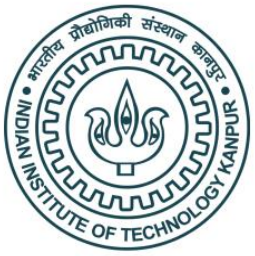
INDIAN INSTITUTE OF TECHNOLOGY, KANPUR

CS-689A Computational Linguistics for Indian Languages

Project Report



Title: Transliteration Correction for Indian Words

Supervised by:

Prof. Arnab Bhattacharya

Submitted by:

Darshan Jain (231110009)
Shaurya Agarwal (231110046)

Acknowledgement:

We would like to extend our sincerest gratitude to our Professor Arnab Bhattaacharya for providing the inspiration and opportunity for this project. Their insightful suggestion to correct Indian words in English text to their correct ISO 15919 transliteration led to a captivating exploration into the realm of linguistic and cultural accuracy. The chance to delve into the Mahabharata text within this context has been both enlightening and enriching. His guidance and support throughout this endeavor have been invaluable.

Contents of the Report

1. ABSTRACT

In today's globalized world, where diverse cultures intersect through various mediums, accurate representation and preservation of cultural heritage are paramount. One significant challenge in this endeavor is the faithful transliteration of indigenous words, especially from languages with non-Latin scripts, like Indian languages, into English.

This project aims to develop a robust system for correcting Indian words in English text to their correct ISO 15919 transliteration and correcting the Indian words to their correct ISO 15919 transliteration on the Mahabharata text. The Mahabharata, one of the two major Sanskrit epics of ancient India, contains a vast array of Indian names and terms often transliterated into English in various ways. However, inconsistencies and inaccuracies in transliteration can lead to confusion and misinterpretation of the text. By leveraging natural language processing techniques and linguistic resources, this project seeks to automatically identify Indian words in English text and transform them into their accurate ISO 15919 transliteration, thereby enhancing the readability and understanding of the Mahabharata for a wider audience.

2. INTRODUCTION

The transliteration of Indian words into English poses a significant challenge due to the diversity and complexity of Indian languages and scripts. Inaccurate transliterations can lead to misinterpretations, confusion, and loss of cultural nuances in texts containing Indian terms. This project aims to address this challenge by developing a system that automatically corrects Indian words in English text to their correct ISO 15919 transliterations.

ISO 15919 (Transliteration of Devanagari and related Indic scripts into Latin characters) is one of a series of international standards for romanization by the International Organization for Standardization. It was published in 2001 and uses diacritics to map the much larger set of consonants and vowels in Devanagari scripts to the Latin script. By adhering to this standard, the project seeks to improve the quality and reliability of transliterations in English texts containing Indian words.

The Mahabharata, an ancient Indian epic, is one of human history's most revered and voluminous literary works. Its profound narratives, rich cultural nuances, and timeless wisdom have transcended geographical boundaries, captivating the minds and hearts of people across the globe. However, as the epic traverses linguistic barriers, the faithful representation of its indigenous terminology in English translations poses a significant challenge.

This project endeavors to bridge the gap between the original Indian words and their English representations in the Mahabharata, aligning them with the ISO 15919 standard. By employing natural language processing techniques and linguistic analyses, we propose a method to automatically correct Indian words in English text to their accurate transliterations. Our approach aims to enhance the linguistic fidelity of Mahabharata translations and facilitate a deeper understanding of its cultural and linguistic heritage. The project has practical applications in fields such as education, research, and information retrieval. By providing a reliable tool for transliterating Indian words in English text, the system can facilitate language learning, enable linguistic analysis, and improve the accessibility of information for speakers of different languages.

Through this report, we present our methodology, experimental results, and insights gained from the correction of Indian words in the English text of the Mahabharata. We believe that our findings contribute to computational linguistics and hold broader implications for cross-cultural communication, textual accuracy, and the preservation of cultural heritage in an increasingly interconnected world.

3. PROBLEM STATEMENT

Given the inherent complexities of transliterating Indic scripts into the Latin alphabet, the lack of standardized practices in existing English translations of Indian words results in inconsistencies and inaccuracies. This inconsistency impedes the text's computational processing and undermines linguistic representations' fidelity. Therefore, there is a critical need for an automated system that can accurately correct Indian words in English text to their ISO 15919-compliant transliterations. This project seeks to develop such a system using natural language processing techniques, enhancing the accuracy of Indian word translations and facilitating computational analysis of its cultural and linguistic content.

