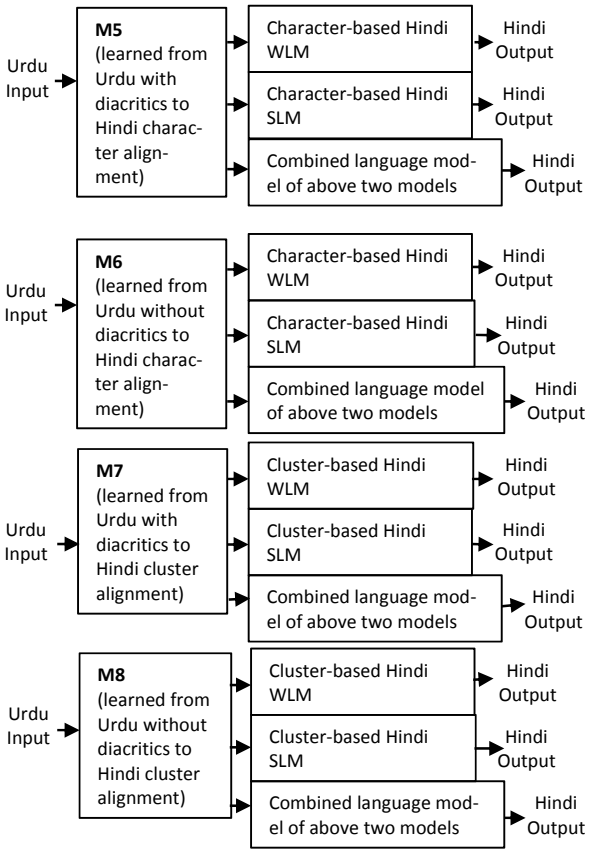

Figure 3: Figure 3: Urdu–to–Hindi Statistical Transliteration (ST) systems

5 Experiments and Results

Urdu – Hindi transliteration models are learnt from Urdu – Hindi parallel lexicons, thus the input to these Statistical Transliteration (ST) systems must be a word and not a running text or a sentence. The Hindi or Urdu input text is pre-processed to generate a list of Hindi or Urdu words before feeding it to the ST systems. The output of the ST systems is then post-processed to generate the final Urdu or Hindi output text.

5.1 Test Sets

We developed three Urdu – Hindi test sets. **Test Set 1** contains 2,753 Hindi–Urdu parallel words.

Test Set 2 contains 200 sentences (4,281 words) of Hindi origin that were extracted at random from a Hindi corpus of more than 3 million words. It is a common practice in the Hindi community to use the characters क [k], ख [k^h], ग [g], ज [ʤ], ड [ɖ], ढ [$ɖ^h$] and फ [p] instead of the characters क़ [q], ख़ [x], ग़ [ɣ], ज़ [z], ड़ [ɽ], ढ़ [$ɽ^h$] and फ़ [f] respectively, due to their shape similarities. In Test Set 2, the extracted Hindi sentences were edited and corrected for these typographical errors. Then, we translated the extracted Hindi sentences into Urdu by using an online Hindi–Urdu transliteration system[9]. These translated Urdu sentences were post-edited to remove errors, and all necessary diacritical marks were introduced in the Urdu text. Diacritical marks are vital for Urdu to Hindi transliteration, but they are sparingly used by people in writing. To compute the performance of our ST systems in this unfortunate but real situation, we developed another Urdu test data by removing the diacritical marks from the post-edited 200 Urdu sentences. These Hindi and Urdu (with and without diacritics) data served as input to our ST systems as well as an output reference for the automatic transliteration evaluations

Test Set 3 contains 226 sentences (4,632 words) of Urdu origin that were extracted at random from the Urdu corpus of more than 2 million words. To build Test Set 3, we first edited the extracted Urdu sentences for any error and restored the missing but necessary diacritics. We also developed a new Urdu test data without diacritics by removing all diacritical marks from the edited Urdu sentences. Then the edited Urdu sentences with diacritics were translated into Hindi using online an Urdu – Hindi transliteration system. The translated Hindi sentences were then post-edited for any error. These data are also used both as an input as well as a reference output.

5.2 Experiments

We experimented with 12 Hindi–to–Urdu and 12 Urdu–to–Hindi ST systems. Each ST system was tuned with 2,500 Urdu – Hindi parallel words using the Moses 'mert-moses.pl' script. Thus we have another set of 24 ST systems with tuning and it raises the total number of ST systems to 48. During the application of these ST systems on our difference test sets, we also varied the re-ordering parameter to analyse its effect on Urdu – Hindi ST systems. Again, this parameter varia-variation doubled the total number of our ST systems, from 48 to 96.

