Volume 6, Issue 6, June 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Transliteration of Name Entities Using Rule Based Approach

Tejpaul S. Sasan, Dr. Shubhnandan S. Jamwal
Department of Computer Science & IT, University Of Jammu,
Jammu and Kashmir, India

Abstract— Transliteration is the conversion of a text from one script to another. It is basically the conversion of one writing system to another. The characters from a source language are mapped to the characters of the target language. During the process of translation from one language to another, the most significant problem is to translate the proper names and technical terms. In this paper a forward rule based transliteration system is developed for detection of proper nouns. The language pair used in the research is Dogri to English and we have achieved an accuracy of 90% for the said language pair.

Keywords— Transliteration, Name entity recognition, Rule based Approach.

I. INTRODUCTION

Name Entity Recognition is the process to detect named entities in a sentence like name of the city, state, country ,person, location , sport, river. NER is the process of identifying the entities (proper noun) in the running text and assigning them predefined categories. NER has following types of categories person, name, location, time, date, months, country, river, sport. Every machine translation system deal with out-of-vocabulary words, like technical terms and proper names of person, places, objects etc. Machine transliteration is an obvious choice for such words. During the process of translation from one language to another, the most significant problem is to translate the proper names and technical terms. The major challenge is the transliteration of out of vocabulary words that are appearing test. The problem is more complicated for language pairs with different scripts like Dogri/English due to the vast different in their character set (i.e. in Dogri total characters are 44 out of which 10 are vowels and 34 are consonants in which 6 are used for sound but in English language the total characters are 26 out of which 5 are vowels and 21 are consonants but it is not as serious in case of language pairs with similar scripts for example in French/English. The representation of characters of a given script into another script in such a way that the information of original script must be same is known as Transliteration.

II. LITERATURE REVIEW

Transliteration is not a new idea. This has been developed since 1885. Lots of work had been done on transliteration in English Language, but not much has been reported for Dogri Language. For Indian Languages,

- **Gurpreet Singh Josan, and Gurpreet Singh Lehal[1]** presented a novel approach to improve Punjabi to Hindi transliteration by combining a basic character to character mapping approach with rule based and Soundex based enhancements. The results show that following the rule based followed by Soundex approach produces the best results and the words can be transliterated with considerable accuracy. A fully accurate transliteration system is not possible due to the inherent problems, missing corresponding letters in two scripts. Although it is possible to transliterate across the scripts preserving the basic sounds of the source language, there will be some variations because the word in the source script may be pronounced somewhat differently in the target script, as per local conventions in each region.
- **Veerpal Kaur, Amandeep kaur Sarao and Jagtar Singh[2]** published a comprehensive survey and proposed a system for transliteration purpose using Statistical machine translation approach. The proposed system works in two phases. First phase comprises of System Training Phase and second is System Transliteration Phase.
- **Deepti Bhalla, Nisheeth Joshi and Iti Mathur[3]** developed transliteration system for English-Punjabi language pair using rule based approach. They have constructed some rules for syllabification. Syllabification is the process to extract or separate the syllable from the words. In this they have calculated the probabilities for name entities (Proper names and location). For those words which do not come under the category of name

entities, separate probabilities are being calculated by using relative frequency through a statistical machine translation Through this approach we attained accuracy of 88.19%.

- **Kamaljeet Kaur and Parminder Singh[4]** described the process of transliteration from English to Punjabi language using a rule based approach. Both source grapheme and phonetic information of words have been considered for rule formation to achieve high performance and more accurate result. Phonetic information proved vital for correct transliteration as well as for ambiguous words. The system is tested on news domain text of more than 10,000 words and achieved accuracy of 95%.
- **Vishal Goyal and Gurpreet Singh Lehal[5]** presented a hybrid translation approach for translating the text from Hindi to Punjabi. The proposed architecture has shown extremely good results and is found to be appropriate for MT systems between closely related language pairs.
- **Gurpreet Singh and Jagroop Kaur[6]** described the transliteration system build on statistical techniques. They worked on the combinations of approaches and the baseline model produces 73.13% accuracy rate and Statistical method shown 87.72% accuracy rate.
- **Navneet Garg, Vishal Goyal and Suman Preet[7]** described the part of speech tagger using rule based technique. Tokenized words are searched in the database and if not found then appropriate rules are applied. There is problem in handling the words that can act as both common noun and proper nouns.
- **Er.Sukhnandan kaur, Ms.Rupinder Kaur and Er.Nidhi Bhalla[8]** presented a system based on hybrid approach. They proposed the architecture of the hybrid system used for transliteration and this architecture reduces the rate of error in transliteration system to a great extent.
- **Kamaljeet Kaur Batra and G. S. Lehal[9]** developed a system to translate simple sentences in legal domain from Punjabi to English using rule based approach of translation. The steps involved are preprocessing, tagging, ambiguity resolution, phrase chunking, translation and synthesis of words in target language. The accuracy is calculated for different phases of the system and the overall accuracy of the system for a particular type of sentences is about 60%.

