International Journal of scientific research and management (IJSRM)
\||Volume||3||Issue||3||Pages|| 2415-2420||2015||
Website: www.ijsrm.in ISSN (e): 2321-3418

English To Punjabi Script Converter System For Proper Nouns Using Hybrid Approach

Devinder Kaur[1], Er.Rishamjot Kaur[2]

[1]Department of Computer Science & Engineering,
Baba Farid College of Engineering and Techonolgy, Bathinda, India
devinderbrar90@gmail.com

[2]Department of Information Technology,
Baba Farid College of Engineering and Techonolgy, Bathinda, India
jotrishamsran@gmail.com

Abstract: *The most commonly faced problem with translators is to translating proper nouns like names and technical terms. In this research work we have proposed a English to Punjabi transliteration system for proper nouns that involves person names including male and female from Indian and foreign origin, continent names, city names, country names, state names etc. A hybrid approach is used to transliterate the proper nouns written in English text into its equivalent Punjabi text. Hybrid approach used in the proposed system is a combination of three approaches which are direct approach which is also known as dictionary lookup approach, rule based approach and statistical machine translation approach. The accuracy of the overall system highly depends upon the data provided for the training phase and data available in the parallel corpus. Various test cases have been performed on the proposed system and accuracy of proposed system is 90.4%. In future, this accuracy can be further improved by increasing the corpus size and using some more rules.*

Keywords: *NLP, Transliteration, SMT, Direct mapping, RBA***.**

1. Introduction

1.1 NLP (Natural Language Processing)

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 semantics

1.2 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.

1.3. 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 phase 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"

Type of transliteration:

- **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
- **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"

2. Problem Definition

Term transliteration means to produce the results from source noun into target noun keeping its pronunciation same. Existing system was generated using linux environment "moses toolkit and GIZA++ " which is difficult to operate and needs to be shifted to windows. Maximum accuracy of existing transliteration system is 63% which needs further improvement. A web based system is required to transliterate proper nouns so that it can be used anywhere in the world.

3. Proposed Work

We design a Graphical User Interface, which accepts input as a English text and will gives output as Punjabi text. The motive is to develop system that can be beneficial for the society and is useful in one's day-to-day life. In all government offices, records are written in English and if we are able to develop such system then it will be easy to convert them into Punjabi. This system will be used by most of government offices for converting their English text into Punjabi text. If this will solve their purpose, it will be implemented in government offices. Our proposed system transliterates most of the know nouns from English to Punjabi.

3.1 Design and Implementation of proposed system

Proposed system transliterates the proper nouns written in English into Punjabi using hybrid approach which is a combination of three approaches direct approach, statistical machine approach, rule based approach. To develop the system various technologies are being used:

- MS ACCESS (To store the training corpus)
- ASP.Net (To create the web Graphical User Interface)
- C#.Net (To be used as programming language)

3.1.1 Direct Approach (Dictionary Lookup)

Using direct approach system try to generate the result with the help of parallel corpus provided for training. It generates only those results which are in the parallel corpus Direct Approach (Dictionary Lookup). We have more than 15000 English Punjabi corpuses from where directly nouns are matched and output is displayed. Leaner search is applied for searching the names from the database. We have single database for all type of nouns.

3.1.2 Statistical Machine Translation

It uses N-Gram approach to train the system and to generate the results. This approach consists of two phases that is training phase and transliteration phase. This approach requires a parallel corpus to train the system. Accuracy of SMT depends on the corpus of the system. For transliteration bi-gram to six - gram are used. It provides better results than that of Rule Based System. Here we have flow chart of both the phases