For Hindi–to–Urdu transliteration, we performed 96 experiments for each test set, that is in total 288 experiments. Each Urdu output was then post-processed (diacritics were removed) to compare it with Urdu references with and without diacritics.

For Urdu–to–Hindi transliteration, we have two types of inputs, with and without diacritics. Again, we conducted 288 Urdu–to–Hindi transliteration experiments, from which we computed character-level, word-level and sentence level accuracies.

5.3 Results and Discussion

Due to space limitation, we will only report and discuss the results of particular interest. We subdivide the results for each test set by the transliteration model, the alignment strategy, and the type of input/output data, because it is difficult to present all results in only one large table.

5.3.1 Hindi–to–Urdu transliteration results

Table 9 shows all the results of ST systems for Hindi–to–Urdu transliteration of transliteraiton models M1 and M2 for Test Set 2. The best results for M1 are 71.5% and 5.5% at the word-level and the sentence level respectively, shown in bold in the upper grid of Table 9. The best result at word-level and sentence level are produced by the ST systems *M1-Urdu SLM+ WLM-WD-No-Reordering* and *M1-Urdu SLM-No-Reordering* respectively.

[9] http://www.sanlp.org/HUMT/index.html

SMT Model	Sentence Accuracy		Word accuracy		Character accuracy	
	default output	Processed output	default output	processed output	default output	processed output
M1-Urdu WLM-WD-With-Reordering	0.5%	3%	26.1%	65.7%	89.1%	93.3%
M1-Urdu WLM-WD-No-Reordering	0.5%	3%	26.1%	65.7%	89.1%	93.3%
M1-Urdu WLM-WD-Tuned-With-Reordering	1%	3%	34.4%	62.9%	88.7%	92.7%
M1-Urdu WLM-WD-Tuned-No-Reordering	1%	3%	34.4%	62.9%	88.7%	92.7%
M1-Urdu SLM-WD-With-Reordering	1%	4%	48.9%	62.2%	84.9%	92.2%
M1-Urdu SLM-WD-No-Reordering	1%	**5.5%**	49.7%	64.2%	85.8%	93.3%
M1-Urdu SLM-WD-Tuned-With-Reordering	0.5%	3.5%	34.2%	63.3%	88.6%	92.6%
M1-Urdu SLM-WD-Tuned-No-Reordering	0.5%	3.5%	34.2%	63.3%	88.6%	92.6%
M1-Urdu SLM+WLM-WD-With-Reordering	1%	4.5%	50.5%	70.9%	89.0%	94.3%
M1-Urdu SLM+WLM-WD -No-Reordering	1%	4.5%	50.8%	**71.5%**	89.2%	**94.5%**
M1-Urdu SLM+WLM-WD -Tuned-With-Reordering	1%	3%	33.9%	62.7%	88.6%	92.6%
M1-Urdu SLM+WLM-WD -Tuned-No-Reordering	1%	3%	33.9%	62.7%	88.6%	92.6%
M2-Urdu WLM-WOD-With-Reordering	3%	3%	63.6%	63.6%	93.3%	93.3%
M2-Urdu WLM-WOD-No-Reordering	3%	3%	63.6%	63.6%	93.3%	93.3%
M2-Urdu WLM-WOD-Tuned-With-Reordering	3%	3%	64.8%	64.8%	92.6%	92.6%
M2-Urdu WLM-WOD-Tuned-No-Reordering	3%	3%	64.8%	64.8%	92.6%	92.6%
M2-Urdu SLM-WOD-With-Reordering	**5.5%**	**5.5%**	63.2%	63.2%	92.9%	92.9%
M2-Urdu SLM-WOD-No-Reordering	**5.5%**	**5.5%**	63.8%	63.8%	93.5%	93.5%
M2-Urdu SLM-WOD-Tuned-With-Reordering	3.5%	3.5%	64.8%	64.8%	92.8%	92.8%
M2-Urdu SLM-WOD-Tuned-No-Reordering	3.5%	3.5%	64.8%	64.8%	92.8%	92.8%
M2-Urdu SLM+WLM-WOD-With-Reordering	5%	5%	**68.5%**	**68.5%**	**93.7%**	**93.7%**
M2-Urdu SLM+WLM-WOD -No-Reordering	5%	5%	**68.5%**	**68.5%**	**93.7%**	**93.7%**
M2-Urdu SLM+WLM-WOD -Tuned-With-Reordering	3%	3%	64.8%	64.8%	92.7%	92.7%
M2-Urdu SLM+WLM-WOD -Tuned-No-Reordering	3%	3%	64.8%	64.8%	92.7%	92.7%

Table 9: Test Set 2 results of Hindi–to–Urdu ST systems (character alignments

Abbreviation: SLM = sentence language model, WLM = word language model, WD = with diacritics, WOD = without diacritics.

