Int. Journal of Engineering Research and Applications www.ijera.com
ISSN: 2248-9622, Vol. 5, Issue 11, (Part - 4) November 2015, pp.06-10

RESEARCH ARTICLE OPEN ACCESS

Punjabi to Hindi Transliteration System for Proper Nouns Using Hybrid Approach

Er.Sahil Malhan , Er.Jasdeep Mann
Department Of Compter Science And Engineering, Bhai Maha Singh College of Engineering,Sri Muktsar Sahib, PUNJAB

Abstract
The language is an effective medium for the communication that conveys the ideas and expression of the human mind. There are more than 5000 languages in the world for the communication. To know all these languages is not a solution for problems due to the language barrier in communication. In this multilingual world with the huge amount of information exchanged between various regions and in different languages in digitized format, it has become necessary to find an automated process to convert from one language to another. Natural Language Processing (NLP) is one of the hot areas of research that explores how computers can be utilizing to understand and manipulate natural language text or speech. In the Proposed system a Hybrid approach to transliterate the proper nouns from Punjabi to Hindi is developed. Hybrid approach in the proposed system is a combination of Direct Mapping, Rule based approach and Statistical Machine Translation approach (SMT). Proposed system is tested on various proper nouns from different domains and accuracy of the proposed system is very good.
Keywords: Rule based approach. Dictionary Lookup approach, Statistical Machine Translation approach (SMT), Transliteration, Natural Language Processing (NLP).

I. Introduction

Machine Translation is the process of enabling a computer to translate nouns or sentences from one language to another. There is a lot of research going on this area of machine translation at present. Many works are concentrated on translating Indian regional languages to Hindi. It is also important for many applications. In India and in foreign, most of the population is not so familiar with Punjabi. Where Hindi is a national language and one of the popular spoken languages in India. The translation of the Punjabi language into a commonly used language is essential for many applications like instant message systems and for all communication systems. The thesis proposes a translation system which translates Punjabi nouns into its corresponding Hindi word.

Transliteration and Translation are both different. Transliteration maps the letters of source script to letters of pronounced similarly in target script. Transliteration is particularly used to translate proper names and technical terms from languages. Transliteration is particularly used to translate proper names and technical terms from languages. Translation is the action of interpretation of the meaning of a text and subsequent production of an equivalent text. Translation is a process that communicates the same message in another language.

One instance of transliteration is the use of a Hindi computer keyboard to type in a language that uses a different alphabet, such as in Russian. While the first usage of the word implies seeking the best way to render foreign words into a particular language, the typing transliteration is a purely pragmatic process of inputting text in a particular language. Transliteration from Hindi letters is particularly important for users who are only familiar with Hindi keyboard layout, and hence could not type quickly in a different alphabet even if their software would actually support a keyboard layout for another language. If relation between sounds and letters are similar in both the languages, then transliteration may be same as transcription. There are also some mixed transliteration systems that transliterate a part of original script and transcribe the rest. Greeklish is an example of such a mixture. Punjabi language is written from left to right using Gurumukhi script and Punjabi language consist of consonants, vowels, halant, punctuation and numerals. Similarly Hindi language is written from left to right using Roman Script and Hindi language also consist of consonants, vowels, and punctuation. Transliteration between Punjabi and Hindi is required for proper names.

II. Approaches to MT

There are four approaches to Machine transliteration. These are discussed as follows:

III. Direct Based Approach

A direct based machine transliteration system carries out word by word translation with the help of

a bilingual dictionary, usually followed by some syntactic rearrangement.

IV. Rule Based Approach

A rule based machine transliteration system parses the source text and produces an intermediate representation, which may be a parse tree or some abstract representation.

V. Corpus Based Approach

Corpus based Machine transliteration requires sentence aligned parallel text for each language pair. The corpus based approach is further classified into statistical and example based machine translation approaches.

VI. Knowledge Based Approach

Early machine systems are characterized by the syntax. Semantic features are attached to the syntactic structures and semantic processing occurs only after syntactic processing. Semantic based approaches to language analysis have been introduced by AI researchers. The approached require a large knowledge base that includes both ontological and lexical knowledge.

VII. Statistical Machine Translation (SMT) Approach

The statistical Machine Translation (SMT) is part of corpus based machine translation.SMT requires less human effort to undertake translation.SMT is a machine transliteration paradigm where translation are generated on the basis of statistical models .These statistical models parameters are derived from the analysis of bilingual text corpora.

Literature Survey

Pankaj kumar and Er.Vinod kumar (2013) has developed Statistical Machine Translation based Punjabi to English Transliteration System for nouns.The process is performed into two parts Segmentation Phase in which words of the source language are segmented into units and the Assembly phase in which segmented characters are mapped to the characters of target language with the help of rules.The overall system works in two phases Training phase and Translation phase. The overall accuracy of the system comes out to be 97%.

