Hindi-Punjabi Machine Transliteration System (For Machine Translation System)

Vishal Goyal[1], Gurpreet Singh Lehal[2]
[1] Lecturer, Department of Computer Science, Punjabi University, Patiala
[2]Professor and Head, Department of Computer Science, Punjabi University, Patiala
[1]vishal.pup@gmail.com, [2]gslehal@gmail.com

Abstract

Transliteration is a process that takes a character string in a source language and generates equivalent mapped character string in the target language. One of the most frequent problems translators must deal with is translating proper names and technical terms. Such terms are basically not translated rather are transliterated. In our paper, we have taken Hindi as source language and Punjabi as target language. Thus, Hindi words will be transliterated into Punjabi words. Hindi and Punjabi are closely related languages and hence it is comparatively easy to develop than the system between very different language pairs like Hindi and English. We have implemented approximately fifty complex rules for making the transliteration between Hindi-Punjabi language pair accurate after studying both the languages in details. Then regrous testing was done by test data covering number of domains like medicine, proper names, city names, country names, castes, surnames, rivers, subject related technical terms etc. The system found to give accuracy of about 98%..

1. Introduction

Hindi and Punjabi are closely related languages with lots of similarities in syntax and vocabulary Both Punjabi and Hindi languages have originated from Sanskrit which is one of the oldest language. Hindi and Punjabi belong to the same subgroup of the Indo-European family i.e. Indo-Aryan family of the languages. In India, the script of Hindi language is Devanagari and the script of Punjabi Language is Gurmukhi. Hindi is one of the most widely spoken languages of the world, possessing speakers of the same order of magnitude as those of English and Russian. In India it has been accorded the status of 'official language' by the central govt. for use for most administrative purposes, and Punjabi being the official language of state Punjab and has been accorded the status of 'official language' by the Punjab government for use for most administrative purposes. Hindi-Punjabi Machine Transliteration system is not only an attempt to undo scriptural divide between Hindi and Punjabi but also an attempt to providean innovative base for the future natural language processing work on both Hindi and Punjabi.

2. Hindi Language

Hindi Language is written in Devanagari Script. Many other languages are using the Devanagari scripts like Sanskrit, Nepal, Marathi etc. In Hindi Language, there are thirty three basic consonantal signs, twelve short vowels, ten full form vowels, two vowel signs are used for nasalization, one punctuation marks (।) is used as a full stop, abbreviations are formed by the use of either a small circle (∘) or a dot after the first syllable of the word, doubled consonant cluster, gemination is written by writing the first component of the consonant cluster as the truncate form of the consonant (which is frequently built from the independent version of the latter consonant by the deletion of the vertical bar that appears on the right side of many Devanagri character) and the second component of the consonant cluster is, the unaltered full symbol for the second consonant.

3. Punjabi Language

Punjabi Language is written in Gurmukhi Script. In Punjabi Language, there are thirty five basic consonants, no dead consonants like in Hindi Language, three types of conjunct consonants, gemination is written by the sign ੱ (addak) above and before the consonant to be doubled, twelve short form vowels, ten full form vowels, two vowels for nasalization, viram (।) is used for end of the sentence, sign : is used to mark abbreviation.

5. Existing Transliteration Systems

5.1 Shahmukhi to Gurmukhi Transliteration System: A Corpus based Approach

This System is developed by Dr. Gurpreet Singh Lehal and Tajinder Singh Saini at Advanced Centre for Technical Development of Punjabi Language, Literature & Culture, Punjabi University, Patiala. It is a corpus based transliteration system for Punjabi language. The existence of two scripts for Punjabi language has created a script barrier between the Punjabi literature written in India and in Pakistan. The corpus analysis program has been run on both Shahmukhi and Gurmukhi corpora for generating statistical data for different types like character, word and n-gram frequencies. Potentially, all members of the substantial Punjabi community will be benefited vastly from this transliteration system. In this system, first of all script mappings are done. In the mappings process, mappings of Simple Consonants, Aspirated Consonants (AC), Vowels, other Diacritical Marks or Symbols are done. After doing the mappings, the transliteration system is virtually divided into two phases. The first phase performs pre-processing and rule-based transliteration tasks and the second phase performs the task of post-processing. In the post-processing phase bi-gram language model has been used. In pre-processing stage Shahmukhi token is searched in the Shahmukhi-Gurmukhi dictionary before performing rule-based transliteration. If the token is found, then the dictionary component will return a weighted set of phonetically similar Gurmukhi tokens and those will be passed on to the bi-gram queue manager. In the post-processing stage, the first task is to perform formatting of the Gurmukhi token according to Unicode standards. The second task in this phase is critical and especially designed to enable this system to work smoothly on Shahmukhi script having missing diacritical marks.