III. RESEARCH METHODLOGY

We have developed Dogri English transliteration system using rule based approach which helps in transliterating Dogri proper nouns into English proper nouns. For this transliteration system, we have developed a Graphical User Interface with which we can enter the input. Input can be directly entered through keyboard. In our system, we have used the dictionary of name entities. Source text has to be passed through various phases to get output string. Rules are created to detect the name entities

A. Preparation of database

In this system dictionary is created on the basis of different categories:

Name Dictionary: It consist of the names of different men or women with 584 proper nouns (person's name) from various sources.

Sport Dictionary: This contains name of the sports with 43 proper nouns (sport's name) from various sources.

Country Dictionary: Country dictionary contains the name of the Country/States/districts with 2133 proper nouns (Country/States/districts name) from various sources.

Surname dictionary: Surname dictionary contains the name of the surname with 855 proper nouns (surname name) from various sources.

Title Dictionary: Title dictionary contains the name of the titles which are given to the persons with 34 proper nouns (surname name) from various sources.

B. Architecture of System

The system takes Dogri text as input and gives English text as output. The algorithm is as follows:

Step1: Input the Dogri text in the system.

Step2: Preprocessing of Dogri text to be performed.

Step3: Division of Dogri text into Tokens on the basis of Spaces between the words.

Step4: Detection of name done using name dictionary and if word matches then go to step 9.

Step5: Detection of sports name is done using the sport dictionary and if word matches then go to step 9.

Step6: Detection of country, states, districts name are performed using dictionary and if word matches then go to step 9.

Step7: The token word is matched with the created rules of rivers, film, awards and if the any rules follow then transliterate the word and got to step 9.

Step8: The token word is matched with title and surname dictionary and if matches are found then names are transliterated.

Step9: End

The Architecture for transliteration is as follows:

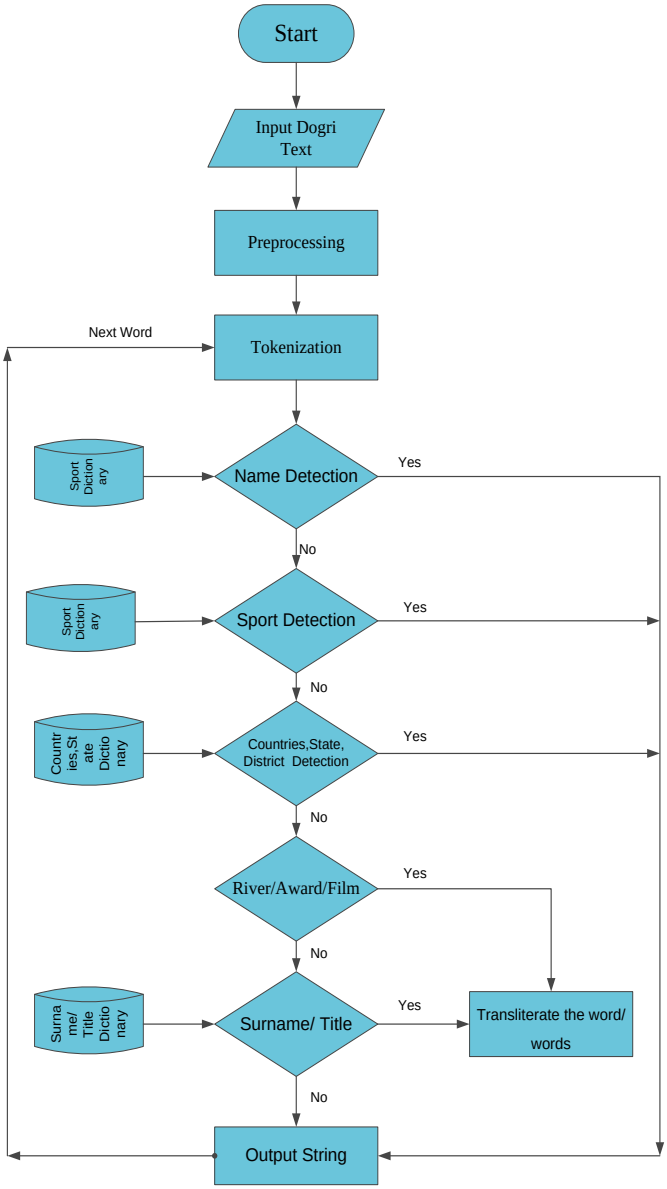

Figure 1: Architecture of Dogri English NER

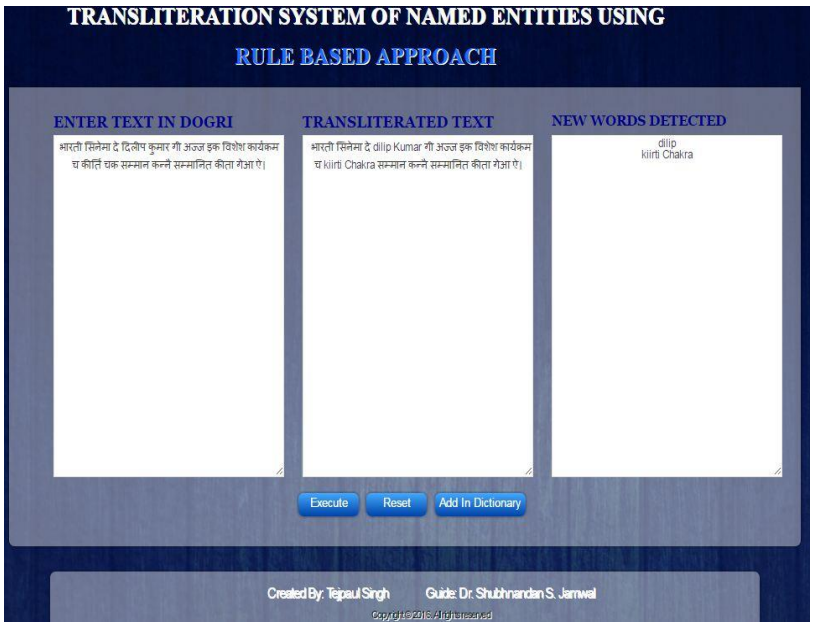

Figure 2: Interface of the NER

IV. RESULTS AND ANALYSIS

Test No.	Domain	No. of words
Test Case 1	Countries	2133
Test Case 2	Surnames & Names	1439
Test Case 3	Others	75
Test Case 4	Awards, Rivers, Films (Rules)	-

NER is the process of identifying the entities in the running text. In the system the four test cases are taken, in which 210(countries), 143(names/ surnames), 7(others), 11(Awards, Rivers, Films) name entities is given to the system of which 333 entities has been correctly identified and transliterated by the system. The system gives accuracy of 90%.

Results for Test Case 1:

Dogri Text	Our System
दिल्ली	DELHI
बांग्लादेश	BANGLADESH
हरियाणा	HARYANA
मेंढर	MENDHAR
कठुआ	KATHUA
आन्ध्र प्रदेश	ANDRA PRADESH
चीन	CHINA
नोएडा	NOIDA
केरल	KERALA
मुंबई	MUMBAI
पुणे	PUNE
महाराष्ट्र	MAHARASHTRA
अमेरिका	AMERICA
न्यूयॉर्क	NEW YORK
मेक्सिको	MEXICO
इटली	ITALY
इराक	IRAQ
नेपाल	NEPAL
स्विजरलैंड	SWITZERLAND
हिमाचल	HIMACHAL
छत्तीसगढ़	CHHATTISGARH
गोवा	GOA
तमिलनाडु	TAMILNADU
जिम्बाब्वे	ZIMBABWE
साउथ अफ्रीका	SOUTH AFRICA
वेस्ट इंडीज	WEST INDIES
कोलकाता	KOLKATA
बंगाल	BENGAL
अफ्रीका	AFRICA
श्रीलंका	SRI LANKA
कर्नाटक	KARNATAKA
उत्तर प्रदेश	UTTAR PRADESH
पठानकोट	PATHANKOT
सहारनपुर	SAHARANPUR
झारखंड	JHARKHAND
पालमपूर	PALAMPUR
पटना	PATNA
जोधपुर	JHODPUR
भारत	BHARAT
लुधियाना	LUDHIANA

Results for Test Case 2:

Dogri Text	Our System
कृष्ण	KRISHAN
गंगा	GANGA
नरेंद्र मोदी	NRENDR MODI
रितिक रोशन	RITIK ROSHAN
सुजैन खान	SUJAIN KHAN
अर्जुन कपूर	ARJUN KAPOOR
भरत	BARAT
बिपिन	BIPIN
बीरेंद्र	BIRENDRA

