REPRESENTATION LEARNING FOR CODE-MIXED LANGUAGES

Aditeya Baral, Allen George Ajith, Mrityunjay Abhijeet Bhanja, Roshan Nayak
New York University
{ab12057, aa12938, mb9348, rn2588}@nyu.edu

ABSTRACT

Globalisation and cultural adaptation have fostered the growth of multilingualism in the modern world, leading to a widespread phenomenon of *code-mixing* (CM), where multiple languages are mixed within a single utterance with the points of inflection being known as *switching points* (SP). In multilingual countries like India, the fusion of Hindi and English (popularly termed *Hinglish*) has become increasingly prevalent in conversational usage both online and offline. Recent studies indicate that the *Hinglish* population has grown steadily at an annual rate of 1.2% with a 2% yearly increase in online usage[1].

As Large Language Models (LLMs) become increasingly integrated into our daily lives, recent years have seen the development of multiple LLMs for a range of diverse languages, including high-resource ones such as English[2] and French [3]. Despite the steady increase in CM speakers today, most LLMs still fail to serve CM needs[4, 5]. This is evidenced by the subpar performance on real-world tasks requiring encoder-only models such as classification, semantic search and information retrieval. Their limitations can be traced back to their inadequate and poor neural representations of these languages [6, 7]. The unique linguistic structure of CM languages makes it a challenge to represent them effectively, limiting their efficacy in multilingual contexts.

Previous approaches have used multilingual models to bridge the gap between the languages in a CM language. However, studies have shown that CM languages being *inherently* multilingual, cannot be replaced by multilingual models since they are not natural code-mixers [8] of monolingual languages. The lack of formal grammar, frequent occurrence of switching points, spelling variants and contextual nuances are some of the problems posed by CM languages. Additionally, models pre-trained on CM data have surprisingly shown inadequate improvements over multilingual models despite increasing the number of parameters [9, 10]. Since scaling up both data and parameters has only yielded marginal improvements, recent studies have moved away from the former and established the necessity of specialised pre-training techniques for CM languages.

Current approaches focus on capturing the linguistic complexities of switching points [11, 12, 13] in a CM sentence. We build upon these and propose a new pre-training approach for neural language modelling of CM text representations. First, we introduce new training objectives like predicting switches, CM Index [14] and cross-lingual alignments to learn grammatical nuances and switching point behaviour. Second, we implement core architectural modifications to the Transformer architecture to encode switching point information and introduce a novel dual-decoder training setup for bilingual language translation modelling (BiLTM). All models will be pre-trained on the L3Cube-HingCorpus[9] dataset and evaluated across multiple classification benchmarks like ICON [15], HASOC [16] and SentiMix [17].

As we will be working with encoder-based models (each with around 600M parameters), the project is computationally feasible requiring not more than two GPUs. With this project, we hope to address a critical research gap by developing specialized pre-training techniques that enable models to support code-mixed and underserved languages effectively. Our approach democratizes language technology and makes it more inclusive by bridging the performance gap in multilingual applications.

References

[1] A. Sengupta, S. Das, M.S. Akhtar, and et al. Social, economic, and demographic factors drive the emergence of Hinglish code-mixing on social media. *Humanities and Social Sciences Communications*, 2024.

[2] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, and Radu Soricut et. al. Gemini: A family of highly capable multimodal models, 2024.

[3] Manuel Faysse, Patrick Fernandes, Nuno M. Guerreiro, António Loison, Duarte M. Alves, Caio Corro, Nicolas Boizard, João Alves, Ricardo Rei, Pedro H. Martins, Antoni Bigata Casademunt, François Yvon, André F. T. Martins, Gautier Viaud, Céline Hudelot, and Pierre Colombo. Croissantllm: A truly bilingual french-english language model, 2024.