5.2.2Hindi Urdu Machine Transliteration System

This system is developed by M.G. Abbas Malik in department of Linguistics, University of Paris 7-Denis Diderot, Paris, France. Hindi Urdu Machine transliteration System is not only an effort to undo the scriptural divide between Hindi and Urdu; also it is an attempt to provide an innovative base for future natural language processing work on both Hindi and Urdu. Hindi and Urdu are more or less the same languages. They share a huge vocabulary, morphological rules, grammar, semantics etc. For transliteration from Urdu to Hindi and vice versa, first Hindi and Urdu will be converted into a common language that may be ASCII encodings of Hindi and Urdu characters or the International Phonetic alphabet (IPA) equivalents of Hindi and Urdu Phonemes. Secondly, Hindi and Urdu will be generated from the common language .This common language can also be used in more sophisticated future tools like morphological analyzer, grammar checker, spell checker, Machine transliteration, etc. for Hindi and Urdu.

5.2 A Roman-Gurmukhi Transliteration System

This system is developed by Rajesh Kumar Verma, in the Department of computer Science, Punjabi University, Patiala. In this system, survey of existing Roman-Indic script transliteration techniques is done. After doing survey, a transliteration scheme based on ISO: 15919 transliteration and ALA-LC is developed. According to linguistics, these are closer to the natural pronunciation of Punjabi words as compared with others. Most of the rules for transliteration in both schemes were same except for Bindi and tippi in case of vowels as compared with consonants.Some modifications are done like, bindi and tippi will now be represented with the same symbol because both produce similar sounds and will be transliterated in the same way.

5.3 xNagari:The Scheme for Transliteration from DevaNagari to Latin

This scheme is lossless; i.e. the DevaNagari text trans-liberated to Roman-Latin text using this scheme can be exactly transliterated back to DevaNagari. Most encodings (including Unicode, ISCII) support at least basic Latin alphabet and punctuation marks.

5.2.5 INSROT – Indian Script to Roman Transliteration

Indian Scripts are phonetic and have similarity in alphabetic correspondence. Transliteration between Indian languages is simple, unambiguous and phonetically similar. However, there is need for a Scheme for transliteration from Indian Scripts to Roman Script. People in India have been using many different schemes for Hindi- Roman transliteration. ITRANS is one of the popular schemes. ITRANS provides choice and uses capital letters also. The present scheme is not suitable for searching the contents on the Internet. There is a need to evolve a Case-insensitive transliteration scheme to facilitate searching on web. INSROT (Indian Script

Roman Transliteration) Table is proposed as a standard transliteration scheme.
It is worth mentioning that there is no transliteration system developed so far Hindi to Punjabi Language Pair.

5. Direct Character Mappings

In this section, direct mapping of Hindi Consonants and Vowels into Punjabi Consonants and Vowels is explained. It is the base of the transliteration process.

अ	ਅ	ड	ड	ਾੋ	ੌ
आ	ਆ	ड़	ੜ	ੀ	ੰ
इ	ਇ	ढ	ढ	ਿ	ਿ
ई	ਈ	ढ़	ੜ੍	ੀ	ੀ
उ	ਉ	ण	ਣ	ਾੱ	ਾ
ऊ	ਊ	त	ਤ	0	0
ऐ	ऐ	थ	ਥ	०	੦
ओ	ਓ	द	ਦ	1	੧
औ	ਔ	न	ਨ	१	੧
क	ਕ	प	ਪ	2	੨
क़	ਕ	फ	ਫ	२	੨
ख	ਖ	फ़	ਫ਼	3	੩
ख़	ਖ਼	ब	ਬ	4	੪
ग	ਗ	भ	ਭ	४	੪
ग़	ਗ਼	म	ਮ	5	੫
घ	ਘ	य	ਯ	५	੫
ङ	ਙ	र	ਰ	6	੬
च	ਚ	ल	ਲ	६	੬
छ	ਛ	व	ਵ	7	੭
ज	ਜ	श	ਸ਼	७	੭
ज़	ਜ਼	ष	ਸ਼	8	੮
झ	ਝ	स	ਸ	८	੮
ञ	ਞ	ं	ं	9	੯
ट	ਟ	ੂ	ੁ	९	੯
ठ	ਠ	ਾ	ਾ	ः	:
ु	ੁ	ੋ	ੈ	े	ै
ै	ै				

6. Complex Rules for transliteration

Following are rules that have been implemented in addition to direct mapping explained in the previous section. It is to mention that some rules has been developed such that they can be helpful during the translation process. This means that while translating any text from Hindi to Punjabi Text, there are some words that are not being translated then they will transliterated as they have been translated by following the similarity between hindi and Punjabi language inflection rules.

