Machine Transliteration in Code-Mixed Indian Social Media Text

Hemanta Baruah
(186155001)
Ph.D Research Scholar

Under the Supervision of

Dr. Sanasam Ranbir Singh
Dr. Priyankoo Sarmah

Centre for Linguistic Science & Technology
Indian Institute of Technology, Guwahati

OUTLINE

I. What is Transliteration ?

II. Types of Transliteration.

III. Translation vs Transliteration.

IV. Challenges in Machine Transliteration.

V. Code-Mixing in Social Media.

VI. Challenges in Code-Mixed Social Media transliteration.

VII. Application areas of Transliteration.

VIII. Dataset collection.

IX. Future Plan of action .

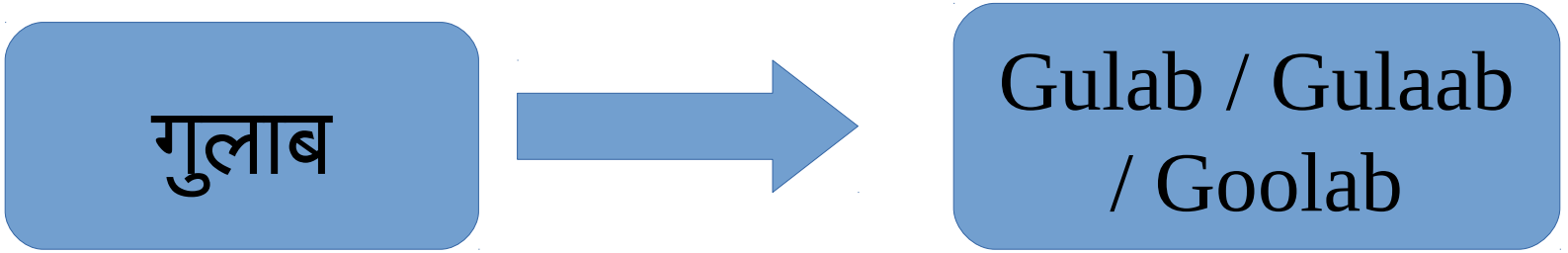
What is Transliteration

- Transliteration is the process of phonetic transformation of the script of a word from a source language to a target language, while preserving pronunciation.
- e.g :

- Transliteration helps people to pronounce words and names in foreign languages.
- In the process of transliteration, there is no loss of meaning or content.

Types Of Transliteration

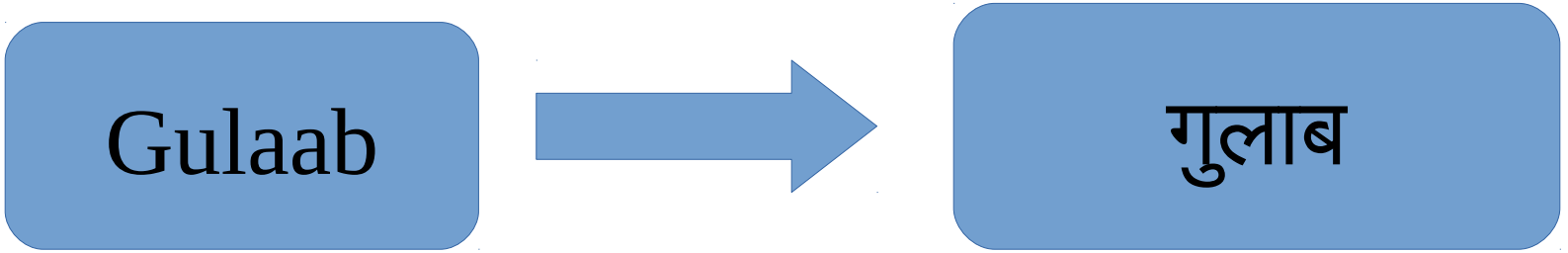
- **Forward Transliteration :**
- When one writes native terms using a non-native or foreign scipts.
- e.g :-



- Script Language → **Hindi** **English**
- Underlying Language → **Hindi** **Hindi**

Types Of Transliteration

- **Back Transliteration :**
- When one represents conversion of a term back to its native script, it is called back-transliteration .
- e.g :-



- Script Language → **English** **English**
- Underlying Language → **Hindi** **Hindi**

Forward Vs Back-Transliteration

- **Forward transliteration** allows for creativity of the transliterator.
- e.g :

