



## English-Dogri Named Entity Recognition Using Statistical Machine Translation

Asmeet Kour, Dr. Shubhnandan S. Jamwal

Department of Computer Science & IT, Jammu University,  
Jammu & Kashmir, India

---

**Abstract**— *Natural Language Processing is a field of Computer Science and Artificial Intelligence. NLP enables a human to interact with computer in natural languages. Machine Transliteration is an important Natural Language Processing task. Transliteration is the conversion of word written in one language to other language without losing its phonological characteristics. Transliteration is used for the conversion of out-of-vocabulary words i.e. proper nouns and technical terms from one language to other language. In this paper we are using MOSES , a statistical machine translation tool for the transliteration of English-Dogri language pair .We have built parallel corpora of 85,965 entries. GIZA++ is used for word alignment over parallel corpora and KenLM is used to build statistical language model. The accuracy of our system for English to Dogri transliteration is 70 % and Dogri to English transliteration is 75%.*

**Keywords**— *Moses, GIZA++, corpus*

---

### I. INTRODUCTION

Natural language processing is a field of Artificial Intelligence that deals with the methods of communicating with computers in natural languages like English, Hindi, Dogri etc. NLP uses various computational and analysing processes which enables the computer to understand the language [1].

Dogri is an Indo-Aryan language. Dogri is mother tongue of 422 million people. It is the second prominent language of J&K State, presence of Dogri language can also be felt in northern Punjab, Himachal Pradesh and other places. There is not much work done in Dogri in the field of machine transliteration, so we are motivated towards this language.

Transliteration is defined as the generation of phonetic equivalents of OOV words in target language. OOV (Out Of Vocabulary) words are mainly named entities that include person, location, organization names etc. Named Entities transliteration is required in many applications, such as machine translation, corpus alignment, cross-language information retrieval and information extraction [2]. Transliteration and Translation are both different. Transliteration maps the letters of source script to letters in target script that have same pronunciation whereas translation is the generation of text in target language that has same meaning as that of source text. Translation is a process that communicates the same message in another language [3]. Transliteration is mainly used to translate proper nouns and technical terms. Transliteration is a two level decoding process; first segmentation of the source string into transliteration units; and then relating the source language transliteration units with units in the target language.

The remaining paper is organized as follows. In section II we describe related work in this field. Section III describes how MOSES is used in transliteration. In section IV we describe the methodology followed in this paper for transliteration between English- Dogri language pair. Section V describes results and evaluation and section VI describes conclusion of the paper.

#### Models of Transliteration:

- *Grapheme-based Transliteration Model:* It is direct orthographical mapping from source graphemes to target graphemes. This model is also referred to as direct/spelling method as it directly maps source language character to target language character.
- *Phoneme-based Transliteration Model:* In it, the transliteration key is pronunciation or the source phoneme rather than source grapheme. This model is basically source grapheme-to-source phoneme transformation and source phoneme-to-target grapheme transformation.
- *Hybrid-based Transliteration Model:* The Hybrid-based simply combines grapheme and phoneme through linear interpolation. It combines grapheme based transliteration probability and phoneme based transliteration probability through linear interpolation.
- *Correspondence-based Transliteration Model:* This model can combine any number of grapheme or phoneme based models but not both [4].

#### Approaches for Transliteration:

There are two approaches for machine transliteration:

- **Rule-Based Approach:** Rule Based Approach is the classical approach to transliteration. It uses rules manually written by linguists. There are several rule-based transliteration systems, [2] containing mainly lexicalized grammar, gazetteer lists, and list of trigger words. The rule based approaches require small amount of training data. The problem with rule based technique is its development is time consuming and changes are hard to accommodate and also these systems are not transferable to other domains.
- **Machine Learning Approach:** Machine learning approaches are most commonly used now a day. They produce output in very short time. These are of two types: supervised and unsupervised approach. In machine learning approach it is quite easy to accommodate changes and can be transferable to other domains.

## II. PREVIOUS WORK

Rakesh Chandra Balabantaray and Deepak Sahoo [2] developed Odia – English and Odia – Hindi machine transliteration system using statistical approach. They have used phrase –based SMT technique for transliteration of Odia-English and Odia- Hindi language pair. They have created two models for syllable based splits (Odia-English, Odia-Hindi) on 50,900 parallel entries and two models for character based splits (Odia-English, Odia-Hindi) on 1,10,000 parallel entries. The statistical approach that they have used is MOSES. To build statistical language model they used SRILM. To perform word alignments over parallel corpora they have used GIZA++. They have achieved an accuracy of 89% for Odia-English and 86% for Odia-Hindi on Syllable based split and 71% for Odia-English and 85% for Odia-Hindi on character based split.

