Architecture of the Transliteration System from ENGLISH to PUNJABI using Hybrid approach

Er. Sukhnandan Kaur
Pursuing M.Tech
From Swami Vivekanand
Engineering college, Punjab.

Ms. Rupinderdeep Kaur
Lecturer in department computer
Science & Engineering, Thapar
University, Patiala, Punjab

Er. Nidhi Bhalla
Lecturer in CSE department,
Swami Vivekanand
Engineering college, Punjab

Abstract: Current web is a huge repository, with its explosion in both the range and quantity of web content. Nowadays there are many transliteration systems available, where some are rule based and rest are based on statistical approach. This leads to an inappropriate transliteration. Keeping this in mind, through this paper we are presenting a system based on hybrid approach. This paper also addresses the actual process of the hybrid system used for transliteration.

Keywords: Architecture, Flow Diagram

1. INTRODUCTION:

Transliteration is an automatic method to generate characters or words in one alphabetical system for the corresponding characters in another alphabetical system. Transliteration is a process that takes a character string in source language as uage and their possible transliterated forms in target language. However, this is not a practical solution since proper nouns and technical terms, which are frequently transliterated, usually have rich productivity [1]. This paper discusses another approach based on machine learning to automate the process of machine transliteration.

2. METHODS OF TRANSLITERATION:

2.1 Rule Based Approach:

Transliteration based on the rules is the solution for appropriate transliteration. Such as a letter ‘a ’ can be transliterated as ‘ਾ’ as well as ‘ਅ’.

There are two levels in this:

1. Segmentation of the source string into transliteration units.
For example:

sapna
as
‘s’,’a’,’p’,’n’,’a’.

input and generates a character string in the target language as output. The process can be seen conceptually as two levels of decoding: segmentation of the source string into transliteration units and relating the source language transliteration units with units in the target language by resolving different combinations of alignments and unit mappings [2]. For example, consider word in source language ‘sapna’ which is segmented into source language transliterated units ‘s’ ’a’ ’p’ ’n’ ’a’, and then these units are transliterated into target language transliterated units ‘ਸ’, ’ਪ’, ’ਨ’ and ‘ਾ’ and finally these target transliterated units into final target language word ‘ਸਪਨਾ’. One possible method to generate transliteration is based on the use of dictionaries, which contains words in source language.

2. Relating the source language transliteration units with units in the target language.

‘s’,’a’,’p’,’n’,’a’.
As
‘ਸ’, ’ਪ’, ’ਨ’ and ‘ਾ’

3. Finally, these target transliterated units are combined to form a final target language word.

2.2 Statistical Approach

Statistical machine translation tries to generate translations using statistical methods based on bilingual text, it uses a dictionary for the transliteration. It reduces the error rate to great extent. The error rate can be reduced by providing maximum number of entries in the dictionary. But it is not always possible to put all the words especially in the dictionary.

2.3 Hybrid Approach

Hybrid machine transliteration strengths of statistical and rule-based translation methodologies. The approaches differ in a number of ways:

- **Rules post-processed by statistics**: Translations are performed using a rules based engine. Statistics are then used in an attempt to adjust/correct the output from the rules engine.

- **Statistics guided by rules**: Rules are used to pre-process data in an attempt to better guide the statistical engine. Rules are also used to post-process the statistical output to perform functions such as normalization. This approach has a lot more power, flexibility and control when translating.

3. DESIGN AND IMPLEMENTATION [3]

The system architecture, given below in Figure 1, consists of various stages through which source language text is passed and converted into target language. These different stages have its own process. They process the data and pass it to the next stage for further processing.

Basically, Hybrid Transliteration Architecture consists of five stages. These are input, Preprocessing, transliteration, post processing and the final stage gives the output of the system. From the above mentioned stages only three are the main stages which are explained further in 3.1,3.2 and 3.3 in detail.

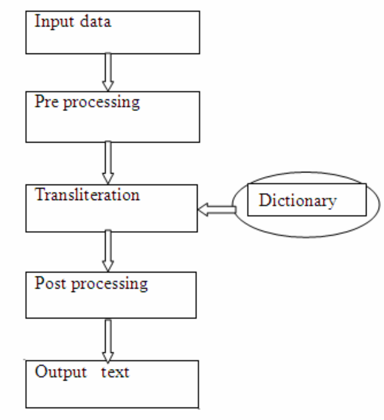

Fig.1 Architecture for transliteration

3.1 **Preprocessing:**

There are following steps to be followed under this :

Step 1: The English word goes through Schwa Deletion Algorithm [5]

The problem in many of the languages is mainly due to the existence of schwa vowel that is sometimes considered and sometimes not. In order to determine the proper transliteration of words, it is necessary to identify which schwas are to be deleted and which are to be retained as shown in Figure 2. The process of schwa deletion is one of the complex and important issue.

For instance,

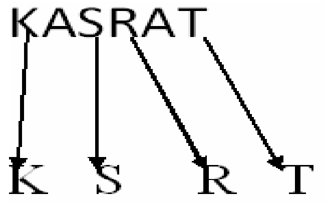

Fig.2 After Appling Schwa Deletion Algorithm

In the above example, a is to be deleted as per Schwa Deletion Algorithm. It is applied for appropriate results.