BLOCK DIAGRAM OF APPROACH:

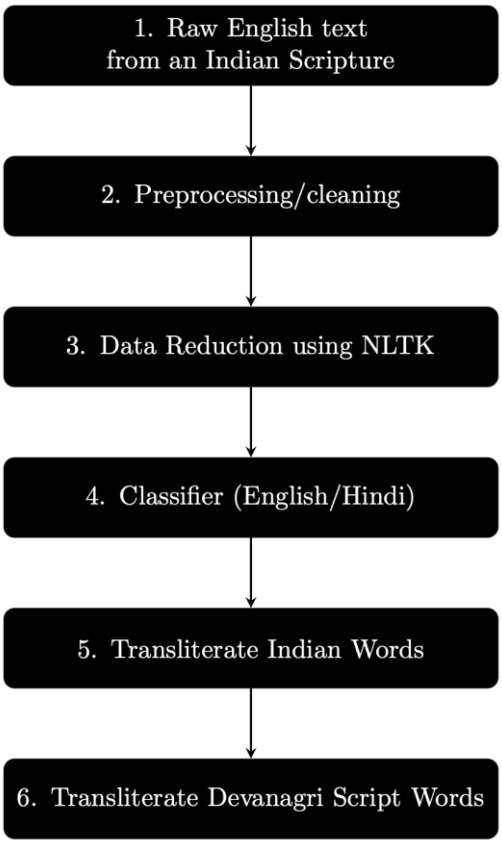


Fig: Work Flow

4. ISO - 15919 TRANSLITERATION RULES[5]

Devanāgarī	IAST	ISO 15919	Devanāgarī	IAST	ISO 15919			
अ	a	a	क	ka	ka	द	da	da
आ	ā	ā	ख	kha	kha	ध	dha	dha
इ	i	i	ग	ga	ga	न	na	na
ई	ī	ī	घ	gha	gha	प	pa	pa
उ	u	u	ङ	ṅa	ṅa	फ	pha	pha
ऊ	ū	ū	च	ca	ca	ब	ba	ba
ए	e	ē	छ	cha	cha	भ	bha	bha
ऐ	ai	ai	ज	ja	ja	म	ma	ma
ओ	o	ō	झ	jha	jha	य	ya	ya
औ	au	au	ञ	ña	ña	र	ra	ra
ऋ	ṛ	r̥	ट	ṭa	ṭa	ल	la	la
ॠ	ṝ	r̥̄	ठ	ṭha	ṭha	व	va	va
ऌ	ḷ	l̥	ड	ḍa	ḍa	श	śa	śa
ॡ	ḹ	l̥̄	ढ	ḍha	ḍha	ष	ṣa	ṣa
अं	ṃ	ṁ	ण	ṇa	ṇa	स	sa	sa
अः	ḥ	ḥ	त	ta	ta	ह	ha	ha
अँ	ã	m̐	थ	tha	tha	ळ	ḻa	l̲a

Devanāgarī	ISO 15919
क्ष	kṣa
त्र	tra
ज्ञ	jña
श्र	śra

Devanāgarī	ISO 15919
क़	qa
ख़	kha
ग़	ġa
ज़	za
फ़	fa
ड़	ṛa
ढ़	ṛha

5. METHODOLOGY

5.1. Dataset Collection

The dataset is obtained from the Mahabharata in English script . It contains several Indian names that are not in ISO 15919 format.

5.2. Data Cleaning

The dataset text is cleaned to remove any irrelevant information or noise that might interfere with the transliteration process.

It includes:

- Removing punctuation marks or any other special characters
- Removing numbers
- Next line character
- Extra white spaces

5.3. Word Classification

The words in the cleaned dataset are classified into two categories: English and Indian (Hindi). The NLTK English dictionary removes common English words from the dataset to achieve this. The remaining words are then subjected to classification using the Microsoft Language Identity Tool.

Even though the Microsoft Lid tool considers the context to detect the word as Hindi or not, for our project purpose, it is sufficient to classify the word without the need for context of the sentence because the words that need to be transliterated are spelled correctly and written correctly. Hence, we passed all the text into the classifier tool word by word. Words identified as Hindi are tagged as Indian words.

After passing the words to the classifier, we get the tags for each word, which are either HI or EN for Hindi and English, respectively. These words with Hindi tags are stored and sent to the next step.

