



A Review of Transliteration system from English to Punjabi

Devinder Brar *

Department of Computer Science and Engineering,
BFCET Bathinda, India

Er. Rishamjot Kaur

Department of Information and Technology,
BFCET Bathinda, India

Abstract: *The most commonly faced problem with translators is to translating proper names and technical terms. For some language pairs there is not much difference between their transliterated forms. For example in Spanish/English, this presents no great challenge: a phrase like Antonio Gil usually gets translated as Antonio Gil. However, the condition is different and more complicated for language pairs that employ very different alphabets and sound systems, such as Punjabi/Hindi. Phonetic translation across these pairs is called Transliteration. Transliteration attempts to be lossless, so that an informed reader should be able to reconstruct the original spelling of unknown transliterated words. To achieve this goal transliteration may define complex conventions for dealing with letters in a source script which do not correspond with letters in a goal script. Transliteration and transcription are opposite to each other. Transcription is which maps the sounds of one language to script of another language. 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. Translation is the action of interpretation of the meaning of a text and subsequent production of an equivalent text also called a translation that communicates the same message in another language.*

Keywords: *Translators, transliteration, translating, Antonio Gil, transcription.*

I. INTRODUCTION

A) NLP

Natural Language Processing holds great promise for making computer interfaces that are easier to use for people, since people will (hopefully) be able to talk to the computer in their own language, rather than learn a specialized language of computer commands. For programming, however, the necessity of a formal programming language for communicating with a computer has always been taken for granted. We would like to challenge this assumption. We believe that modern Natural Language Processing techniques can make possible the use of natural language to (at least partially) express programming ideas, thus drastically increasing the accessibility of programming to non-expert users. To demonstrate the feasibility of Natural Language Programming, this paper tackles what are perceived to be some of the hardest cases: steps and loops. We look at a corpus of English descriptions used as programming assignments, and develop some techniques for mapping linguistic constructs onto program structures, which we refer to as programmatic semantic.

B) Goal

The goal of NLP is “to accomplish human-like language processing”. The choice of the word ‘processing’ is very deliberate, and should not be replaced with ‘understanding’. For although the field of NLP was originally referred to as Natural Language Understanding (NLU) in the early days of AI, it is well agreed today that while the goal of NLP is true NLU, that goal has not yet been accomplished. A full NLU System would be able to:

- Paraphrase an input text
- Translate the text into another language
- Answer questions about the contents of the text
- Draw inferences from the text

While the entire field is referred to as Natural Language Processing, there are in fact two distinct focuses – language processing and language generation. The first of these refers to the analysis of language for the purpose of producing a meaningful representation, while the latter refers to the production of language from a representation. The task of Natural Language Processing is equivalent to the role of reader/listener, while the task of Natural Language Generation is that of the writer/speaker. While much of the theory and technology are shared by these two divisions, Natural Language Generation also requires a planning capability. That is, the generation system requires a plan or model of the goal of the interaction in order to decide what the system should generate at each point in an interaction.

C) Transliteration

The most commonly faced problem with translators is to translating proper names and technical terms. For some language pairs there is not much difference between their transliterated forms. For example in Spanish/English, this presents no great challenge: a phrase like Antonio Gil usually gets translated as Antonio Gil. However, the condition is different and more complicated for language pairs that employ very different alphabets and sound systems, such as Punjabi/Hindi. Phonetic translation across these pairs is called **Transliteration**. Transliteration attempts to be lossless, so that an informed reader should be able to reconstruct the original spelling of unknown transliterated words. To achieve this objective transliteration may define complex conventions for dealing with letters in a source script which do not correspond with letters in a goal script.

Transliteration and transcription are opposite to each other. Transcription is which maps the sounds of one language to script of another language. 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. Translation is the action of interpretation of the meaning of a text and subsequent production of an equivalent text also called a translation that communicates the same message in another language. Like Transcription and transliteration, 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 the process of converting a word written in one language into another language keeping its pronunciation same. Transliteration is not translation. It's research area belongs to NLP(Natural Language Processing).

For Example:

Name "Satnam" can be transliterated to "ਸੱਤਨਾਮ"

D) Type of Transliteration

Transliteration can be further divided into:

- 1) *Forward Transliteration*: Transliterating any word from source language to its target language is known as forward transliteration. -For Ex: "ARMAN" will be transliterated into "ਅਰਮਾਨ"
- 2) *Backward Transliteration*: Transliterating any word from target language to its source language is known as backward transliteration -For Ex: "ਅਰਮਾਨ" will be transliterated into "ARMAN"

E) Application of Transliteration

- Information is present in selected number of languages.
- Effective knowledge transfer across linguistics requires bringing down language barriers.
- Plays an important role in cross-lingual applications
- Can be used in various Government or Private Offices to transliterate nouns

-For Ex: ARMAN will be transliterated into ਅਰਮਾ

-Transliterating any word from target language to its source language is known as backward transliteration

-For Ex: ARMAN will be transliterated into ਅਰਮਾ

II. EXISTING APPROACHES

English to Punjabi transliteration can achieve by using various techniques. In transliteration there are following techniques:

- Direct mapping
- Rule based approach
- Statistical machine translation approach

A) Direct mapping Approach

When two languages are structurally similar and have similar vocabulary then direct approach is the best choice. The performance of a direct MT system depends on the quality and quantity of the source-target language dictionaries, text processing software, and word-by-word transliteration with minor grammatical adjustments on word order and morphology.