Kamaldeep and Dr. Vishal Goyal (2011) have developed hybrid approach for Punjabi to English transliteration system. The system has been developed with the mixture of rule based and direct mapping approach. The system has three layers. First one is tokenization in which Punjabi words are taken as input and break into the individual characters. Second one is rule based and direct mapping approach in which rules are applied. Third one is compression with dictionary in which a dictionary is created that contains the spelling of names that are commonly used in real life. The overall accuracy of the system comes out to be 90% with rules and direct mapping and compare with dictionary approach comes with 95.23%.

Kamaljeet kaur batra and G S lehal (2010) has developed rule based machine translation of Punjabi to English. The system has analysis, translation and synthesis component. The steps involved are pre processing, tagging, ambiguity, resolution, translation and synthesis of words in target language. The accuracy is calculated for each step and the overall accuracy of the system is calculated to be about 85% for a particular type of noun phrases. After training the system with about 2000 phrases, testing is performed with 500 sentences and accuracy at different levels are calculated.

VIII. Proposed Methodology (Hybrid Approach)

There are various approaches with which transliteration can be performed from Gurumukhi Script to Devnagri Script. These techniques are as follows:

Direct Mapping:

This is a simplest approach to perform transliteration in which proper nouns in Gurumukhi Script along with their equivalent meaning in Hindi language. Direct mapping is performed from the database for the source noun.

The accuracy depends on the number of proper nouns stored in the database.

IX. Rule Based Approach

In this approach rule based approach various rules are created according to the phonetic equivalence of both source and target language. In this approach direct character to character mapping along with transliteration rules are used to obtain the results.

The accuracy of the rule based system is highly depends on the created rules for transliteration.

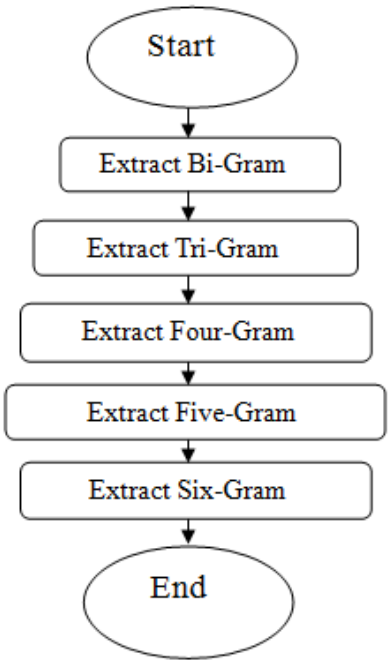
X. Statistical Machine Translation (SMT)

Statistical Machine translation system makes use of a parallel corpus of source and target language pairs. This parallel corpus is necessary requirement before undertaking training in Statistical Machine Translation. The system has used parallel corpus of Punjabi and Hindi names. A parallel corpus of 15000+ names has been developed. We use Statistical Machine translation System to transliterate

proper nouns written in the Gurumukhi script. The system is implemented with .Net using C# language. The transliteration system implementation goes through two stages. First is Training phase and other one is Transliteration phase.