We have taken Mahabharata Dataset from Prof. Arnab Bhattacharya

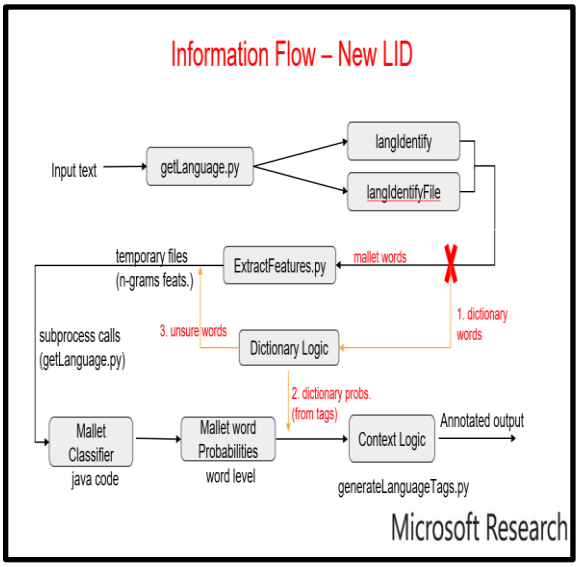


Fig: Microsoft LID Tool flow chart[1]

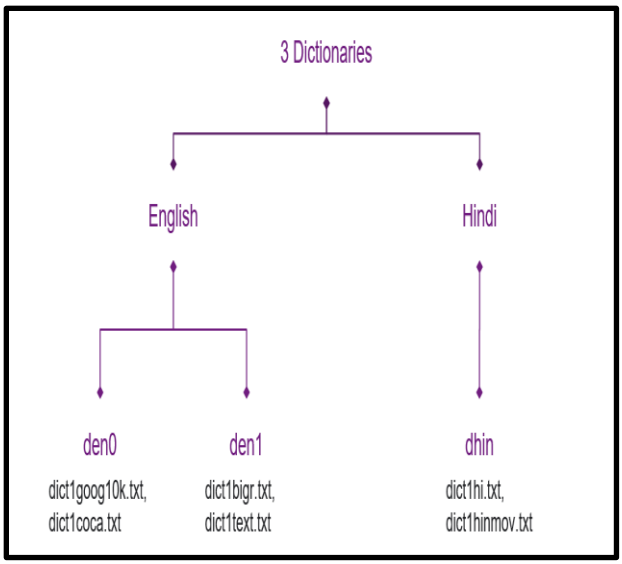


Fig: Dictionaries used by Lid tool[1]

5.4. Transliteration to Hindi

The Indian words identified in the previous step are transliterated from English to Devanagari script using the XlitEngine from ai4bharat.transliteration. This step ensures that the Indian words are represented accurately in the Hindi script. It considers top-k items that can be the correct transliteration of a given word in Devanagari. We see that on using the k value as 5, the topmost predicted value is closest to the actual Indian word, pronounced correctly. We took the word with the highest probability from the predicted words as our chosen word in the Devanagari script.

5.5. Transliteration Back to Latin ISO 15919

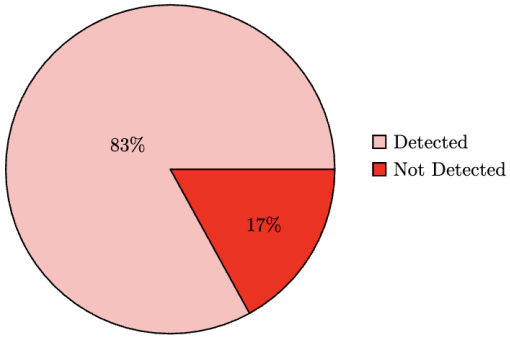
The transliterated Indian words in the Devanagari script are then transliterated back to the Latin script using the transliteration module from the Akshara Mukha library. This step ensures that the corrected Indian words are represented accurately in the Latin script, making them readable and usable within English text. It takes a word in Devanagari Script, breaks the word into characters, and then internally maps characters to ISO 15919, recombines the mapping and gives the corrected word as output.