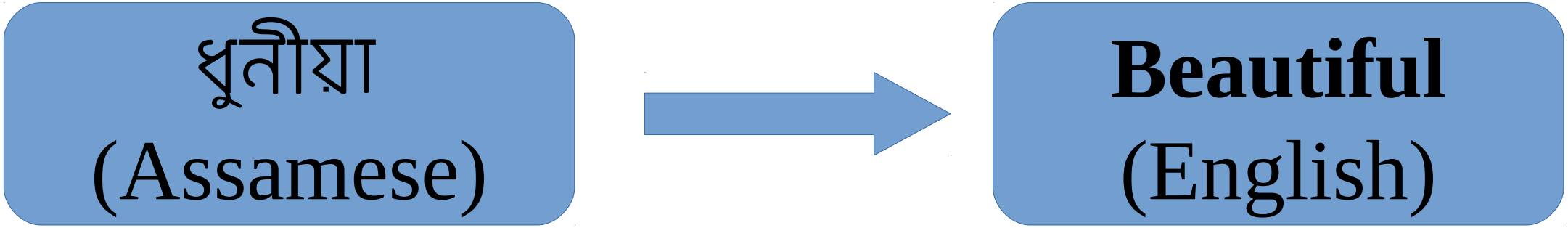
শোভাযাত্রা → Subhajatra / Hubhajatra / Shubhayatra / khubhajatra

- Whereas **Back-transliteration** is ideally strict and expects the **same initial word** to be generated.
- e.g :

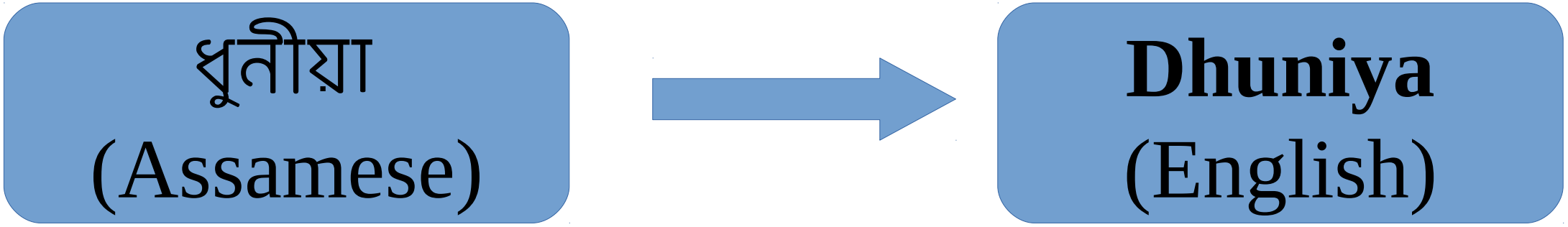
Xuoni / Khuwoni / Huwoni → শুরনি

Translation VS Transliteration

- **Translation :** Transfer of meaning takes place from one language (the source) to another language (the target).



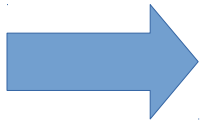
- **Transliteration :** Phonetically translating words from one source language to a target language alphabet.



Pure Vs Code-mixed Transliteration

- **Pure Transliteration :** Terms present in the sentences are from single language and written in non-native scripts.
- e.g

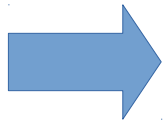
मुझे पहले से ही पता था
(Language - Hindi)



Mujhe pehle se hi pata tha
(Transliterated Hindi in English)

- **Code-mixed Transliteration :** Candidate terms are from different languages and might be in more than one languages.
- e.g

मे confident था
(Code-mixed Hindi with English)



Mei confident tha
(Transliterated Hindi +English)

Pure Vs Code-mixed Transliteration

- **Pure transliteration** follows some standard transliteration guidelines for the language under consideration. So the text are mostly formal.
- **Code-mixed transliteration** uses orthography of the scripts based on word pronunciation mixed with the terms of other language. Text are mostly informal in nature.
- Code-mixed transliterated text found abundantly in User generated text in social media.
- Language identification may not require in Pure transliteration.
- Language Identification is required before transliteration in code-mixed transliteration.

Challenges in Machine Transliteration

1) **Script specifications:** Knowledge of different character encoding, Direction of writing.
2) **Missing sounds**.
3) **Transliteration variants**.
4) **To decide whether to transliterate or not**: NEs are out-of-dictionary words where both translation and transliteration can be necessary.

e.g : **Congress** *Parliamentary Committee.*

Code Mixing in Social Media

1) **Code Mixing:** Embedding of linguistic units such as phrases, words and morphemes of one language into an utterance of another language.

2) No formally defined grammar for a code-mixed hybrid language.

3) A code-mixed sentence retains the underlying grammar and script of one of the languages it is comprised of.

e.g : grammer(**Assamese**) and script(**English**)

---> **Actually** moi aji **party** loi naahilu hoi **but** hi muk **forced** karile aahibo.