The best results for the ST systems, developed from M2 are 68.5% and 5.5% at word-level and sentence-level, respectively, shown in bold in the lower grid of Table 9. In this case also, the best results at word-level and sentence-level are produced by different ST systems.

Figure 4 shows a sample Hindi source text from Test set 2 of Hindi origin.

भारतीय सांस्कृतिक परंपराएँ कपिला वात्स्यायन कलाओं के क्षेत्र में सृजनात्मक कार्य प्रश्न बहन जी , आप का विद्याध्ययन प्रतिभासूचक रहा है और उस के बाद भारत सरकार के संस्कृतिविभाग के अध्यक्ष के नाते भी आप का कार्य उच्चकोटि का रहा है ;
इस पद पर कार्य करते हुए आप पर भारत के अन्तर्राष्ट्रीय सांस्कृतिक सम्बन्धों और भारत के भीतर सांस्कृतिक नीति के निर्माण का दायित्व रहा है , इस क्षेत्र में ऐसी भिन्नभिन्न चीज़ें आती हैं जैसे संग्रहालय , पुरातत्व , पुरालेख , ग्रंथालय और कलाएँ ; केन्द्रीय स्तर पर भी और राज्य स्तर पर भी ।
आप का सरकारी कार्य तो भारत और विदेशों में सुविदित रहा है पर भारत की प्रदर्शनकारी और सुघट्य प्लास्टिक कलाओं के वर्गीकरण और व्याख्या करने के क्षेत्र में आप के सृजनात्मक योगदान के बारे में कम से कम औसत स्तर के शिक्षित सामान्य व्यक्ति को कुछ अधिक जानकारी नहीं है ।
क्या आप हमारे पाठकों के लिए भारतीय कलाओं और सौन्दर्यशास्त्र के क्षेत्र में किये गये अपने सृजनात्मक और अन्वेषणात्मक कार्य का संक्षिप्त विवरण देने की कृपा करें गी ?

Figure 4: A sample Hindi input text for Hindi–to–Urdu transliteration

Table 10 shows the Urdu output of the ST systems M1-Urdu SLM+WLM-WD-No-Reordering and M1-Urdu SLM-WD-No-Reordering that produced the best results at word-level and sentence-level, respectively.

Urdu reference without diacritics	Processed Urdu output of M1-Urdu SLM+WLM-WD-No-Reordering
بھارتیہ سانسکرتک پرنپراءیں کپلا واتسیاین کلاءوں کے کشیتر میں سرجناتمک کاریہ پرشن بہن جی ، آپ کا ودیادھیین پرتبھاسوچک رہا ہے اور اس کے بعد بھارت سرکار کے سنسکرتوبھاگ کے ادھیکش کے ناتے بھی آپ کا کاریہ اچکوٹی کا رہا ہے ؛ اس پد پر کاریہ کرتے ہوءے آپ پر بھارت کے انتراشٹریہ سانسکرتک سمبندھوں اور بھارت کے بھیتر سانسکرتک نیتی کے نرمان کا دایتو رہا ہے ، اس کشیتر میں ایسی بھنبھن چیزیں آتی ہیں جیسے سنگرہالیہ ، پراتنو ، پرالیکھ ، گرنتھالیہ اور کلاءیں ؛ کیندریہ سطر پر بھی اور راجیہ سطر پر بھی - آپ کا سرکاری کاریہ تو بھارت اور ودیشوں میں سودت رہا ہے پر بھارت کی پردرشنکاری اور سگھٹیہ پلاسٹک کلاءوں کے ورگیکرن اور ویاکھیا کرنے کے کشیتر میں آپ کے سرجناتمک یوگدان کے بارے میں کم سے کم اوسط سطر کے شکشت سامانیہ ویکتی کو کچھ ادھک جانکاری نہیں ہے - کیا آپ ہماری پاٹھکوں کی لی بھارتیہی کلاوں اور سوندریشاسترہ کی کشیتر میں کیے گیے اپنی سرجناتمک اور انویشناتمک کاریہ کا سنکشپت ویورنہ دینی کی کرپا کریں گی ؟	بھارتیہی سانسکرتیک پرنپرہیں کپیلا واتسیاین کلاوں کی کھیتر میں سرجناتمک کاریہ پرشن بہن جی ، آپ کا ویدیادھیین پرتیبھاسوچکا رہا ہی اور اس کی یاد بھارت سرکار کی سنسکرتیوبھاگ کی ادھیکش کی ناتی بھی آپ کا کاریہ اچکوٹی کا رہا ہی ؛ اس پد پر کاریہ کرتی ہوای آپ پر بھارت کی انتراشٹریے سانسکرتیک سمبندھوں اور بھارت کی بھیتر سانسکرتیک نیتی کی نرمان کا دایتو رہا ہی ، اس کھیتر میں ایسی بھینابھینین چیزیں آتی ہیں جیسے سنگرہالی ، پرانتو ، پرالیکھ ، گرنتھالی اور کلایں ؛ کیندریہ سر پر بھی اور راجی سر پر بھی - آپ کا سرکاری کاریہ تو بھارت اور ویدیشوں میں سویدیت رہا ہی پر بھارت کی پردرشنکاری اور سگھٹی پلاسٹک کلاوں کی ورگیکرن اور ویاکھایا کرنے کی کھیتر میں آپ کی سرجناتمک یوگدان کی بہاری میں کم سی کم اوست سر کی شکسیت سامانی ویکتی کو کچھ ادھیک جانکاری نہیں ہی - کیا آپ ہماری پاٹکوں کی لی بھارتیہی کلاوں اور سوندریشاستر کی کھیتر میں کہیے گیے اپنے سرجناتمک اور انوےشناتمک کاریہ کا سنکشت ویوران دینو کی کرپا کریں گی ؟