B) Rule-based Approach

The rule-based approach is the first strategy that was developed. A Rule-Based Machine Translation system consists of collection of rules, called grammar rules, a bilingual or multilingual lexicon, and software programs to process the rules.

C) Statistical machine translation approach

Statistical machine translation is a data-oriented statistical framework for translating text from one natural language to another based on the knowledge. During translation, the collected statistical information is used to find the best

translation for the input sentences, and this translation step is called the decoding process. There are three different statistical approaches in MT, Word-based Translation, Phrase-based Translation, and Hierarchical phrase based model. Statistical MT model take the view that every sentence in the target language is a translation of the source language sentence with some probability. The best translation, of course, is the sentence that has the highest probability.

The key problems in statistical MT are: estimating the probabilities of translation, and efficiently finding the sentence with the highest probability. The other problems include sentence alignment, compound words, idiom translation, morphology and out of vocabulary words.

III. PROBLEM DEFINATION

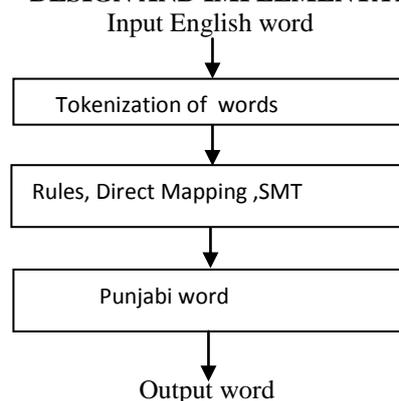
Term transliteration means to produce the results from source noun into target noun keeping its pronunciation same. Maximum accuracy of existing transliteration system is 63% which needs further improvement. N-Gram is used up to six-gram which has to extend to nine-gram to obtain accurate results. A web based system is required to transliterate proper nouns so that it can be used anywhere in the world.

Some of the names which are not correctly transliterated by "RULE BASED TRANSLITERATION SCHEME FOR ENGLISH TO PUNJABI" [Table1] is as follows:

TABLE 1
INCORRECT OUTPUT

Falkanvai	ਫਲਕੰਵਆ
Gowryharan	ਗੋਰੀਹਰਣ
Seerajuddeen	ਸੀਰਾਜੁਦਿਨ
Kolkandkar	ਕੋਲਕੰਡਕਰ

IV. DESIGN AND IMPLEMENTATION



V. CONCLUSION

After reviewing related papers we concluded that not much work is done for English to Punjabi transliteration. Further improvements can be done in this transliteration system from English to Punjabi. One of major weakness of transliteration from English to Punjabi is dealing with multiple mapped characters as discussed earlier. Multiple-mapping leads to some problems in transliteration process. This problem also affects the accuracy of this transliteration system.

REFERENCES

- [1] Deepti Bhalla , Nisheeth Joshi and Iti Mathur, "Rule based transliteration scheme for English to Punjabi " , International Journal on Natural Language Computing (IJNLC) Vol. 2, No.2, April 2013
- [2] Jasleen kaur and Gurpreet Singh josan, "Statistical Approach to Transliteration from English to Punjabi" International Journal on Computer Science and Engineering (IJCSSE)
- [3] Gurpreet Singh Josan and Gurpreet Singh Lehal, "A Punjabi to Hindi Machine Transliteration System" , Computational Linguistics and Chinese Language Processing Vol. 15, No. 2, June 2010, pp. 77-102..
- [4] Vishal GOYAL and Gurpreet SINGH LEHAL, " Evaluation of Hindi to Punjabi Machine Translation System" , IJCSI International Journal of Computer Science Issues, Vol. 4, No. 1, 2009 ISSN (Online): 1694-0784 .
- [5] Kamal Deep, Dr.Vishal Goyal, "Hybrid Approach for Punjabi to English Transliteration System", International Journal of Computer Applications (0975 – 8887) Volume 28– No.1, August 2011
- [6] Sumita Rani, Dr.Vijay laxmi , "A Review on Machine Transliteration of related languages Punjabi to Hindi " , International Journal of Science, Engineering and Technology Research (IJSETR) Volume 2, Issue 3, March 2013.
- [7] Sato (2009), "Web-Based Transliteration of Person Names", IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology – Workshops, pp-273-278.

- [8] 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.
- [9] 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.
- [10] 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.
- [11] Lehal and Singh (2008) ,“Shahmukhi to Gurmukhi 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.
- [12] 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.
- [13] Verma(2006), “A Roman-Gurumukhi Transliteration system”, proceeding of the Department of Computer Science, Punjabi University, Patiala, 2006.
- [14] 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.
- [15] Knight, Graehl (2005), “English-Japanese Transliteration system”, Computational Linguistics, Volume 24,Number 4, pp.599-612.
- [16] 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.
- [17] Hong, Kim, Lee and Chang(2009) ,“A Hybrid Approach to English-Korean Name Transliteration” , Proceedings of the 2009 Named Entities Workshop, ACL-IJCNLP 2009, pages 108–111,Suntec, Singapore, 7 August 2009 ACL and AFNLP.
- [18] Ali and Ijaz(2009), “English to Urdu Transliteration System”, Proceedings of the Conference on Language & Technology 2009., pp: 15-23.
- [19] Wei, Xu Bo(2008), “Chinese-English Transliteration Using Weighted Finite-state Transducers ” , ICALIP ,pp-1328 – 1333.