Code-Mixed Dialogue Systems: Bridging Sociolinguistics and AI

Chandravallika Murarisetty
University of Windsor
murarisc@uwindsor.ca

Saima Khatoon
University of Windsor
khatoons@uwindsor.ca

Rajkumar Pankajkumar Patel
University of Windsor
patel9qb@uwindsor.ca

ABSTRACT

Code-switching in conversational AI presents significant challenges in understanding and generating multilingual dialogue. This study focuses on improving code-switching performance through two key areas: leveraging Large Language Models (LLMs) and personalizing conversations. Targeting English, Hindi, and Hinglish languages, the research aims to enhance semantic understanding and language model adaptability by considering linguistic nuances, user demographics, and cultural contexts.

KEYWORDS

Code switch, Hinglish, LLM, Demographics, Dialogue System

1 INTRODUCTION

Code- switching, the practice of blending two or more languages within a single conversation, has become increasingly common in multilingual societies, particularly in informal communication on digital platforms. In conversational AI, understanding the context of a single sentence comprising multiple languages and responding vice-versa is a critical goal to achieve to increase efficiency and make the model more personalized. [1, 4] The pre-trained models such as mBert, BART, M2M, and T5 were used for multiple languages to achieve code-switching. However, understanding the semantics and applying the language rules to the models is always a complex task. [2] Incorporating the LLM with linguistic theory to increase the quality of the model is the research that is in great demand due to the enormous versatility of LLMs in the current market. To make the conversation more personalized, considering the user's age, linguistic proficiency, and understanding of the regional and cultural context is also important [2]. This study aims to improve the performance of code-switching by exploring the above three areas with the help of recently developed LLMs. In this study, we are focussing on the languages English, Hindi, and Hinglish because English and Hindi are two out of the ten most spoken languages all over the world[3].

2 MOTIVATION

A critical challenge in AI is ensuring accessibility for users from diverse linguistic backgrounds. However, most Large Language Models (LLMs) fail to dynamically adapt to these variations, often producing monolingual outputs or struggling with transitions.

This research builds upon [2] and explores how LLMs can adapt based on user demographics. We aim to develop a mechanism that identifies multiple languages (Hindi, English and Hinglish) and tailors responses to factors like linguistic proficiency, age, and regional context, enhancing the AI's natural, user-centric communication.

- **Linguistic Proficiency:** Assessing the user's fluency in each language and adjusting responses accordingly.
- **Age Considerations:** Recognizing variations in multilingual interaction preferences across different age groups.
- **Regional and Cultural Context:** Customizing language usage based on geographical and socio-cultural influences.

By integrating these factors, we propose an NLP-driven approach that enhances AI's ability to interact naturally with multilingual users.

2.1 Motivating Example

Consider Rohan, a bilingual user who switches between Hindi and English, known as Hinglish. When interacting with an AI assistant, he might say:

> *"Bhai, I was working on this project kal raat tak, aur ab mujhe ek solution chahiye."*

A conventional AI will respond entirely in either English or Hindi, failing to match Rohan's style. Most existing dialogue systems are monolingual and fail to account for the complexities of code-switching driven by factors like linguistic proficiency, age, and regional and cultural context.

3 PROBLEM DEFINITION

We aim to develop a model capable of performing code-switching, taking into account linguistic proficiency, age, and regional or cultural context. Such a model would adapt its code-mixing patterns to reflect the sociolinguistic factors influencing real-world communication. By creating or adapting datasets (e.g., CM-DailyDialog), leveraging multilingual transformer models like mBART[1], and establishing standardized evaluation metrics.

> *"Haan bhai, agar tune kal raat tak kaam kiya hai then you should take a break."*

This advancement will enable more natural, engaging, and contextually appropriate dialogue systems for multilingual societies, making conversational AI more inclusive and adaptable to diverse user needs.

4 TEAM JUSTIFICATION

- **Chandravallika** - Training LLMs, Benchmarking
- **Saima** - Code Switch framework, Fine Tuning Dataset
- **Rajkumar** - Demographics Integration, Linguistics

REFERENCES

[1] Vibhav Agarwal, Pooja Rao, and Dinesh Jayagopi. 2021. Towards Code-Mixed Hinglish Dialogue Generation. , 271-280 pages. https://doi.org/10.18653/v1/2021.nlp4convai-1.26

[2] Garry Kuwanto, Chaitanya Agarwal, Genta Winata, and Derry Wijaya. 2024. Linguistics Theory Meets LLM: Code-Switched Text Generation via Equivalence Constrained Large Language Models. https://doi.org/10.48550/arXiv.2410.22660

[3] Marcus Lu. 2024. MiscRanked: The Top Languages Spoken in the World. https://www.visualcapitalist.com/top-languages-spoken-in-the-world/ [Accessed: (Jan 23 2024)].

[4] Suneeta Thomas. 2021. Code-Switching in Spoken Indian English: A Case Study of Sociopolitical Talk. , 7-40 pages. https://doi.org/10.37834/JCP2141007t