Word Level Language Identification for Romanized Code-Mixed Indian Languages

Devika Singh (210101036)
Dr. Sanasam Ranbir Singh

Department of Computer Science and Engineering
Indian Institute of Technology Guwahati

25 November 2024

Contents

Introduction

Language identification

- **To identify the language of a given input**
- **Different levels of granularity:**
  - Document Level
  - Sentence Level
  - Word Level
  - Character Level
- We tackle at word level, each word in a sentence is tagged with its respective language

Transliteration

Definition: Convert words to another script based on phonetic sounds, representing original language sounds in a different alphabet.

CODE-MIXING – ALSO CALLED *Code-switching*

- **Definition**: Blending of linguistic elements from multiple languages within a single utterance, sentence, or conversation.
- **Types of Code-Mixing:**
  – Single-Script Code-Mixing
  – Mixed-Script or Intra-sentential Code-mixing
  – Inter-sentential Code-Mixing
- Multilingual communities on social media generate code-mixed data.

Main kal school nahi jaungi because I am feeling sick.

Translation: "I will not go to school tomorrow because I am feeling sick."

FIGURE: Single-script code-mixing in Romanized Hindi and English languages.

Related Works

Cite Key	Work
(1)	Bilingual Hindi-English Twitter analysis
(2)	Code-mixing NLP challenges in multilingual settings
(3)	Twitter code-switching via word-level detection
(4)	Language ID in code-switched data task
(5)	Multilingual dataset web-crawl audit
(6)	Resources and models for Indian languages
(7)	Efficient text classification techniques
(8)	Defining Indian internet with languages
(9)	Multilingual token classification for Indian languages
(10)	Multilingual Indian language representations
(11)	Llama 3 model series

Table: Summary of Works in Language Identification

Related Works (Contd.)

fastText

FastText is a library developed by Facebook for efficient learning of word representations and text classification.

- **Word Vectors:** Generates vectors using subword info.
- **Model Training:** Fast neural network for large datasets.
- **Subword Information:** Uses character n-grams.
- **Language Support:** Supports multiple languages.

Related Works (Contd.)

Bhasha-Abhijnaanam

- **Dataset Overview:** Targets sentence-level language identification for 22 Indic languages; clean, labeled data in native and romanized scripts.
- **Linear fastText Model:** Efficient, split into native and romanized script versions; uses character n-grams to distinguish languages.
- **BERT and Hybrid Models:** BERT offers high accuracy; hybrid combines fastText speed and BERT accuracy for optimal performance.

CHALLENGES

- **Lack of Dataset:**
  – Gap in widespread use of code-mixed language vs. data availability; scarcity due to rare appearance in formal web texts.
  – Privacy restrictions on platforms hinder code-mixed dataset generation.
  – Synthetic generation of most code-mixed datasets limits capture of natural complexity of code-mixing.
  – Scarce pre-training data for low-resource languages in large multilingual corpora.
- **Multilinguality:**
  – Spelling variations from phonetic transliterations in code-mixing, internet slang, and non-standard spellings.
- **Borrowed words:**
  – Difficult to determine if words belong to the "borrowing" or "native" language.
- **Homographs:** Identical spellings across languages with different meanings in code-mixed text challenge word-level language identification, requiring context for accuracy.

Problem Statement and Proposed Approach

Problem Statement

- Creating a word-level, Romanized code-mixed language-labelled dataset for 22 Indian languages
- Develop a model that can identify the language of each word in the code-mixed sentences.

Proposed Approach

Dataset

- Dataset proposed with sentences in Romanized 22 Indian languages.
- Each word in the sentences is tagged with its language.
- Each record includes a label tag structured as a key-value pairs, mapping words to their respective languages.

Proposed Approach (Contd.)

Model

- We propose a word-level tagging system enhanced on out sentence-level language identification, utilizing a word-labeled dataset for training.
- Individual words are analyzed using a Unicode-Range script classifier to determine their scripts, focusing on Devanagari, Bengali, Arabic, and Latin.
- Words not matching these major scripts are tagged directly based on their script.
- For major script words, language identification employs fastText models trained on word-level data specific to each script.
- This method integrates script and language classification to accurately tag each word in code-mixed sentences.