Eng-gloss : Actually I would not have come to the party today but he has forced me to come.

Different types of Code Mixing in Social Media

1. **Inter-Sentential :**
2. **e.g :** Fear cuts deeper than sword…… bukta fete jachche :(……
3. **Eng-Gloss :** Feaar cuts deeper than a sword…..it seems my heart will blow up….. :(
4. **Intra-Sentential :**
5. **e.g : Dakho sune 2mar kharap lagte pare but it is true that u are confused**
6. **Eng-Gloss:** You might feel bad hearing this but it is true that you are confused.

Different types of Code Mixing in Social Media Contd...

1.3. Tag :

2.e.g : Ami majhe majhe **fb** te on9 hole ei confession page tite aasi.

3.Eng-Gloss : While I get online on **facebook** I do visit this confession page very often.

4.4. Intra-word :

5.e.g : Tomar osonkkhho admirer **der** modhhe ami ekjon nogonno manush.

6.Eng-Gloss: Among your numerous admirer**s** I am the negligible one.

7. In this example the plural suffix of admirer (i.e. **admirers**) has been bengalified to der.

Challenges in Code-Mixed Social Media Transliteration

1) Very informal nature of code-mixed social media text.
2) Social media text suffers from several phenomena code-mixing, code-switching, lexical borrowings etc.
3) Other challenges like spelling errors, auto-correction, creative spellings (e.g: **gr8** for **great**), word play ("**goooooood**" for "**good**"), abbreviations ("**OMG**" for "**oh my GOD!**"), meta tags (**URL**s, **Hashtag**s) and so on.
4) Non-standard roman spelling variations for the words in a language in Social media.
5) In a code-mixed sentence, word-ordering is lost; and thus an important feature for sentence analysis is lost.

Application areas of Machine Transliteration

1) Machine Translation (**MT**).
2) Parts-Of-Speech (**POS**) tagging.
3) Mixed script information retrieval (**MSIR**).
4) Sentiment Analysis (**SMA**).
5) Language Identification
6) Code-mixed information retrieval (**CMIR**)

Machine Translation

1) Traditionally used in Machine Translation to translate Named Entities, NEs and Out Of Vocabulary, OOV words .

2) Building of different linguistic tools for low resource language to get the inside of the data .

Parts-Of-Speech (POS) tagging

1) POS tagger for any language is an important linguistic tool for performing any NLP task .

2) Researh on building POS tagger for code-mixed social media text.

3)

4) Language specific code-mixed roman transliteration should be done before subjecting it to **POS** tagging.

Mixed script information retrieval

1) Text document contains multiple scripts involving multiple languages
.

2) Each language may use its own native script within a single document.

3) Spelling variations can occur across queries and documents, even within a single document.

4) To resolve them it is necessary to bringing them to a common form

Sentiment Analysis

1) Multi-lingual users on Social Media usually generates code-mixed sentiment bearing transliterated text.

2)

3) No formally defined grammar for a code-mixed hybrid language in Social Media.

4) Traditional approaches to Sentiment Analysis(**SA**) does not work very well on code-mixed content.

Language Identification

1) For any multilingual NLP task, language identification is always the first step to start with.

2) Language identification for code-mixed Social Media content is a difficult task due to its inherent characteristics.

3) For the transliterated contents either we can do the transliteration first then identification or we can do the reverse.

Code-mixed information retrieval

1) Multi lingual users create multi lingual documents.

2) Code-mixed information retrieval faces multilingual issues and term mis-matching.

3) Combine effort of language identification, translation/transliteration helps to address the problem of code-mixed information retrieval, CMIR.

Why this problem is important

1) Rapid growth of multi-lingual users as well as user generated transliterated contents all over the internet.
2) These informal text contains a very good amount of useful information.
3) Before applying any NLP techniques, user generated noisy text requires some pre-processing. (translation or transliteration)
4) Transliterated search on web by multi-lingual users.
5) Very few existing research on low resource Indian languages in the field of code-mixed machine transliteration.

Dataset Preparation

1) Currently collecting English – Assamese transliterated data from YouTube video comments.

2) Have collected available Eng-Hindi code-mixed transliterated data from existing research work.

3) Data annotation is going on for existing transliterated Assamese, Bengali and Hindi text collected from Facebook.

Future Work Plan

Duration	Work Plan
Year wise collection of all previous research papers related to text transliteration and translation domain in general and code-mixed social media text in specific, collection of online available datasets, in-house collection of datasets.	Aug – Oct , 2019
Study and explore all state-of-the-art NLP techniques used in Machine Translation and Transliteration.	Aug – Oct , 2019

- **Thank You.**