Table 10: Sample Hindi–to–Urdu outputs with for the best ST systems for model M1

On average, there are 9.4 and 16 errors per sentence in the Urdu outputs of the ST systems M1-Urdu SLM+WLM-WD-No-Reordering and M1-Urdu SLM-WD-No-Reordering. In terms of usability of the output text, these outputs are not usable and require a huge amount of effort for post-editing. Therefore, these SMT systems would be ranked quite low by a user. From here onward, we will not give all results like we did in Table 9, but report only the results of particular interest.

For Test Set 3, the best results produced by the ST systems for models M1 and M2 are shown in Table 11.

SMT Model	Sentence Accuracy		Word accuracy		Character accuracy	
	default output	processed output	default output	processed output	default output	processed output
M1	0.9%	**4%**	57.3%	**71.1%**	90.6%	93.3%

Table 11: Test Set 3 results of Hindi–to–Urdu ST systems (character alignments)

For Test Set 2 and 3, the best results produced by the ST systems for models M3 and M4 are given in Table 12.

HU Test Set 2

SMT Model	Sentence accuracy		Word accuracy	
	default output	processed output	default output	processed output
M3	1%	**5.5%**	53.4%	66.6**%**
M4	**5.5%**	**5.5%**	65.3%	65.3%
M4	**5.5%**	**5.5%**	66.2%	66.2%
M4	**5.5%**	**5.5%**	69.5%	69.5%
M4	**5.5%**	**5.5%**	**69.7%**	**69.7%**

HU Test Set 3

SMT Model	Sentence Accuracy		Word accuracy	
	default output	processed output	default output	processed output
M3	0.9%	**4.9%**	58.0%	69.3%
M4	3.5%	3.5**%**	**68.0%**	**68.0%**
M4	3.5%	3.5**%**	**68.0%**	**68.0%**

Table 12: Test 2 and 3 results of Hindi–to–Urdu ST systems (cluster alignments)

For Test Set 1, the best results are 78.4% and 79.7% for the default and the processed Urdu output by the ST systems M4-Urdu SLM+WLM-WOD-Tuned-No-Reordering and M3-Urdu SLM+WLM-WD-Tuned-No-Reordering respectively. Test Set 1 consists of a word list, so there is no meaning of sentence-level results here. For Test Set 1, the best results are 78.3% and 80.2% for the default and the processed Urdu output by the ST systems M4-Urdu SLM+WLM-WOD-Tuned-No-Reordering and M3-Urdu SLM+WLM-WD-Tuned-No-Reordering respectively.

The ST systems developed from Urdu – Hindi parallel data without diacritics in the Urdu data are performing well compared to the systems developed from data that contains diacritical marks, because the removal of diacritical marks reduces the complexity of the transliteration problem.

5.3.2 Urdu–to–Hind transliteration results

Table 13 and Table 14 shows the best results of Urdu–to–Hindi transliteration of our ST systems for models M5 and M6 for HU Test Set 2 and 3 respectively.