Progress

Native Script Dataset

Features sentences in multiple Indic languages, each tagged with one of 31 distinct language and script identifiers for sentence-level language identification.

Romanized code-mixed dataset

- **Data Collection:** Gathered from social platforms like Twitter, Facebook, and YouTube using location-based filters.
- **Pre-processing Challenges:** Include only sentences in Latin script. Issues due to purely English sentences or code-mixed languages like Hinglish from Tamil speaking region.
- **Language Detection:** Uses Llama 3 to identify sentence languages; dataset pending human verification and future word-level tagging.
- **Data Collection Progress:** Currently gathered data for Hindi, Tamil, Telugu, Malayalam, and Kannada.

PROGRESS (CONTD.)

sentence	lang_tag
Kal meeting hogi.	hindi
nuv on screen medha em chestawo hit padali seenaa	kannada
ldf divides udf unites #udfwilltransformkerala	malayalam
anna oru poster aachu vidunga new year ku #et	tamil
vaathi coming response maa tanuku lo #masterfdfs	telugu

FIGURE: Snapshot of the Romanized Code-Mixed dataset.

Progress (Contd.)

fastText based Model

Reproduction and Training: Replicated Bhasha-Abhijnaanam((12)) using fastText for sentence-level language identification on native scripts.

fastText and Unicode-Range based Model

- **Preprocessing:** Sentences with multiple scripts are removed.
- **Script Classification:** A Unicode-Range based classifier identifies scripts such as Devanagari, Bengali, Arabic, or Latin. Other scripts, unique to one language due to dataset nuances, receive immediate identification at this stage.
- **fastText Training and Comparison:** fastText classifiers are trained on script-specific data. Use the script's corresponding classifier for efficient and optimal classification.

PROGRESS (CONTD.)