Step 2: Clustering: The output generated from schwa deletion algorithm is passed into clustering phase. Clustering means to form groups in source language words on the basis of certain information in target language. For clustering, input string is divided into characters. Individual characters or tokens are extracted from both English and Punjabi strings. Tokens are separated by space character. In Text tokenization unit, string is divided into characters. Individual characters or tokens are extracted from both English and Punjabi strings. Tokens are separated by space character. Tokens generated from text tokenization unit are passed to Transliteration unit. Clustering is shown in Figure.3.

Fig.3 Clustering

3.2 **Transliteration module:** The transliteration module either in Statistical Approach uses the dictionary that contains the spellings of names that are commonly used in real life. The output from previous step is compared with this dictionary and system will show only that spelling of those names that are given in the dictionary. If the no match found in dictionary then system will follow machine transliteration, which follow rule based approach which maps the characters as shown in Figure 4 with the corresponding character. For instance,

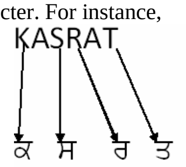

Fig.4 Mapping of character.

3.3.**Post Processing:**

This takes the input from its previous module, then apply the appropriate rules to provide the output as desired by the user.

4. FLOWCHART OF THE TRANSLITERATION SYSTEM USING HYBRID APPROACH:

Hybrid Approach uses both statistical and rule based approach shown in Figure 5.

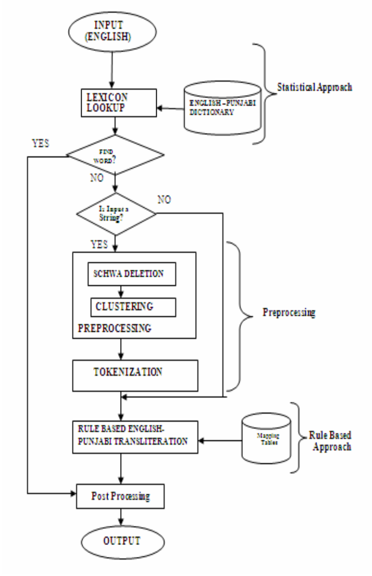

Fig.5 Working Flow Diagram of the Proposed System

Firstly, it follows the statistical approach which uses the bilingual dictionary. If it does not find the word in the dictionary then it uses the Rule based approach, which further undergoes many phases of the transliteration system. It is shown in Fig.5.

Based on the above process the expected error rate can be reduced to a great extent.

Conclusion:

With the advancement of technology and the ocean of information on web, it has become very common for one to adopt any foreign word into their own language. Transliteration is very much useful for those who are lame about the script of the language, although knows to speak and understand it. In this paper, we have addressed a hybrid approach for transliteration System. This reduces the rate of error in transliteration system to a great extent.

References:

1. Development of a Punjabi to English transliteration system, Kamal Deep and Vishal Goyal, International Journal of Computer Science and Communication Vol. 2, No. 2, July-December 2011, pp. 521-526
2. Statistical Approach to Transliteration from English to Punjabi, Jasleen kaur Gurpreet Singh josan, International Journal on Computer Science and Engineering (IJCSE), ISSN : 0975-3397 Vol. 3 No. 4 Apr 2011.
3. Hybrid Approach for Punjabi to English Transliteration System, Dr.Vishal Goyal, Kamal Deep, International Journal of Computer Applications (0975 – 8887) Volume 28– No.1, August 2011.
4. Rule Based Machine Translation of Noun Phrases from Punjabi to English,Kamaljeet Kaur Batra and G S Lehal,, IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 5, September 2010,
5. A Rule Based Schwa Deletion Algorithm for Punjabi TTS System, Parminder Singh, Gurpreet Singh Lehal,, Communications in Computer and Information Science, Volume 139, Part 1, 98-103, DOI: 10.1007/978-3-642-19403-0_16, 2011.
6. Machine transliteration, Knight, Kevin and Graehl, Jonathan. 1997.. In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, pp. 128-135.
7. Statistical transliteration for English-Arabic Cross LanguageInformation Retrieval, AbdulJaleel, Nasreen and Larkey, Leah S..CIKM 2003: Proceedings of the twelfth international conference on information and knowledge management, New Orleans, LA, 139-146.
8. Automatic transliteration for Japanese-to-English text retrieval. Yan Qu, Gregory Grefenstette and David A. Evans, SIGIR 2003: 353-360.
9. Vijaya, V.P., Shivapratap and K.P. CEN(2009), “English to Tamil Transliteration using WEKA system”, International Journal of Recent Trends in Engineering, May 2009, **1**, No. 1, pages: 498-500..Manoj Kumar Chinnakotla and Om P. Damani.

Sukhnandan Kaur is currently pursuing her M.Tech in Computer science and Engineering .from Swami Vivekanand institute of Engineering and technology, Banur. She holds the degree of B.Tech in Computer science and Engg. from Baba Banda Singh Bahadur Engineering college, Fatehgarh Sahib. She holds the OCP(Oracle Certified Professional) certification . She was the member of British Council of India.

Rupinderdeep Kaur is Lecturer in the department of Computer Science and Engineering at Thapar University, Patiala. She has almost two years of academic experience. She has received her B.Tech in Computer Sciences from Chandigarh Engineering College, Mohali and M.E. in Software Engg. Form Thapar University, Patiala. Her area of interest include Natural Language Processing and Database Management System.

Nidhi Bhalla is Lecturer in the department of Computer Science and Engineering at swami Vivekanand engineering college. She has almost six years of academic experience. She has done her B.Tech in information technology from Punjab technical university and M.Tech. From lovely professional university.