1. If very first character in the word is य, it will be transliterated into ਯ .For example यशोदा will be transliterated to ਯਸ਼ੋਦਾ.

2. If current character is fourth last character in the word and it is ੂ matra, next character is ਗ , next character is ਾ and next to this is ਂ, the current character i.e. ੂ will be transliterated to ਾ. For example: खपूगां will be transliterated to ਖਪਾਗਾਂ.

3. If current character is fourth last character in the word and it is ੂ matra, next to it is ं , next to it is ग then ਾ matra, current character ੂ will be transliterated into ਾ. For example: मारुंगा will be transliterated into ਮਾਰਾਂਗਾ.

4. If current character is 'ऊ' and is at fourth last position in the word, it will be transliterated into ਵਾ For example खाऊंगा will be transliterated into ਖਾਵਾਂਗਾ

5. If at the end of the word, यों or यें or यां or याँ substring is present, then this substring will be transliterated into ਆਂ For example बेटीयां will be transliterated to ਬੇਟੀਆਂ.

6. If second last character is य, and the last character is matra ा, and before य there are more than one character, या will be transliterated into ਾ. For example रिया will be transliterated to ਰਿਆ.

Rule 7:-If second last character is current character and it is character य and matra after य is े, it will be transliterated into ਏ. For Example अक्षय will be transliterated into ਅਕਸ਼ਏ.

Rule 8:-If second last character is current character and it is matra ो then ं, it will be transliterated into ਿ+ਅ+ ਾ +ਂ. For example नदियों will be transliterated into ਨਦੀਆਂ

"Testing for last character in word"

Rule 9:-If last character is matra ो, it will be transliterated into ਾ and ं. For example चीजों will be transliterated into ਚੀਜਾਂ

Rule 10:-If last character is character य, it will be transliterated into ਏ . For Example अक्षय will be transliterated into ਅਕਸ਼ਏ.

"Half words"

Rule 11:-If there is Half Word and character य and then ा matra, it will be converted into ਿ+ Full Word + ਅ+ਾ. . For Example प्यास will be transliterated into ਪਿਆਸ.

Rule 12:- If there is Half Word and character य and ू matra, it will be transliterated into ਿ + Full Word+ ੳ + ੂ matra. For Example अभिमन्यू will be transliterated to ਅਭਿਮਨਿਊ

Rule 13:-If there is Half Word and character य and ु matra, it will be transliterated into ਿ + Full Word+ ੳ + ੁ matra. For Example मृत्यु will be transliterated into ਮ੍ਰਿਤਉ

Rule 14:-If there is Half Word and character य and matra ो or ौ, it will be transliterated into ਿ +Full Word+ ੳ + ੁ matra. For Example ज्यादती will be transliterated into ਜਿਉਤੀ

Rule 15:-If there is Half Word and character य and े matra, it will be transliterated into ਿ +Full Word+ े. For Example प्रत्येक will be transliterated into ਪ੍ਰਤਿਏਕ

Rule 16:-If there is Half Word and character य and ै matra, it will be transliterated into + ਿ + Full Word + ਐ.

Rule 17:-If there is Half Word and character य and े matra, it will be transliterated into ਿ + Full Word + ਅ.

Rule 18:-If there is Half Word and character य and 'No' matra, it will be transliterated into ਿ +Full Word+ ਅ.