[4] Varad Srivastava. DweshVaani: An LLM for detecting religious hate speech in code-mixed Hindi-English. In Kengatharaiyer Sarveswaran, Ashwini Vaidya, Bal Krishna Bal, Sana Shams, and Surendrabikram Thapa, editors, *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 46–60, Abu Dhabi, UAE, January 2025. International Committee on Computational Linguistics.

[5] Dilip Venkatesh, Pasunti Prasanjith, and Yashvardhan Sharma. BITS pilani at SemEval-2024 task 10: Fine-tuning BERT and llama 2 for emotion recognition in conversation. In Atul Kr. Ojha, A. Seza Doğruöz, Harish Tayyar Madabushi, Giovanni Da San Martino, Sara Rosenthal, and Aiala Rosá, editors, *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 811–815, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[6] Debajyoti Mazumder, Aakash Kumar, and Jasabanta Patro. Revealing the impact of synthetic native samples and multi-tasking strategies in hindi-english code-mixed humour and sarcasm detection, 2024.

[7] Shruti Jagdale, Omkar Khade, Gauri Takalikar, Mihir Inamdar, and Raviraj Joshi. On importance of code-mixed embeddings for hate speech identification, 2024.

[8] Ruochen Zhang, Samuel Cahyawijaya, Jan Christian Blaise Cruz, Genta Indra Winata, and Alham Fikri Aji. Multilingual large language models are not (yet) code-switchers, 2023.

[9] Aryan Patil, Varad Patwardhan, Abhishek Phaltankar, Gauri Takawane, and Raviraj Joshi. Comparative study of pre-trained bert models for code-mixed hindi-english data. In *2023 IEEE 8th International Conference for Convergence in Technology (I2CT)*, page 1–7. IEEE, April 2023.

[10] Sebastin Santy, Anirudh Srinivasan, and Monojit Choudhury. BERTologiCoMix: How does code-mixing interact with multilingual BERT? In Eyal Ben-David, Shay Cohen, Ryan McDonald, Barbara Plank, Roi Reichart, Guy Rotman, and Yftah Ziser, editors, *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 111–121, Kyiv, Ukraine, April 2021. Association for Computational Linguistics.

[11] Mohsin Ali, Kandukuri Sai Teja, Sumanth Manduru, Parth Patwa, and Amitava Das. Pesto: Switching point based dynamic and relative positional encoding for code-mixed languages, 2021.

[12] Mohsin Ali, Kandukuri Sai Teja, Neeharika Gupta, Parth Patwa, Anubhab Chatterjee, Vinija Jain, Aman Chadha, and Amitava Das. Conflator: Incorporating switching point based rotatory positional encodings for code-mixed language modeling, 2023.

[13] Ayan Sengupta, Tharun Suresh, Md Shad Akhtar, and Tanmoy Chakraborty. A comprehensive understanding of code-mixed language semantics using hierarchical transformer, 2022.

[14] Vivek Srivastava and Mayank Singh. Challenges and limitations with the metrics measuring the complexity of code-mixed text, 2021.

[15] Braja Gopal Patra, Dipankar Das, and Amitava Das. Sentiment analysis of code-mixed indian languages: An overview of sail_code-mixed shared task @icon-2017, 2018.

[16] Sarah Masud, Mohammad Aflah Khan, Md. Shad Akhtar, and Tanmoy Chakraborty. Overview of the hasoc subtrack at fire 2023: Identification of tokens contributing to explicit hate in english by span detection, 2023.

[17] Parth Patwa, Gustavo Aguilar, Sudipta Kar, Suraj Pandey, Srinivas PYKL, Björn Gambäck, Tanmoy Chakraborty, Thamar Solorio, and Amitava Das. SemEval-2020 task 9: Overview of sentiment analysis of code-mixed tweets. In Aurelie Herbelot, Xiaodan Zhu, Alexis Palmer, Nathan Schneider, Jonathan May, and Ekaterina Shutova, editors, *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 774–790, Barcelona (online), December 2020. International Committee for Computational Linguistics.