HU Test Set 2

SMT Model	Sentence Accuracy		Word accuracy	
	with diacritics	without diacritics	with diacritics	without diacritics
M5	**5.5%**	2%	**72.2%**	57.9**%**
M5	**5.5%**	2%	**72.2%**	57.9**%**
M6	0.5%	5%	50.1%	**77.0%**
M6	0.5%	5%	50.1%	**77.0%**

Table 13: Test Set 2 best results for Urdu–to–Hindi ST systems (character alignments)

Urdu references with and without diacritics of Table 10 are our sample Urdu inputs for Urdu–to–Hindi transliteration. Table 15 shows the Hindi output for the sample Urdu input text with diacritics of the ST system M5-Hindi SLM+WLM-Tuned-No-Reordering with its Hindi reference. On average, the Hindi output of Table 15 contains 10.8 errors per sentence. A real user of the system would rate this output very low or even totally unacceptable.

HU Test Set 3

SMT Model	Sentence Accuracy		Word accuracy	
	with diacritics	without diacritics	with diacritics	without diacritics
M5	5.3%	0.4%	**77.8%**	57.9%
M5	5.3%	0.4%	**77.8%**	57.9%
M6	0%	0.4%	44.8%	**60.1%**
M6	0%	0.4%	44.8%	**60.1%**

Table 14: Test Set 3 best results for Urdu–to–Hindi ST systems (character alignments)

Urdu references with and without diacritics of Table 10 are our sample Urdu inputs for Urdu–to–Hindi transliteration. Table 15 shows the Hindi output for the sample Urdu input text with diacritics of the ST system M5-Hindi SLM+WLM-Tuned-No-Reordering with its Hindi reference. On average, the Hindi output of Table 15 contains 10.8 errors per sentence. A real user of the system would rate this output very low or even totally unacceptable.

Table 16 shows the Hindi output of the ST system M6-Hindi SLM+WLM-No-Reordering for the sample Urdu input without diacritics. The Hindi output of Table 16 also contains 10.8 errors per sentence.

Hindi reference	Hindi output from the Urdu text with diacritics M5-Hindi SLM+WLM-Tuned-No-Reordering
भारतीय सांस्कृतिक परंपराएँ कपिला वात्स्यायन कलाओं के क्षेत्र में सृजनात्मक कार्य प्रश्न बहन जी , आप का विद्याध्ययन प्रतिभासूचक रहा है और उस के बाद भारत सरकार के संस्कृतिविभाग के अध्यक्ष के नाते भी आप का कार्य उच्चकोटि का रहा है ;	भारतीय सांसंकृतिक परंपराईं कपिला वातंसंयायन कलावं के कंृषेतंर में सरज्ञातंमक कारंृय पंरषंन बहन जी , आप का विदंयाधंयेन पंरतिभासूचक रहा है और उस के बअद भारत सरकार के संसंकृतिविभाग के अधंेकंष के नाते भी आप का कारंृय उच्चकोटी का रहा है ;
इस पद पर कार्य करते हुए आप पर भारत के अन्तर्राष्ट्रीय सांस्कृतिक सम्बन्धों और भारत के भीतर सांस्कृतिक नीति के निर्माण का दायित्व रहा है , इस क्षेत्र में ऐसी भिन्नभिन्न चीज़ें आती हैं जैसे संग्रहालय , पुरातत्व , पुरालेख , ग्रंथालय और कलाएँ ; केन्द्रीय स्तर पर भी और राज्य स्तर पर भी ।	इस पद पर कारंृय करते हूए आप पर भारत के अनंतर्राषंटंरीया सांसंकृतिक समंबनंधों और भारत के भीतर सांसंकृतिक नीति के निरंमान का दायितंवो रहा है , इस कंृषेतंर में ऐसी भिन्नभिन्न चीज़ें आती हैं जैसे संगंृराल्य , पुरातततंवो , पुरालेख , गंरंथाल्य और कलाईं ;
आप का सरकारी कार्य तो भारत	केनंदंरीया संतर पर भी और

और विदेशों में सुविदित रहा है पर भारत की प्रदर्शनकारी और सुघट्य प्लास्टिक कलाओं के वर्गीकरण और व्याख्या करने के क्षेत्र में आप के सृजनात्मक योगदान के बारे में कम से कम औसत स्तर के शिक्षित सामान्य व्यक्ति को कुछ अधिक जानकारी नहीं है । क्या आप हमारे पाठकों के लिए भारतीय कलाओं और सौन्दर्यशास्त्र के क्षेत्र में किये गये अपने सृजनात्मक और अन्वेषणात्मक कार्य का संक्षिप्त विवरण देने की कृपा करें गी ?	राजंय संतर पर भी । आप का सरकारी कारंय तो भारत और विदेशों में सुविदित रहा है पर भारत की पंरदरंशंकारी और सुघटंय पंलासंटिक कलावं के वरंगीकरण और वंयाखंया करने के कंषेतंर में आप के सरज्ञातंमक योगदान के बारे में कम से कम औसत संतर के षिकंषित सामानंय वंेकंती को कुछ अधिक जानकारी नहीं है । कंया आप हमारे पाठकों के लिये भारतीय कलावं और सौनंदरंेासंतंर के कंषेतंर में किये गऽे अपने सरज्ञातंमक और अनंवेशनातंमक कारंय का सनकंषिपंत विवरन देने की कृपा करें गी ?