Rule 19:- If within word there is य and next character is matra ा, it will be transliterated into ਇ + ਆ. For Example उठाया will be transliterated into ਉਠਾਇਆ

Rule 20:-If within word there is य and next character is े matra, it will be transliterated into ਏ. For Example आयेगा will be transliterated into ਆਏਗਾ

Rule 21:-If within word there is य and next character is matra ौ ,it will be transliterated into ਿ and ੳ. For Example व्यौरा will be transliterated into ਬਿਉਰਾ

Rule 22:-If there is ਾ +ਨ +े at the end of word, it will be transliterated into ਾ + ਣ. For Example बनाने will be transliterated into ਬਨਾਣ

Rule 23:-If there is ੌ + ਨ +े at the end of word, it will be transliterated into ਿ +ਣ +ਆ. For Example खिलौने will be transliterated into ਖਿਲੌਣਿਆ.

Rule 24:-If there is ਨ+ े at the end, it will be transliterated into ਣ. For Example खाने will be transliterated into ਖਾਣ

Rule 25:-If there is ਤ + े at the end, it will be transliterated into ਤ + े. For Example □ते will be transliterated into ਪੱਤਾ

Rule 26:-If there is ਤ +ੀ at the end, it will be transliterated into ਤ + ੀ . For Example जयन्ती will be transliterated into ਜੈਂਤੀ.

Rule 27:-If current character is matra ਾ and next character is ਯ and next character is ਿ, it will be transliterated into ਇ. For Example रसायनिक will be transliterated into ਰਸਾਇਣਿਕ

Rule 28:- If current character is matra ਾ and next character is ਯ and next character is ੀ, it will be transliterated into ਈ. For Example अनुयायी will be transliterated into ਅਨੁਯਾਈ.

Rule 29:- If current character is matra ਾ and next character is ਯ and next character is matra ੋ, it will be transliterated into ਾ + ਯ + ੋ matra. For Example आयोजन will be transliterated into ਆਯੋਜਨ

Rule 30:-If current character is matra ਾ and next character is ਯ and next character is matra ਾ, it will be transliterated into ਾ matra + ਇ + ਆ. For Example आया will be transliterated into ਆਇਆ

Rule 31:-If current character is matra ਾ and next character is ਯ and next character is matra े,it will be transliterated into ਾ matra + ਏ. For Example मनायेगा will be transliterated into ਮਨਾਏਗਾ

Rule 32:-If current character is matra ਾ and next character is ਯ and next character is any other character, it will be transliterated into ਾ matra + ਏ. For Example अध्याय will be transliterated into ਅਧਿਆਏ

Rule 33: If current character is ਿ and next character is ਯ and next character is ਿ it will be transliterated into ਇ.

Rule 34:-If current character is ਿ and next character is ਯ and next character is matra ਾ or ੋ or े, it will be transliterated into ਆ.

Rule 35:-If current character is ਿ and next character is ਯ and next is any other character it will be transliterated into ਿ and ਅ.

Rule 36:-If current character is matra ੈ and next character is ਯ and next character is matra ਾ, it will be transliterated into ਿ + ਆ. For Example तैयार will be transliterated into ਤਿਆਰ

Rule 37:-If current character is ं and next character is ਯ, it will be transliterated into ਿ + ੰ + ਅ.

Rule 38:-If current character is ਯ and next character is ं, it will be transliterated into ਿ + ਅ+ ੰ.

Rule 39:-If there is half ਰ and full ਯ, it will be transliterated into ਰ +ਜ. For Example कार्य will be transliterated into ਕਾਰਜ

Rule 40:-If there is half ਕ and full ਖ, it will be transliterated into ੱ. For example मक्खन will be transliterated into ਮੱਖਣ

Rule 41:-If there is half ਚ and full ਛ, it will be transliterated into ੱ. For example इच्छा will be transliterated into ਇੱਛਾ.

Rule 42:-If there is half ਟ and full ਠ, it will be transliterated into ੱ. For example मट्ठी will be transliterated into ਮੱਠੀ

Rule 43:-If there is half ਗ and full ਘ, it will be transliterated into ੱ. For example मग्घर will be transliterated into ਮੱਘਰ

Rule 44:-If there is half ज and full झ, it will be transliterated into ੱ. For example निज्झर will be transliterated into ਨਿੱਝਰ

Rule 45:-If there is half त and full थ, it will be transliterated into ੱ. For example □त्थर will be transliterated into ਪੱਥਰ

Rule 46:-If there is half द and full ध, it will be transliterated into ੱ. For example सिद्धार्थ will be transliterated into ਸਿੱਧਾਰਥ

Rule 47:-If there is half प and full फ, it will be transliterated into ੱ.

Rule 48:-If there is half ब and full भ, it will be transliterated into ੱ.

Rule 49:-If there is half ड and full ढ, it will be transliterated into ੱ.