दानिश	DANISH
धर्मवीर	DHARAMVIR
अमित	AMIT
गगन	GAGAN
मनीष	MANISH
अरविंद केजरीवाल	ARVIND KEJRIWAL
अर्जुन कपूर	ARJUN KAPOOR
आयुष्मान	AYUSHMAN
रितेश	RITESH
सुब्रमण्यन स्वामी	SUBRMANYAN SWAMI
प्रकाश सिंह	PRAKASH SINGH
इमरान	IMRAN
सायना नेहवाल	SAYANA NEHWAL
यूसेन बोल्ट	YAOOSEN BOULT
रामदेव	RAMDEV
रतन टाटा	RATAN TATA
रविशंकर प्रसाद	RVISHNKR PRASAD
अरविंद	ARVIND
विजय माल्या	VIJAY MALLYA
रणवीर सिंह	RNVEER SINGH
करीना कपूर	KREENA KAPOOR
अनुपम खेर	ANUPAM KHER
अनुराग	ANURAG
प्रियंका चोपड़ा	PRIYANKA CHOPRA
नव्या नवेली नंदा	NAVYA NVELII NANDA
अक्षय कुमार	AKSHAY KUMAR
शाहिद कपूर	SHAHID KAPOOR
कोमल	KOMAL
वरुण धवन	VRUN DHAWAN
आलिया भट्ट	AALIYA BHATT
विद्या बालन	VIDYA BALAN
अरविंद देसाई	ARVIND DESAI
दीपिका पादुकोण	DEEPIKA PADUKONE
सलमान खान	SALMAN KHAN
संजय दत्त	SANJAY DUTT

त्रिशाला दत्त	TRISHALA DUTT
श्रद्धा कपूर	SHRDDHA KAPOOR
जरीन खान	JREEN KHAN
अनु मलिक	ANU MALIK
सेलिना जेटली	SELINA JAITLEY
सुरेश चटवाल	SURESH CHATWAL
करण सिंह	KARAN SINGH
राधिका	RADHIKA
करण जौहर	KARAN JOHAR
बराक ओबामा	BRAK OBAMA
राज ठाकरे	RAJ THAKREY
गोपाल राय	GOPAL ROY
अशोक	ASHOK
अनिल कुंबले	***C.*** ANIL KUMBLE
क्रिस गेल	***D.*** KRIS GAYLE
महेंद्र सिंह धोनी	MAHENDR SINGH DHONI
राहुल	RAHUL
लोकेश	LOKESH
राहुल	RAHUL
जीत रावल	JIT RAWAL
युवराज सिंह	YUVRAJ SINGH
सौरव गांगुली	SAURV GANGULY
दिलीप	DILIP
सचिन तेंदुलकर	SACHIN TENDULKAR
जुनैद खान	JUNAID KHAN
रवि शास्त्री	RAVI SHASTRI
लक्ष्मण	LAXMAN
ब्रायन लारा	BRAYAN LARA
सुनील गावस्कर	SUNIL GAVASKAR
सुनील	SUNIL
दिनेश चंदीमल	DINESH CHANDIMAL
विराट कोहली	VIRAT KOHLI
अभय शर्मा	ABHAY SHARMA
शाहरुख खान	SHAHRUKH KHAN
मोहम्मद हफ़ीज़	MOHAMMAD HAFEEZ
भूपिंदर सिंह	BHOOPINDR SINGH

For Test Case 3:

Dogri Text	Our System
कबाली फिल्म	KBALEE FILM
सुपरस्टार रजनीकांत	SUPERSTAR RJNEEKANT
डॉ सुभाष चंद्रा	DR. SUBHASHA CHANDRA
स्केट	SKATE

राष्ट्रमंडल खेल	COMMONWEALTH GAMES
बास्केटबाल	BASKETBALL
क्रिकेट	CRICKET

For Test Case 4:

Dogri Text	Our System
महावीर चक्र	*MAHAVIR CHAKRA*
शौर्य चक्र	*SHAURY CHAKRA*
पद्म विभूषण	*PDM VIBHOOSHAN*
कीर्ति चक्र	*KEERTI CHAKRA*
लोहित नदी	*LOHIT RIVER*
वायु सेना	*VAYU SENA*

शनिवार	*SATURDAY*
जनवरी	*JANUARY*
सितंबर	*SEPTEMBER*
सोमवार	*MONDAY*
शुक्रवार	*FRIDAY*