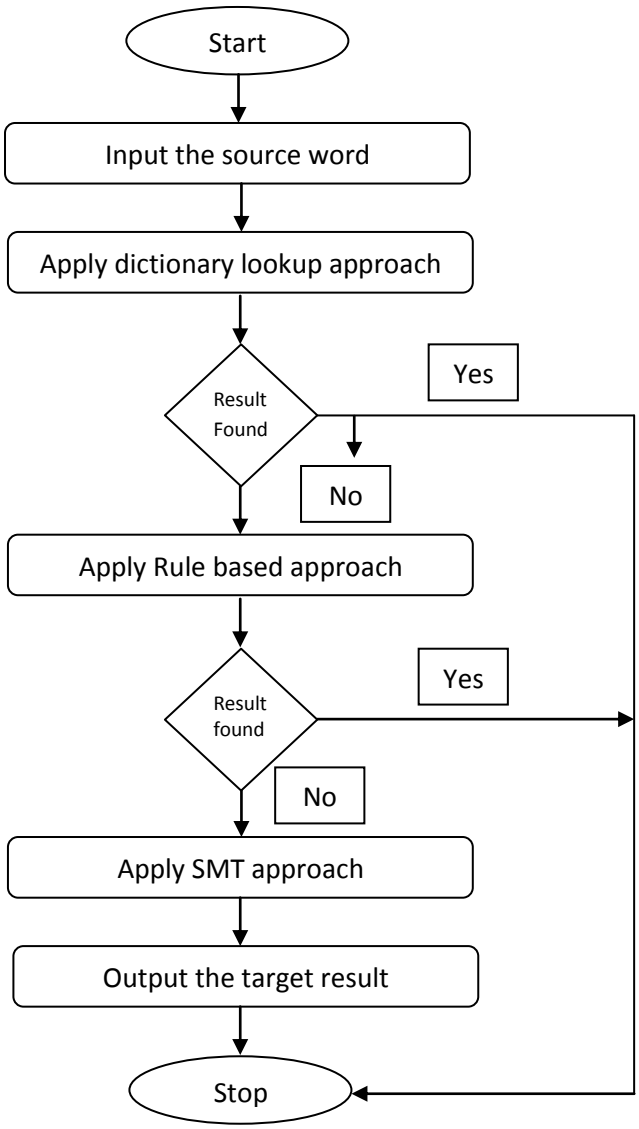
XI. Training phase

In system training phase training is given to the system on the basis of names stored into the database and generate the tables as shown above we store more than 15000 names to which the system is trained. We use algorithms to give training to the system. Training is given with the help of names stored in the database for different tables. In Enhanced N-Gram approach we have used unigram, bi-gram, tri-gram, four-gram, five-gram and six-gram for transliteration purpose. We store the unigrams in database manually and system is proposed to extract bi-gram, tri-gram, four-gram, five-gram and six-gram and then store it into the database along with their frequency of occurrence. These all grams are then stored into the database for transliteration purpose.



XII. Transliteration phase

In transliteration phase system tries to find the word directly into the database and if word is found then system gives output otherwise with the help of above generated tables system can transliterate new word..In N-Gram approach we have used uni-gram, bi-gram, tri-gram, four-gram, five-gram, six-gram for transliteration purpose. We store the uni-grams in database manually and system is proposed to extract bi-gram, tri-gram, four-gram, five-gram and six-gram and then store it into the database. When system try to transliterate a new word, it first try to find the same word from the database and if proper name is not in the database then with the help of bi-gram, tri-gram, four-gram, five-gram and six-gram created, system tries to transliterate the proper noun. A proper name is transliterated with the help of the seven tables mentioned in the training phase. In very first stage the system try to find the name directly from the tables which contain the names for training and if name found then transliteration is completed and result is given to the user but if name does not present in the database in direct manner then system split the proper name in various grams in all the possible combination and tries to find the combinations.



Proposed work use more than 15000 names in the database. Training is given to these names. Different types of tables are used for N-gram extraction. These N-gram tables contain 50000 N-gram entities.

Set	No. of Examples
Names Entity	16000+
N-Gram Extracted	50000+
Result Accuracy	94%

Statistics of dataset

XIII. Evaluation Metrics

In this thesis manual evaluation has been done on the bases of following parameters.

XIV. Transliteration Accuracy Rate

TAR (Transliteration Accuracy Rate) is used for evaluation to capture the performance at word level. Accuracy Rate is the percentage of correct transliteration from the total generated transliterations by the system.

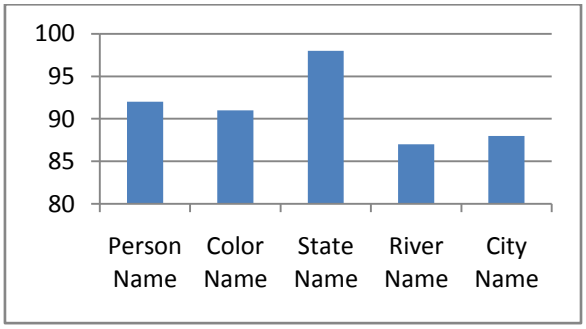
$$\text{Accuracy Rate} = \frac{\text{Number of correct Transliteration}}{\text{Total No. of Generated Transliteration}} * 100$$

XV. Results

We have 16000+ entries in our database for Punjabi words. And we have tested our software on 3000 Punjabi names. Test cases developed and their results are shown in following sections. The system has been test thoroughly using test cases designed for number of various domains like proper names, City names, State names, color names etc. The system has been tested to convert a doc file containing Punjabi names into its Hindi equivalent. The system has option to upload doc file containing Punjabi names. It transliterates those names into Hindi. System is tested by taking names from magazines, telephone directories, newspaper etc. The system is also tested on different color names and state names.TAR for color names, State names and Person names shown below:

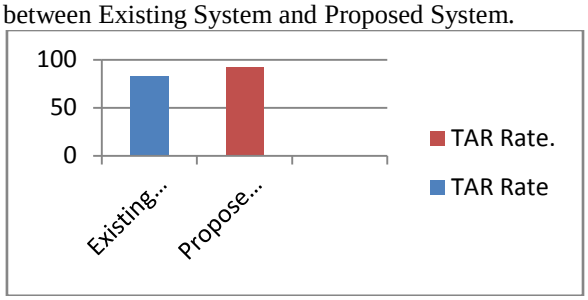
Domains	TAR
Person Name	92%
Color Name	91%
State Name	96%
River Name	87%
City Name	88%

TARS for Domains



Graph for TAR for Different Domains

The following graph shows the Comparison between Existing System and Proposed System.