MODEL

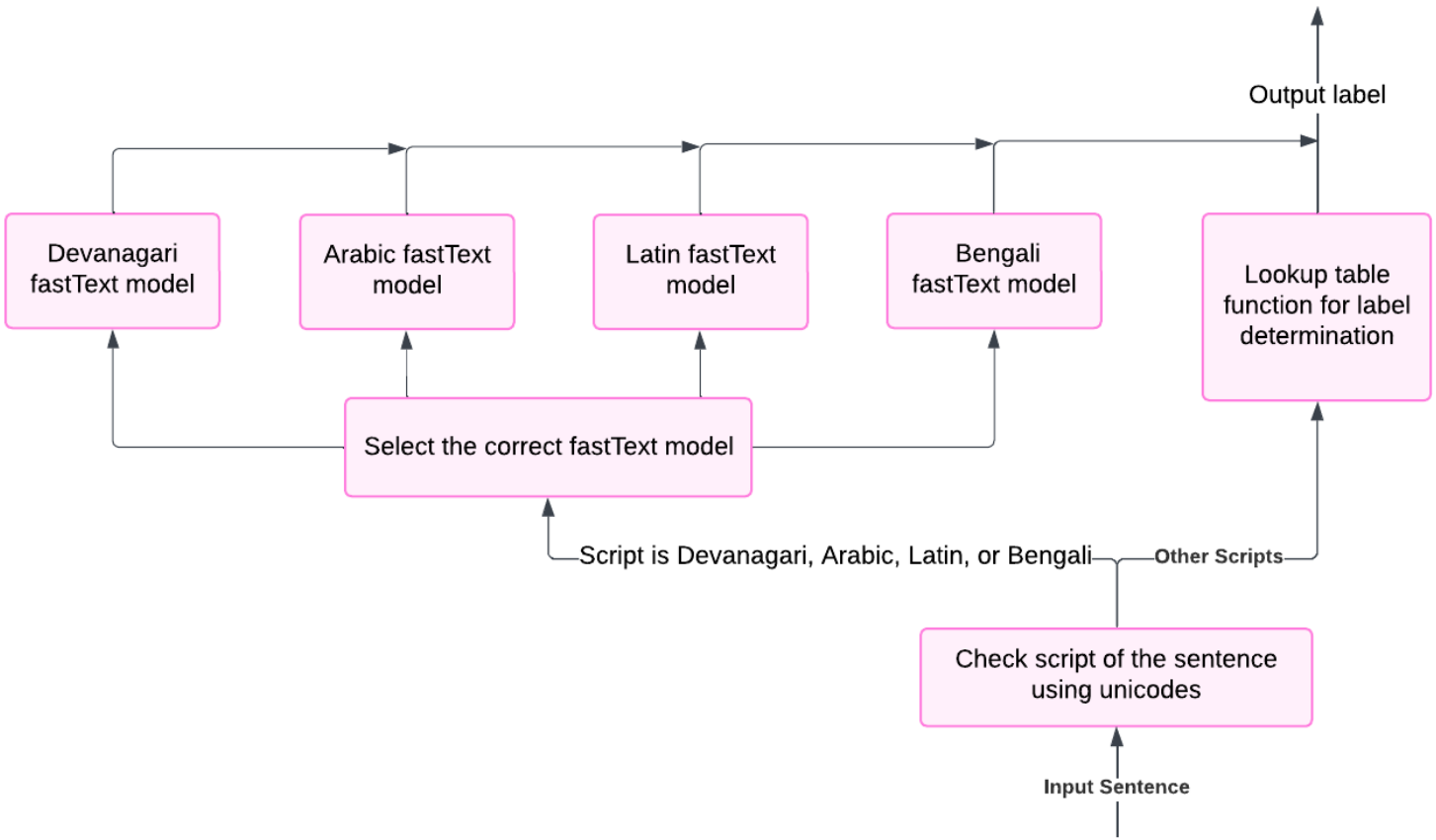

FIGURE: fastText + Unicode-Range based sentence based language identification for native languages in a single script dataset.

References

[1] J. Sharma, M. Gupta, and M. Choudhury, "Understanding script-mixing: A case study of hindi-english bilingual twitter users," in *Proceedings of the Second Workshop on Computational Approaches to Linguistic Code-Switching*. Association for Computational Linguistics, 2020, pp. 36–44. [Online]. Available: https://aclanthology.org/2020.calcs-1.5.pdf

[2] V. Srivastava and M. Singh, "Challenges and considerations with code-mixed nlp for multilingual societies," 2021. [Online]. Available: https://arxiv.org/pdf/2106.07823

[3] S. Rijhwani, R. Sequiera, M. Choudhury, K. Bali, and C. S. Maddila, "Estimating code-switching on twitter with a novel generalized word-level language detection technique," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017.

[4] T. Solorio, E. Blair, S. Maharjan, S. Bethard, M. Diab, M. Ghoneim, A. Hawwari, F. AlGhamdi, J. Hirschberg, A. Chang *et al.*, "Overview for the first shared task on language identification in code-switched data," in *Proceedings of the First Workshop on Computational Approaches to Code Switching*, 2014.

[5] J. Kreutzer, I. Caswell, Y. Wang, A. Wahab, and S. Wu, "Quality at a glance: An audit of web-crawled multilingual datasets," *arXiv preprint arXiv:2201.08239*, 2022.

[6] D. Kakwani, A. Kunchukuttan, D. Golla, A. Bhattacharjee, M. Khapra, and P. Kumar, "Indicnlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020.

[7] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," *arXiv preprint arXiv:1607.01759*, 2016.

[8] KPMG and Google, "Indian languages - defining india's internet," 2017, https://www.kpmg.com.

References (Contd.)

[9] S. Khanuja *et al.*, "Supervised models for multilingual token classification on low-resource indian languages," *arXiv preprint arXiv:2205.03983*, 2022.

[10] ——, "Muril: Multilingual representations for indian languages," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021.

[11] A. Dubey *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024. [Online]. Available: https://arxiv.org/abs/2407.21783

[12] Y. Madhani, M. M. Khapra, and A. Kunchukuttan, "Bhasha-abhijnaanam: Native-script and romanized language identification for 22 indic languages," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, 2023, pp. 816–826. [Online]. Available: https://aclanthology.org/2023.acl-short.71

Thank You