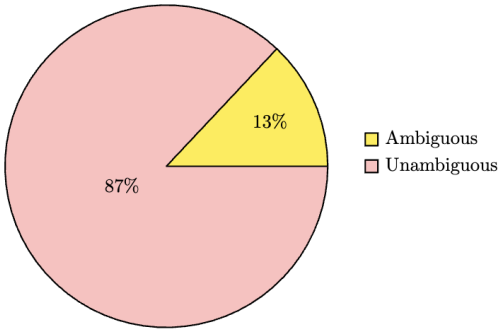
Finally, the complete English text is generated with a corrected ISO 15919 representation of Hindi words.

6. RESULTS

6.1. Classification:

Our Model was able to detect the language of 83% of the actual Hindi words correctly as Hindi. It was not able to classify 17% of the actual Hindi words as Hindi words. Actually, some of the Hindi words were detected as English by the Microsoft LID tool, possibly because those words are nowadays very common in English texts.

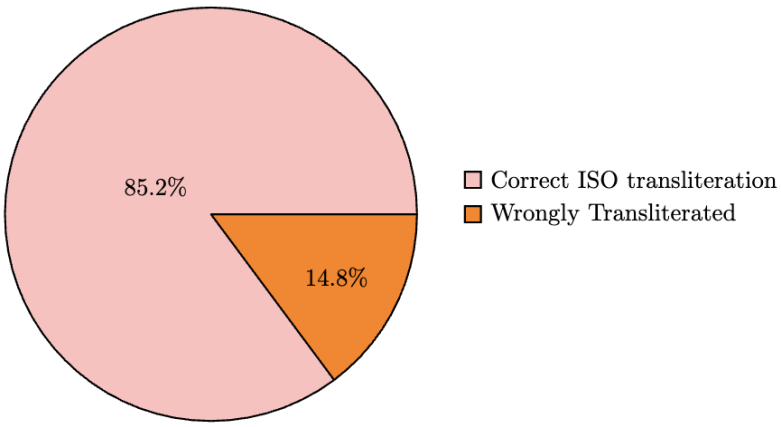




NOTE: 13% of the detected words were not actual Hindi words but some hindi words mixed with some english, they were some manipulation of Hindi words.
Example: Bhishmas, Kauravas, devas, vedas, etc. are considered as ambiguous words.
So, we will look at our score calculations on unambiguous words for now.

6.2. Correct transliteration of Unambiguous Hindi words to ISO -15919 format:

On unambiguous Hindi words detected by the model, 85.2% of the words were correctly transliterated to their correct ISO 15919 format.



7. LIMITATIONS

1: Some misclassification of words into English/Hindi was noticed. Basically, the words that were either too small like “Rik”, “son” or “so”. Classifiers misclassified them as English. Reason the words being written in English and Hindi both or unable to detect small words.
2: Mistransliteration of some Indian words into Hindi because of misspelling.
3: Currently, it works for correcting Hindi words only.
4: Manually corrected words to ISO 15919 might have some errors.
5: Our model is libraries and module dependent as of now, further work acn be done to make it more independent and robust.

8. FUTURE SCOPE

1: Better results may be obtained if a more powerful transliteration model is included.
2: Proposed model can be tested on other regional languages as part of future research.
3: Better results may be obtained if a more diverse and large corpora is included to train the transliteration model
4: Better results may be obtained if a more powerful classifier for language tags is used.
5. If the classifier takes the context of the sentences, it will give better results for language identification of words

9. CONCLUSION

Our approach shows promising results in enhancing cultural representation and linguistic accuracy in the translation of Indian epics like the Mahabharata. Future iterations can address limitations and further contribute to computational linguistics and cross-cultural communication.

10. REFERENCES

[1]. Microsoft LId Tool: https://github.com/microsoft/LID-tool

[2]. AksharMukha: https://pypi.org/project/aksharamukha/

[3]. [“ye word kis lang ka hai bhai?” Testing the Limits of Word level Language Identification](https://aclanthology.org/W14-5151) (Gella et al., ICON 2014)

[4]. S. Biradar, S. Saumya and A. Chauhan, "Hate or Non-hate: Translation based hate speech identification in Code-Mixed Hinglish data set," 2021 IEEE International Conference on Big Data (Big Data), Orlando, FL, USA, 2021, pp. 2470-2475, doi: 10.1109/BigData52589.2021.9671526. keywords: {Training;Social networking (online);Hate speech;Big Data;Writing;Linguistics;Feature extraction;hate speech;deep learning;mBERT;Trans-former}

[5]. Devanagari transliteration : https://en.wikipedia.org/wiki/Devanagari_transliteration