V. CONCLUSION & FUTURE WORK

Name Entity Recognition is the process to detect named entities in a sentence like name of the city, state, country, person, location, sport, river. In this paper an approach is formulated for named entity recognition that uses dictionaries and rules to detect named entity. A rule based approach is used to transliterate proper nouns of Dogri language into its English equivalent names. We developed this new algorithm for detection and transliteration from Dogri to English text. In the current research we have achieved an accuracy of 90%. In the Future we will add more rules/name entities in the dictionary and can improve our system by inheriting other techniques, so that better result can be achieved

REFERENCES

[1] Gurpreet Singh Josan, and Gurpreet Singh Lehal. "A Punjabi to Hindi Machine Transliteration System" The Association for Computational Linguistics and Chinese Language Processing Vol. 15, No. 2, June 2010, pp. 77-102.

[2] Veerpal Kaur, Amandeep kaur Sarao and Jagtar Singh. "A Review on Hindi to English Transliteration System for Proper Nouns Using Hybrid Approach" International Journal of Emerging Trends & Technology in Computer Science (IJETTCS) Volume 3, Issue 5,September-October 2014.

[3] 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.

[4] Kamaljeet Kaur and Parminder Singh. "English to Punjabi Transliteration using Orthographic and Phonetic Information".

[5] Vishal Goyal and Gurpreet Singh Lehal. "Hindi to Punjabi machine translation system".

[6] Gurpreet Singh Josan ,Jagroop Kaur."Punjabi to Hindi Statistical Machine Transliteration" International Journal of Information Technology and Knowledge Management, July-December 2011, Volume 4, No. 2, pp. 459-463.

[7] Navneet Garg,Vishal Goyal and Suman Preet. "Rule Based Hindi Part Of Speech Tagger" Proceedings of Coling 2012: Demonstration Papers,pages 163-174,Coling 2012 ,Mumbai ,December 2012.

[8] Er.Sukhnandan kaur, Ms.Rupinder Kaur and Er.Nidhi Bhalla. "English to Punjabi using Hybrid Approach" International Journal of Computer Science and Information Technology & Security, ISSN:2249-9555 Vol.2, No.2 April 2012 Page 482-485.

[9] Kamaljeet Kaur Batra and G. S. LEHAL. "Punjabi to English for Simple Sentences in Legal Domain" International Journal of Translation, Vol. 23, No. 1, Jan-Jun 2011 page 79-98.

[10] Er.Sahil Malhan, Er.Jasdeep Mann. "Punjabi to Hindi Transliteration System for Proper Nouns Using Hybrid Approach" International Journal of Engineering Research and Applications, ISSN: 2248-9622, Vol. 5, Issue 11, (Part - 4) November 2015, pp.06-10.

[11] Aditi Kalyani. "A Review of Machine Translation Systems in India and different Translation Evaluation Methodologies" International Journal of Computer Applications (0975 – 8887),Volume 121 – No.23, July 2015

[12] Artem Boldyrev. "Dictionary-Based Named Entity Recognition" Master's Thesis in Computer Science, University at des Saarlandes, December, 2013.

[13] Savita Singla, Prof. Seema Baghla. "Hybrid Approach for English to Punjabi Translation System for News Paper Headlines in a Specific Domain" International Journal of Engineering Research & Technology (IJERT), Vol. 2 Issue 11, November – 2013.

[14] B. Sasidhar, P. M. Yohan, Dr. A. Vinaya Babu3, Dr. A. Govardhan. "A Survey on Named Entity Recognition in Indian Languages with particular reference to Telugu" IJCSI International Journal of Computer Science Issues, Vol. 8, Issue2, March 2011.

[15] "Dogri Script." Internet: http://www.ciil-lisindia.net/Dogri/Dogri.html.

[16] Abdul Jaleel, N. and Larkey, L. "Statistical transliteration for English-Arabic cross language information retrieval", in proceedings of the 12th International Conference on Information and Knowledge Management, New York, USA, 2003, pp. 139-146.

[17] Deep, K. and Goyal, V. "Development of a Punjabi to English Transliteration System", International Journal of Computer Science and Communication, Vol. 2, No. 2, 2011, pp. 521-526.

[18] Dhore, M., Dixit, S. and Dhore, R. "Hindi and Marathi to English NE Transliteration Tool using Phonology and Stress Analysis", in proceedings of 24th International Conference on Computational Linguistic, Mumbai, India, 2012, pp. 111-118.