Kamaldeep, Vishal Goyal [6] developed the system named “Punjabi to English Transliteration System” using a rule based approach and achieved accuracy of 93.23%. Transliteration scheme uses grapheme based method to model the transliteration problem. Their rule based approach involve a set of character mapping rules between the languages involved that are Punjabi and English. The system is evaluated for names from the different domains like Person names, City names, State names, River names, etc. The dataset is divided into two parts: one is training dataset other is testing dataset. Test Case 1 consists of person names and has accuracy of 95% whereas Test Case 2 consists of city names, state names, river names and have accuracy of 91.40%.

Lehal and Singh [7] developed Shahmukhi to Gurumukhi Transliteration System based on Corpus approach. This system has been virtually divided into two phases. The first phase performs pre-processing and rule-based transliteration tasks and the second phase performs the task of post-processing. The overall accuracy of system has been reported to be 91.37%.

Chinnakotla and Damani [8] developed Transliteration systems for English to Hindi, English to Tamil and English to Kannada. They have used a Phrase-Based Statistical Machine Translation approach to transliteration where the words are replaced by characters and sentences are replaced by words. They have used GIZA++ for word alignments and Moses for learning the phrase tables and decoding. The overall accuracy of English to Hindi is 49.0%, English to Tamil is 41.0% and English to Kannada is 36.0%.

Jasleen Kaur and Gurpreet Singh Josan [9] have used MOSES for transliteration from English to Punjabi using statistical approach. For transliteration they used three- tier architecture, which include pre-processing, transliteration unit and post processing. Training is done with the help of 3200 names in the both English and Punjabi in tokenized form. Average accuracy and Bleu score of this transliteration system without applying transliteration rules is 50.22% and 0.4123 respectively. After applying transliteration rules, average accuracy and Bleu score comes out to be 63.31% and 0.4502 respectively.

Pankaj Kumar and Er.Vinod Kumar [10] developed Statistical Machine Translation Based Punjabi to English Transliteration System for Proper Nouns. The transliteration system is divided into two parts – learning and transliteration. In learning phase training is given to the system on the basis of names stored into the database and generates tables, they store more than 15,000 unique names on which the system is trained. They test our system on more than 1000 names and the system has accuracy of 97%. System is also checked on those names which are not in the database of the system.

## III. TRANSLITERATING WITH MOSES

Moses (Koehn et al., 2007) [14] is an open source statistical machine translation engine that can be used to train statistical model of text translation from a source language to a target language. It requires parallel corpus for the languages for which you want to train it. Once you have the trained model, an efficient search algorithm quickly finds the translation having the highest probability among the exponential number of choices. Moses gives better principled method, both for learning useful phrases and combining them in the final process of transliteration [2].

Tools used by Moses are GIZA++ and KenLM

### A. Giza++

GIZA++ (Och and Ney, 2003) [12] is an extension of the program GIZA which we have used to perform word alignment over parallel corpora. The alignments are then used to learn the phrase transliteration probabilities which are estimated using the scoring function given best in (Koehn et al., 2007).

### B. KenLM

KenLM is a toolkit for building and applying statistical language model. We have used KenLM to build statistical language model. Language model is built with target language so as to ensure fluent output and in this 3-gram language model has been generated. [12]

#### IV. METHODOLOGY

The methodology of our system is:

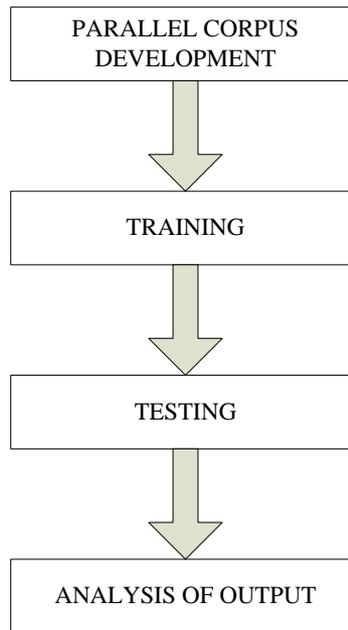


Fig1. Flow Chart of system

##### *Parallel Corpus Preparation*

We have created a parallel corpus of English –Dogri. The parallel corpus contains 85965 entries. We have collected data from newspaper and certain online sources. Our corpus preparation is in two steps:

- Firstly translation of Hindi sentence to English sentence and
- Then the same Hindi sentence is translated to Dogri sentence.

Then we make certain manual corrections to data so as to bring it in desired format like the files are required to be following UTF-8 encoding.