Table 15: A sample Hindi output with its reference from Urdu input with diacritics

Hindi reference	Hindi output from the Urdu text without diacritics M6-Hindi SLM+WLM-Tuned-No-Reordering
भारतीय सांस्कृतिक परंपराएँ कपिला वात्स्यायन कलाओं के क्षेत्र में सृजनात्मक कार्य प्रश्न बहन जी ,आप का विद्याध्ययन प्रतिभासूचक रहा है और उस के बाद भारत सरकार के संस्कृतिविभाग के अध्यक्ष के नाते भी आप का कार्य उच्चकोटि का रहा है ; इस पद पर कार्य करते हुए आप पर भारत के अन्तर्राष्ट्रीय सांस्कृतिक सम्बन्धों और भारत के भीतर सांस्कृतिक नीति के निर्माण का दायित्व रहा है ,इस क्षेत्र में ऐसी भिन्नभिन्न चीज़ें आती हैं जैसे संग्रहालय ,पुरातत्व ,पुरालेख , ग्रंथालय और कलाएँ ; केन्द्रीय स्तर पर भी और राज्य स्तर पर भी । आप का सरकारी कार्य तो भारत और विदेशों में सुविदित रहा है पर भारत की प्रदर्शनकारी और सुघट्य प्लास्टिक कलाओं के वर्गीकरण और व्याख्या करने के क्षेत्र में	भारतीय सांसंकृतिक परंपराई कपिला वातंसंायन कलावं के कंषेतंर में सरज्ञातंमक कारंय पंरषंन बहन जी ,आप का विदंयाधंयेन पंरतिभासूचक रहा है और उस के बअद भारत सरकार के संसंकृतिविभाग के अधंेकंष के नाते भी आप का कारंय उच्चकोटी का रहा है ; इस पद पर कारंय करते हूए आप पर भारत के अनंतर्राषंटंरीया सांसंकृतिक समंबनंधों और भारत के भीतर सांसंकृतिक नीति के निरंमान का दायितंो रहा है ,इस कंषेतंर में ऐसी भिन्नभिन्न चीज़ें आती हैं जैसे संगंरालय , पुरातंो ,पुरालेख ,गंरंथाल्य और कलाईं ; केनंदंरीया संतर पर भी और राजंय संतर पर भी । आप का सरकारी कारंय तो भारत और विदेशों में सुविदित रहा है पर भारत की पंरदरंशंकारी और सुघटंय पंलासंटिक कलावं के वरंगीकरण और

आप के सृजनात्मक योगदान के बारे में कम से कम औसत स्तर के शिक्षित सामान्य व्यक्ति को कुछ अधिक जानकारी नहीं है । क्या आप हमारे पाठकों के लिए भारतीय कलाओं और सौन्दर्यशास्त्र के क्षेत्र में किये गये अपने सृजनात्मक और अन्वेषणात्मक कार्य का संक्षिप्त विवरण देने की कृपा करें गी ?	वंयाखंया करने के कंषेतंर में आप के सरज्ञातंमक योगदान के बारे में कम से कम औसत संतर के षिकंषित सामानंय वंेकंती को कुछ अधिक जानकारी नहीं है । कंया आप हमारे पाठकों के लिये भारतीय कलावं और सौनंदरंेासंतंर के कंषेतंर में किये गऽे अपने सरज्ञातंमक और अनंवेशनातंमक कारंय का सनकंषिपंत विवरन देने की कृपा करें गी ?

Table 16: A sample Hindi output with its reference from Urdu input without diacritics

In general, the sentence-level accuracy of an ST system is always between 5% to 10%. The reason behind this very low accuracy might be the training data. In our case, the training data is an Urdu – Hindi parallel lexicon and not a parallel corpus (a usual case for a classical SMT system). Unfortunately, we do not have any Hindi-Urdu parallel corpus to test our hypothesis that the accuracy of Hindi-Urdu transliteration can be improved by training the ST models with Hindi-Urdu parallel corpus instead of using a Hindi-Urdu parallel lexicon.