- **Training phase**

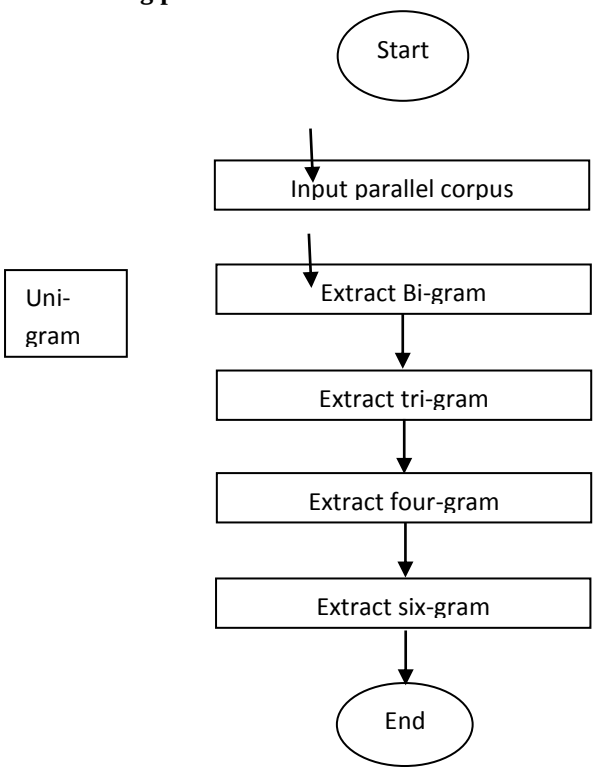

Fig 3.1 flow chart of training phase

- **Transliteration phase**

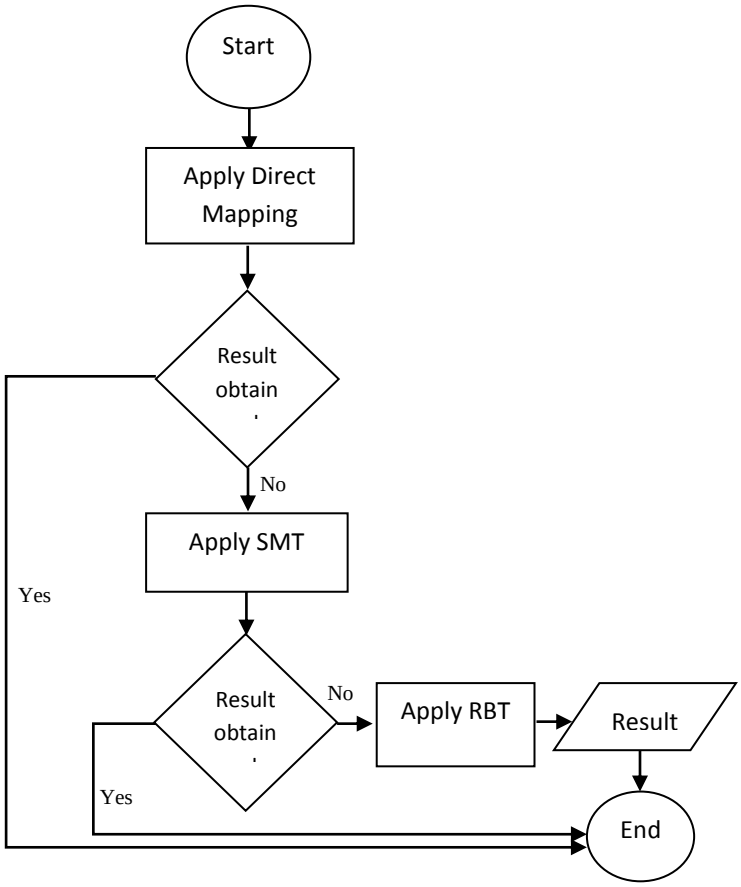

Fig 3.2 flow chart of transliteration phase

3.1.3 Rule Based

In this approach handcrafted rules are created to perform the task of transliteration. Rules are created by considering the properties of the source and target language. Rules-based approaches take time, money and trained personnel to make and test the rules. The rule base is a database consisting of the structural transformation rules, ambiguity rules, phrase rules etc. The knowledge base contains the rules for resolving the ambiguity of number of grammatical categories of words on the basis of type of surrounding words. Rules, not only check the grammatical category, but also number, gender or person in some cases. Rule base also contains the information about its synthesis, that while it is of same order or different. All the rules in the database are arranged according to priority. Phrase Rules are represented as context free grammar. Since these are recursive in nature, the number of rules is not very large, but in some cases, priorities are set depending upon the type of phrases for which the system is being made. [20]

4. Evaluation and results

We have more than 15000 entries in our database for English proper nouns. And we have tested our software on various English Proper nouns. 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, country names, continents names. Accuracy of all the test cases calculated using three parameters precision(P),recall(R) and f-measure(F).

$$P=\frac{no.of\ outputes\ generated\ by\ the\ proposed\ system}{no.of\ inputes\ given\ to\ the\ system}*100$$

$$R=\frac{no.of\ correct\ outputs}{thotal\ no.of\ inputes}*100$$

$$F=\frac{2*RP}{P+R}$$

4.1 Test case 1

Testing of proposed system for names of males

Jaspal singh	ਜਸਪਾਲ ਸਿੰਘ
Amandeep grewal	ਅਮਨਦੀਪ ਗਰੇਵਾਲ
Raajveer sandhu	ਰਾਜਵੀਰ ਸੰਧੂ
Sahil Sharma	ਸਾਹਿਲ ਸ਼ਰਮਾ
Rakesh	ਰਾਕੇਸ਼
Gurpal brar	ਗੁਰਪਾਲ ਬਰਾੜ
Sunil nagpal	ਸੁਨੀਲ ਨਾਗਪਾਲ
Mukesh bansal	ਮੁਕੇਸ਼ ਬਾਂਸਲ