TAR for Existing and Proposed System

Results Generated by our system

Punjabi	Hindi
ਨਵਪ੍ਰੀਤ	नवप्रीत
ਅਮਰਪ੍ਰੀਤ	अमरप्रीत
ਅਮਨ	अमन
ਨਾਜੀਆ	नाजीआ
ਬਠਿੰਡਾ	बठिंडा
ਪਟਿਆਲਾ	पटियाला
ਬੰਗਾਲ	बंगाल
ਜੰਮੂ	जम्मू
ਕਸ਼ਮੀਰ	कश्मीर
ਲਾਲ	लाल
ਨੀਲਾ	नीला
ਪੀਲਾ	पीला

XVI. Conclusion

In this thesis, Punjabi to Hindi machine transliteration system has been developed. The SMT is a part of corpus based MT system which requires parallel corpus. Before undertaking transliteration parallel corpus of 15000 Punjabi and Hindi noun was used to train the system. The SMT system developed accepts Punjabi as input and generates corresponding transliteration in Hindi. The translation was tested on different domains like person name, city name, and state name etc. using human evaluation method.

On the parameter of Transliteration accuracy rate the overall accuracy of the system comes 94%.The quality of the transliterated noun can be depends upon the size of corpus.

XVII. Future Scope

This system is giving promising results and this can be further used by the researchers working on Punjabi and Hindi Natural Language Processing tasks. State govt. Punjabi Documents, Punjabi Literature and other documents in Punjabi of one's interest can be transliterated into Hindi for use on the click on a button.

There can be following futures directions for Punjabi to Hindi SMT system.

- The work can be extended to include multilingual corpus of different languages in the source-target pair. The target and source languages can be increased from present one language.
- The system can also be put in the web based portal to translate content of one web page in Punjabi to Hindi.
- A mobile application can also be developed in which message containing Punjabi noun is sent to the client in Hindi language.
- The corpus can be processed to change its clause structure for improving quality of transliteration.

References

[1] Vijaya,VP, Shivapratap and KP CEN(2009) "*English to Tamil Transliteration using WEKA system"* International Journal of Recent Trends in Engineering, May 2009, Vol. 1, No. 1, pp: 498-500.

[2] Haque,Dandapat,Srivastava,Naskar and Way(2009) "*English—Hindi Transliteration Using Context-Informed PB-SMT: the DCU System for NEWS 2009"* Proceedings of the 2009 Named Entities Workshop, ACL-IJCNLP 2009, pages 104–107,Suntec, Singapore, 7 August 2009. ACL and AFNLP.

[3] Jia, Zhu, and Yu(2009), "*Noisy Channel Model for Grapheme-based Machine Transliteration*", Proceedings of the 2009 Named Entities Workshop, ACL-IJCNLP 2009, pages 88–91.

[4] Lehal and Singh (2008) *"Shahmukhi to Gurumukhi Transliteration System: A Corpus based Approach"* proceeding of Advanced Centre for Technical Development of Punjabi Language, Literature & Culture,Punjabi University, Patiala 147 002, Punjab, India, pp-151-162.

[5] Malik (2006) *"Punjabi Machine Transliteration System":* In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL (2006) 1137-1144.

[6] Verma(2006) *"A Roman-Gurumukhi Transliteration system"* proceeding of the Department of Computer Science, Punjabi University, Patiala, 2006.

[7] UzZaman , Zaheenand ,Khan(2009) "*A Comprehensive Roman (English)-To-Bangla Transliteration Scheme"* A Comprehensive Roman (English) to Bangla Transliteration Scheme, Proc. International Conference on Computer Processing on Bangla (ICCPB-2006), 17 February, 2006, Dhaka, Bangladesh.

[8] Knight, Graehl (2005) *"English-Japanese Transliteration system"*Computational Linguistics, Volume 24,Number 4, pp.599-612.

[9] Sato (2009) "*Web-Based Transliteration of Person Names"* IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology –Workshops, pp-273-278

[10] Malik, Besacier, Boitet, Bhattacharyya(2009) *"A Hybrid Model for Urdu Hindi Transliteration"* Proceedings of the 2009 Named Entities Workshop, ACL-IJCNLP 2009, pages 177–185,Suntec, Singapore, 7 August 2009ACL and AFNLP.

[11] Ali and Ijaz(2009), "*English to Urdu Transliteration System",* Proceedings of the Conference on Language & Technology 2009., pp: 15-23.