6 Conclusion

A mere transliteration between Urdu and Hindi can serve the purpose of translation between Urdu and Hindi. Urdu and Hindi are linguistically the same or almost the same languages because of their common grammatical structure, morphology and roughly 60 to 70% common vocabulary. Thus Urdu Hindi transliteration is very important for more than 1,000 million people on the globe.

In this paper we have reported our experiments on Urdu↔Hindi transliteration using Statistical Machine Translation (SMT) techniques and an Urdu – Hindi parallel lexical resource. We have performed total 576 experiments and have reported results of significant interest. From Hindi–to–Urdu transliteration, we have achieved the maximum word-level accuracy of 71.5%. From Urdu–to–Hindi transliteration, the maximum word-level accuracy is 77.8% when the input Urdu text contains all necessary diacritical marks and 77% when the input Urdu text does not contain all necessary diacritical marks. At character-level, transliteration accuracy is more

than 90% in both directions. At the sentence-level, the accuracy of an ST system is always between 5% to 10%. This is a very low in terms of readability and usability of a transliterated text. These results can be improved by building an Urdu Hindi parallel corpus and by building context sensitive transliteration models. We will show in future that a better high quality Urdu – Hindi transliteration system can be built by combining non-deterministic finite-state transliteration models and finite-state language models.

Acknowledgments
We are thankful to Mr. James Nye, project director of Digital South Asia Library (DSAL), University of Chicago for sharing their digital data of "A dictionary of Urdu, classical Hindi and English" with us.

References

AbdulJaleel, Nasreen and Larkey, L. S. 2003. Statistical Transliteration for English-Arabic Cross Language Information Retrieval. *12th international Conference on information and Knowledge Management (CIKM 03)*, pp. 139-46. New Orleans: ACM.

Al-Onaizan, Yaser and Knight, Kevin. 2002. Machine Transliteration of Names in Arabic Text. *Workshop on Computational Approaches To Semitic Languages, the 40th Annual Meeting of the ACL*, pp. 1-13. Philadelphia, Pennsylvania: ACL.

Al-Onaizan, Yaser and Papineni, Kishore. 2006. Distortion Models for Statistical Machine Translation. *21st International Conference on Computational Linguistics (COLING) and 44th Annual Meeting of the ACL*, pp. 529–36. Sydney, Australia.

Brown, Peter F., Cocke, John, Pietra, Stephen A. Della, Pietra, Vincent J. Della, Jelinek, Fredrick, Lafferty, John D., Mercer, Robert L. and Roossin, Paul S. 1990. A Statistical Approach to Machine Translation. *Computational Linguistics* 16(2): 79-85.

Brown, Peter F., Pietra, Stephen A. Della, Pietra, Vincent J. Della and Mercer, Robert L. 1993. The Mathematics of Statistical Machine Translation: parameter estimation. *Computational Linguistics* 19(2): 263-312.

Durrani, Nadir, Sajjad, Hassan, Fraser, Alexander and Schmid, Helmut. 2010. Hindi-to-Urdu Machine Translation Through Transliteration. *the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 465–74. Uppsala, Sweden.

Ekbal, Asif, Naskar, Sudip Kumar and Bandyopadhyay, Sivaji. 2006. A Modified Joint Source-channel Model for Transliteration. *21st International Conference on Computational Linguistics (COLING) and the 44th Annual Meeting of the ACL*, pp. 191-98. Sydney: ACL & ICCL.

Finch, Andrew and Sumita, Eiichiro. 2009. Transliteration by Bidirectional Statistical Machine Translation. *workshop on Named Entities (NEWS-09), Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing ACL/IJCNLP*, pp. 52-56. Singapore.

Fraser, Alexander and Marcu, Daniel. 2007. Measuring Word Alignment Quality for Statistical Machine Translation. *Computational Linguistics* 33(3): 293-303.

Huang, Fei. 2009. Confidence Measure for Word Alignment. *Joint Conference the 47th Annual Meeting of the ACL and the 4th IJCNLP*, pp. 932-40. Singapore.

Hussain, Sarmad. 2004. Letter-to-Sound Conversion for Urdu Text-to-Speech System. *Workshop on Computational Approaches to Arabic Scriptbased Languages, International Conference on Computational Linguistics (COLING)*, pp. 74-79. Geneva: ICCL.

Khan, Abdul Jamil. 2006. *Urdu/Hindi: an artificial divide*. New York: Algora Publishing.