Table 4.1 tabular representation for names of males

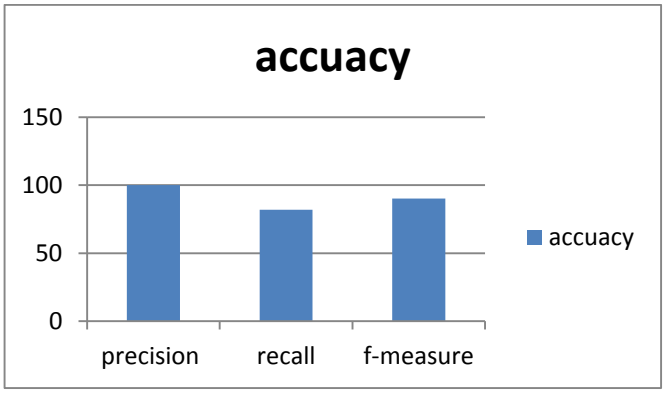

Fig 4.1 graphical representation for names of males

4.2 Test case 2

Testing of proposed system for names of females

Devinder kaur	ਦੇਵਿੰਦਰ ਕੌਰ
Sakshi goyal	ਸਾਕਸ਼ੀ ਗੋਇਲ
Kulveer kaur dhaliwal	ਕੁਲਵੀਰ ਕੌਰ ਧਾਲੀਵਾਲ
Satveer sidhu	ਸਤਵੀਰ ਸਿੱਧੂ
Kiran	ਕਿਰਨ
Jyoti bansal	ਜਯੋਤੀ ਬਾਂਸਲ
Sania	ਸਾਨੀਆ

Neelam	ਨੀਲਮ

Table 4.2 tabular representation for names of females

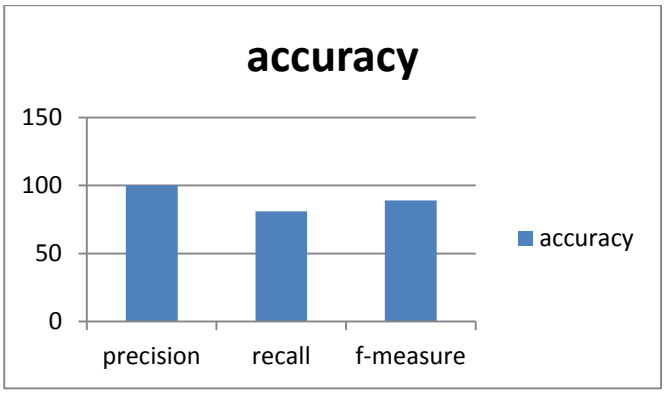

Fig 4.2 graphical representation for names of females

4.3 Test case 3

Testing of proposed system for names of Cities of India

Agra	ਆਗਰਾ
Gujarat	ਗੁਜਰਾਤ
Moga	ਮੋਗਾ
Bathinda	ਬਠਿੰਡਾ
Bikaner	ਬੀਕਾਨੇਰ
Batala	ਬਟਾਲਾ
Delhi	ਦਿੱਲੀ
Amritsar	ਅਮ੍ਰਿਤਸਰ

Table 4.3 tabular representation for names of Cities of India

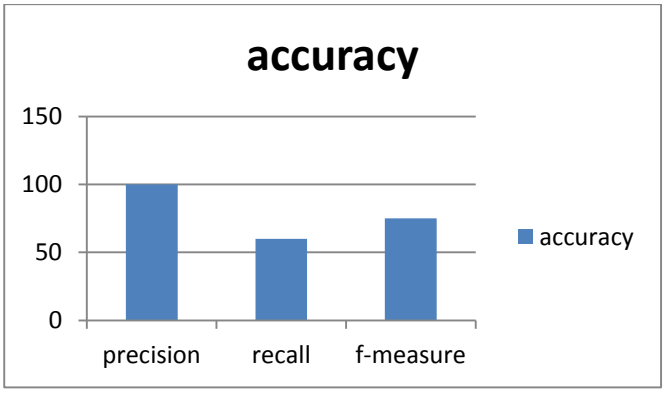

Fig 4.3 graphical representation for names of Cities of India

4.4 Test case 4

Testing of proposed system for name of States of India

Punjab	ਪੰਜਾਬ
Uttarakhand	ਉਤਰਾਖੰਡ
Tripura	ਤ੍ਰਿਪੁਰਾ
Sikkim	ਸਿੱਕਮ
Rajasthan	ਰਾਜਸਥਾਨ
Haryana	ਹਰਿਆਣਾ
Telangana	ਤੇਲੰਗਾਨਾ
Manipur	ਮਨੀਪੁਰ