##### *Training*

Before preparing data for training following steps are performed:

- *Tokenization*: Tokenization is done so as to create spaces between the words and punctuation. Tokenization can be done on the basis of spaces, commas, punctuation marks etc. We have tokenized our data on the basis of spaces.
- *Truercasing*: In truecasing initial words in each sentence are converted to their most probable casing and this is done to reduce data sparsity.
- *Cleaning*: Long sentences, misaligned sentences and empty sentences are removed in cleaning so as to clear the data as these sentences can create problem in training pipeline. We have limited the length of sentences up to 80 words.

Then after this 3 –gram language model is generated. It is built with target language so as to ensure fluent output. Finally we train the model.

##### *Tuning*

Tuning refers to the process of finding the optimal weights for the linear model, where optimal weights are those which maximize translation performance on a small set of parallel sentences. [13] In it we have taken 1000 sentences other than those that are in parallel corpus.

##### *Testing*

Then the output is finally tested and the generated output is analyzed so as to check its correctness.

#### V. RESULTS

In this paper it has been shown that MOSES can be used to perform transliteration on any pair of languages. MOSES require parallel corpus. In this study a parallel corpus of 85,965 entries has been generated and GIZA++ has been used to perform word alignment over parallel corpus along with KenLM to build language model. The MOSES tool has been trained and tested on certain test cases. This tool has been tested on different domains like person name, city name, and state name etc. using manual evaluation method. The accuracy of the result depends upon the size of corpus. The accuracy of our system for English to Dogri transliteration is evaluated to be 70% and for Dogri to English transliteration is 75%.

## VI. CONCLUSIONS

From this study it is concluded that the accuracy of the transliteration system can be increased by developing a big parallel corpora and also one can adopt hybrid approach i.e. the combination of rule based approach and statistical approach for getting more accurate results. More work is needed to be done in this domain to improve its accuracy. In future, we will work in this direction to enhance its performance and accuracy.

## REFERENCES

- [1] Deepti Chopra, Sudha Morwal, “*Named Entity Recognition in Punjabi using HMM*”, International Journal of Computer Science & Engineering Technology (IJCSSET).
- [2] Rakesh Chandra Balabantaray, Deepak Sahoo , “ *Odia Transliteration engine Using Moses*” , Business and Information Management (ICBIM), 2nd International Conference on 9-11 Jan. 2014 ,pages 27 – 29, 2014.
- [3] Er.Sahil Malhan , Er. Jasdeep Mann, “ *Punjabi to Hindi Transliteration System for Proper Nouns Using Hybrid Approach*”, Int. Journal of Engineering Research and Applications(IJERA), Vol. 5, Issue 11, (Part - 4), pp.06-10, November 2015.
- [4] M L Dhorel,R M Dhore, P H Rathod , “*Survey On Machine Transliteration And Machine Learning Models*”, International Journal on Natural Language Computing (IJNLC) Vol. 4, No.2, April 2015.
- [5] Verma, “*A Roman-Gurmukhi Transliteration system*”, proceeding of the Department of Computer Science, Punjabi University, Patiala, 2006.
- [6] Kamal Deep, Dr.Vishal Goyal, 2011, “*Development of a Punjabi to English Transliteration System*”, International Journal of Computer Science and Communication, Vol. 2, No. 2, pp. 521-526, July-December 2011.
- [7] Lehal and Singh, “*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.
- [8] Chinnakotla and Damani , “*English-Hindi, English-Tamil and English-Kannada Transliteration Tasks*”. Proceedings of the 2009 Named Entities Workshop, ACL-IJCNLP 2009, pp. 44–47, Suntec, Singapore, 7 August 2009.
- [9] Jasleen Kaur and Gurpreet Singh Josan, “*Statistical Approach to Transliteration from English to Punjabi*”. Published in International Journal on Computer Science and Engineering (IJCSSE). ISSN: 0975-3397 Vol. 3 No. 4, pp. 1518-1527, Apr 2011.
- [10] Pankaj Kumar and Er. Vinod Kumar, “*Statistical Machine Translation Based Punjabi to English Transliteration System for Proper Nouns*”. Published in International Journal of Application or Innovation in Engineering & Management (IJAEM), Volume 2, Issue 8, pp.318-321, August 2013.
- [11] Ali and Ijaz, “*English to Urdu Transliteration System*”, Proceedings of the Conference on Language & Technology, pp. 15-23, 2009.
- [12] Hoon Oh, and Key-Sun Choi “*An English- Korean transliteration model using pronunciation and contextual rules*”. In Proc. of the 19th International Conference on Computational Linguistics (COLING 2002), pp. 393–399,2002.
- [13] ‘Tuning’, [Online]:<http://www.statmt.org/ Moses /?n= Factored Training. Tuning> [24 June, 2016].
- [14] Koehn et al. Moses: Open Source Toolkit for Statistical Machine Translation. In ACL, volume 45, pp.2-10, 2007.