Rule 50:-If there is half न and full न, it will be transliterated into ੰ. For example □न्ना will be transliterated into ਪੰਨਾ

Rule 51:-If there is Halant (੍) then don't ignore it, if next character is र like premee, it will be transliterated into ੍. For example प्रेमी will be transliterated into ਪ੍ਰੇਮੀ

Rule 52:-If there is half न or half म or half ण, it will be transliterated into ੰ. For example **हिन्दी** will be transliterated into ਹਿੰਦੀ

Rule 53:-If there is Devnagari ज्ञ it will be transliterated in Gurmukhi into ਿ +ਗ + ਅ.

Rule 54:-If there is halant (੍) in between and around halant (੍) previous character is half and after halant (੍) character is full but around halant (੍) character are same, it will be transliterated into ੰ.

Rule 55:-If there is half द, it will be transliterated into ੱ.

"Ignore Halant"

Rule 56:-If there is ੰ, it will be transliterated into ੰ or ੱ. For example कंचन will be transliterated into ਕੰਚਨ.

Rule 57:-If there is ੱ, it will be transliterated into ੰ or ੱ. For example चांद will be transliterated into ਚੰਦ

6. Design and Implementation

The Hindi-Punjabi Transliteration system has been developed using ASP.net and MS-Access. The system will be openly available online for use. First the various rules are checked for every Hindi word, If rule is matched with the character combinations present in the word, rule is applied and at the end, direct mapping is applied for rest of the characters.

7. Results and Discussion

The system has been tested thoroughly using test cases designed for number of domains like proper names, City names, country names, river names, fruit names, color names, day names, computer technical terms, medicine related technical terms, newspaper news related terms, literature, sports, other subjects technical terms. Approx. 100,000 words were tested and all were transliterated accurately. Following sample of text shows the results given by the system:

अम्बर	ਅੰਬਰ	अमृत	ਅਮ੍ਰਿਤ
अरविन्द	ਅਰਵਿੰਦ	अश्विनी	ਅਸ਼ਵਿਨੀ
धर्मेन्द्र	ਧਰਮੇਂਦ੍	धनंजय	ਧਨੰਜਏ
एकलव्य	ਏਕਲਵਿਅ	घनश्याम	ਘਨਸ਼ਿਆਮ
कार्तिकेय	ਕਾਰਤਿਕੇਏ	मुकुन्द	ਮੁਕੁੰਦ
निर्गुन	ਨਿਰਗੁਨ	प्रस्तुति	ਪ੍ਰਸਤੁਤਿ
संदर्भ	ਸੰਰਦਭ	कंप्यूटर	ਕੰਪਿਊਟਰ

8. Conclusion and Future Work

Being Hindi and Punjabi closely related language, Hindi-Punjabi transliteration system is very beneficial for removing the language and scriptural barrier. This system is giving promising results and this can be further used by the researchers working on Hindi and Punjabi Natural Language Processing tasks. Central govt. Hindi Documents, Hindi Literature and other documents in Hindi of one's interest can be transliterated into Punjabi for use on the click on a button.

9. References

[1] Bharati, Akshar, Vineet Chaitanya and Rajeev Sangal. (1995). Natural Language Processing: A Paninian Perspective, Prentice-Hall of India, New Delhi.

[2] Bharati, Akshar, Amba P. Kulkarni, Vineet Chaitanya. (1998a). Challenges in Developing Word Analyzers for Indian Languages, Presented at Workshop on Morphology, CIEFL, Hyderabad, July 1998.

[3] Bharati, Akshar, Rajeev Sangal and S.M. Bendre (1998b). Some Observations on Corpora of Some Indian Languages. Knowledge Based Computing Systems, Tata McGraw-Hill.

[4] Goldsmith, John. (2001). Unsupervised Learning of the Morphology of a Natural Language. Computational Linguistics, Vol 27, No. 2, pp 153-198.

[5] Daniel Jurafsky, James H. Martin.Speech and Language Processing: An introduction to speech recognition, natural language processing, and computational linguistics.

[6] LTRC, IIIT Hyderabad http://ltrc.iiit.ac.in

[7] Gill Mandeep Singh, Lehal Gurpreet Singh, Joshi S.S., A full form lexicon based Morphological Analysis and generation tool for Punjabi,International Journal of Cybernatics and Informatics, Hyderabad,India,October 2007, pp 38-47