Table 4.4 tabular representation for name of States of India

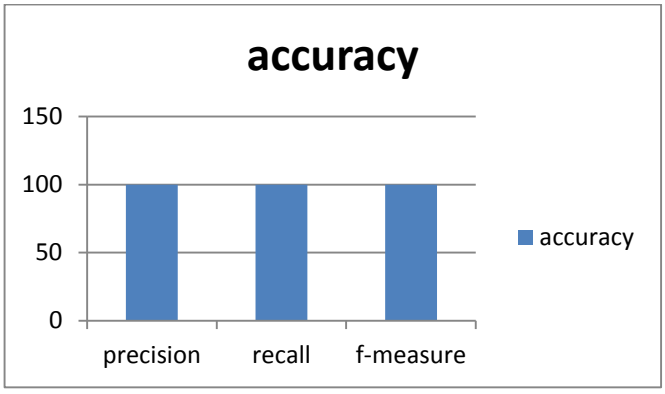

Fig 4.4 graphical representation for name of States of India

4.5 Test case 5

Testing of proposed system for name of continents of world

Asia	ਏਸ਼ੀਆ
Africa	ਅਫ਼ਰੀਕਾ
north America	ਉੱਤਰੀ ਅਮਰੀਕਾ
Europe	ਯੂਰੋਪ
south America	ਦੱਖਣੀ ਅਮਰੀਕਾ
Antarctica	ਐਂਟਾਰਕਟਿਕਾ
Australia	ਆਸਟ੍ਰੇਲੀਆ

Table 4.5 tabular representation for name of continents of world

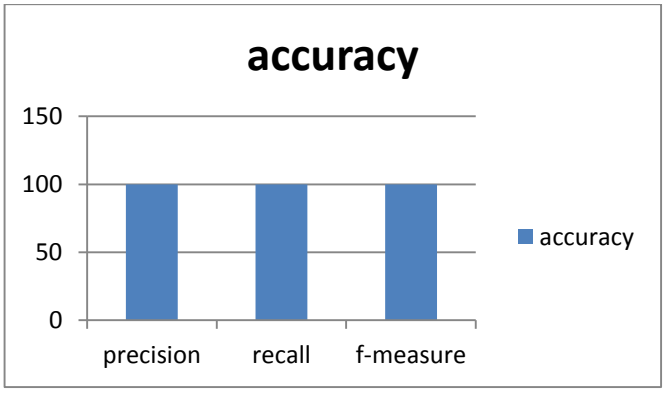

Fig 4.5 graphical representation for name of continents of world

4.6 Overall accuracy

Data set	Accuracy
Test case 1- for names of males	90.1%
Test case 2- for names of females	89%
Test case 3- for names of Cities of India	75%
Test case 4- for names of States of India	100%
Test case 5- for names of continents of world	100%
Overall %	90.82

Table 3.5 tabular representation for overall accuracy

5. Conclusion and Future Scope

Transliteration is the task for automatically converting words in one language into phonetically equivalent ones in another language. There is various machine transliteration models used for Transliteration. Each model has its own requirements for implementation. After studying number of works done by various researches in the area, we have developed new algorithm based on statistical machine translation for transliteration from English to Punjabi and the accuracy comes out to be approx. 90.82%.

This system is giving promising results and this can be further used by the researchers working on English and Punjabi Natural Language Processing tasks. Proper nouns from the State govt. English Documents, English Literature and other documents in English of one's interest can be transliterated into Punjabi for use on the click on a button. Now, as future work, database can be improved by including more names to improve the accuracy.

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 (IJCSE)

[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.

[20] Kamaljeet kaur batra and GS Lehal (2010), "Rule Based Machine Translation of Noun Phrases from Punjabi to English", IJCSI International Journal of Computer Science Issues,Vol.7,Issue5.

Er.Rishamjot kaur , I have done my B-Tech degree in Information technology from Baba Banda Singh Bahadur Engineering College, Fatehgarh Sahib. M-Tech in computer science & Engineering from University College of Engineering, Punjabi university Patiala.

Er.Devinder kaur, I have received my B-Tech degree in Information Technology from Baba Farid College of Engineering and Tecnology, Deon (Bathinda) in 2012 and pursuing M-Tech in Computer Science & Engineering from Baba Farid College of Engineering and Technology, Deon (Bathinda) My Research area is Natural Language Processing.