Kirschenbaum, Amit and Wintner, Shuly. 2009. Lightly Supervised Transliteration for Machine Translation. *12th Conference of the European Chapter of the ACL*, pp. 433-41. Athens: ACL.

Knight, Kevin and Graehl, Jonathan. 1997. Machine Transliteration. *8th Conference of EACL*, pp. 128-35. Madrid: ACL.

Koehn, Philipp. 2010. *Statistiacal Machine Translation*. Cambridge University Press.

Koehn, Philipp, Hoang, Hieu, Birch, Alexandra, Callison-Burch, Chris, Federico, Marcello, Bertoldi, Nicola, Cowan, Brooke, Shen, Wade, Moran, Christine, Zens, Richard, Dyer, Chris, Bojar, Ondřej, Constantin, Alexandra and Herbst, Evan. 2007. Moses: open source toolkit for Statistical Machine Translation. *47th Association for Computational Linguistics 2007 Demo and Poster Sessions*, pp. 177-80. Prague.

Koehn, Philipp, Och, Franz Josef and Marcu, Daniel. 2003. Statistical Phrase-Based Translation. *Human Language Technology (HLT), NAACL-2003*, pp. 48-54. Edmonton, Canada.

Kumar, Shankar, Och, Franz and Macherey, Wolfgang. 2007. Improving Word Alignment with Bridge Languages. *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 42-50.

Lee, J. S. and Choi., K. S. 1998. English to Korean Statistical Transliteration for Information Retrieval. *Computer Processing of Oriental languages* 12(1): 17-37.

Lewis, M. Paul. 2009. Ethnologue: Languages of the World. In M. P. Lewis (ed.). Dallas: SIL International.

Li, Haizhou, Zhang, Min and Su, Jian. 2004. A Joint Source-channel Model for Machine Transliteration. *42nd Annual Meeting on ACL*, Barcelona: ACL.

Lopez, Adam. 2008. Statistical Machine Translation. *ACM Computing Surveys* 40(3).

Malik, M. G. Abbas, Besacier, Laurent, Boitet, Christian and Bhattacharyya, Pushpak. 2009. A Hybrid Model for Urdu Hindi Transliteration. *Joint conference of the 47th Annual Meeting of the Association of Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of NLP ACL/IJCNLP Workshop on Named Entities (NEWS-09)*, pp. 177–85. Singapore.

Malik, M. G. Abbas, Boitet, Christian and Bhattacharyya, Pushpak. 2008. Hindi Urdu Machine Transliteration using Finite-state Transducers. *22nd International Conference on Computational Linguistics (COLING)*, pp. 537-44. Manchester: ICCL.

Nabende, Peter. 2009. Transliteration System using pair HMM with Weighted FSTs. *Joint conference of the 47th Annual Meeting of the Association of Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of NLP ACL/IJCNLP Workshop on Named Entities (NEWS-09)*, pp. 100–03. Singapore: IJCNLP/ACL.

Och, Franz Josef and Ney, Hermann. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics* 29(1): 19-51.

Och, Franz Josef and Ney, Hermann. 2004. The Alignment Template Approach to Statistical Machine Translation. *Computational Linguistics* 30(4): 417-49.

Platts, John T. 1884. A Dictionary of Urdu, Classical Hindi and Englsih. London: W. H. Allen & Co.

Platts, John T. 1909. *A Grammar of the Hindustani or Urdu Language*. Crosby Lockwood and Son.

Rabinder, Lawrence R. 1989. A tutorial on Hidden Markon Models and selected applications in speech recognition. *IEEE*, pp. 257-85.

Rabiner, L. and Juang, B. H. 1986. An Introduction to Hidden Markov Models. *IEEE ASSP Magazine* 3(1): 4-16.

Rahman, T. 2004. Language Policy and Localization in Pakistan: Proposal for a Paradigmatic Shift. *Crossing the Digital Divide, SCALLA Conference on Computational Linguistics*, Katmandu.

Rai, Alok. 2000. *Hindi Nationalism*. New Delhi: Orient Longman Private Limited.

Stall, B. and Knight, K. 1998. Translating Names and Technical Terms in Arabic Text. *Workshop on Computational Approaches to Semitic Languages, COLING/ACL*, pp. 34-41. Montreal.

Zens, Richard and Ney, Hermann. 2003. A Comparative Study on Reordering Constraints in Statistical Machine Translation. *41st Annual Meeting on Association for Computational Linguistics*, pp. 144-51. Sapporo, Japan.

Zia, Khaver. 1999. Standard Code Table for Urdu. *4th Symposium on Multilingual Information Processing (MLIT-4)*, Yangon: CICC.