Enhancing Machine Translation for Code-Mixed Language Text

Kishansinh Jitendrasinh Rathod
krathod@usc.edu

Nikhil Bola Kamath
nikhilbo@usc.edu

Pavle Medvidovic
medvidov@usc.edu

Steven Melgar
stmelgar@usc.edu

Tejas Jambhale
tjambhal@usc.edu

Xingyu Zhao
xzhao911@usc.edu

Abstract

Machine translation is a rapidly evolving field in Natural Language Processing (NLP) with applications ranging from improving cross-lingual communication to assisting global business operations. In this project, we address the complex challenge of translating code-mixed language text. Our aim is to develop a robust system that can accurately translate mixed language text (Spanglish and Hinglish) into English, recognizing the nuances of language switching within a sentence. With bilingualism being a significant aspect of many team members' lives, this project has a personal and practical relevance that extends beyond the research realm. Moreover, this work serves as a proof of concept, potentially paving the way for enhanced mixed language translation in various applications.

1 Project Domain and Goals

Multilingualism and code-mixing have become standard communication features in our increasingly interconnected world. Code-mixing refers to seamless integration of elements from multiple languages within a single utterance or text. "Spanglish," a prime example, is a hybrid language used by bilingual speakers of English and Spanish. These instances pose a unique challenge for machine translation systems, as they require translation and language identification within a sentence. With this problem at hand, we aim to build an end-to-end standalone translation model in this project.

Why NLP (Natural Language Processing)? NLP is a multidisciplinary field that combines linguistics, computer science, and artificial intelligence to enable computers to understand, interpret, and generate human language. Its relevance transcends traditional machine translation, encompassing various applications such as sentiment analysis, chatbots, and language modeling. In this project, NLP is the indispensable tool that enables us to address the complexities of code-mixed language translation.

Machine translation, a classic application of NLP, is pivotal in bridging language barriers. While automated translation services are prevalent on social media platforms, search engines, and websites, they often falter when confronted with code-mixed data. This project attempts to demonstrate the feasibility of accurately translating code-mixed language text (Spanglish and Hinglish), primarily focusing on translating it into English.

Upon completing this project, we can extrapolate this idea to create an end-to-end translation model for speech instead of text. From an application point of view, we can eliminate the language barrier prevalent in conversations in international conferences to our daily discussions. This project would also serve as a foundational model for the cross-linguistic translation.

2 Related Work

Gautam et al. (2021), in their work, use mBART and fine-tune it to achieve Hinglish to English translation. Yet another piece that motivates this project is by (Huber et al., 2022). This work introduces us to an end-to-end speech translation model called LAST, which converts a code-mixed audio signal into a target language. (Pratapa et al., 2018) presents how embeddings can help us improve syntactic and semantic code-mixed processing tasks. (Sitaram et al., 2019) discusses various computational approaches for solving the problem of code-switched text. Generating a dataset is a crucial step towards training the model. Papers by (Dhar et al., 2018) and (Srivastava and Singh, 2020) explain methods that can be used to create/annotate datasets. (Samanta et al., 2019) is another novel work that aims at generating large volumes of language-tagged code-switched text. In this project, we aim to leverage large language models such as those mentioned in the work discussed above, along with improved contextual embeddings to achieve better code-mixed translations.

3 Datasets

Several datasets online can be used to achieve the task of code-mixed translation. Most of them highlight Spanglish and Hinglish to English mappings. In this project, we will be mainly using the data points found in LinCE by (Aguilar et al., 2020). Furthermore, once we develop an end-to-end pipeline, we look forward to scrapping data from various news articles and blogs from websites/online publications and creating custom Spanglish-English and Hinglish-English datasets. The dataset cited in this proposal shows that a significant amount of preprocessing is required before training the model. We plan on leveraging existing tool such as NLTK to perform the preprocessing task. Removing emojis, punctuations, words with less frequent occurrences, URLs, and incorrect spellings are some of the steps that may be necessary. Further, we tokenize the text into sentences and words to facilitate processing. This is especially important in the case of datasets extracted from Twitter, which often include slang and niche words that add noise. We ensure that all input text is in Roman text and that the input code mixed sentence has words from both languages.

4 Technical Challenge

One of the many problems faced in code-mixed translation is the ambiguity in the language detected. Code-mixed text often lacks a clear distinction between languages, making it challenging for machine translation systems to determine which language a particular word or phrase belongs to. For example, consider the Spanglish phrase "Voy a book the flight." Here, "Voy a" is Spanish for "I am going to," while "book" and "flight" are English words. Identifying the language of each segment is non-trivial. Also, code-mixed sentences can introduce subtle shifts in context and meaning. Translating individual words or phrases without considering the context can lead to inaccurate translations; for instance, the term "vuelo" in Spanish can mean both "flight" and "I fly." A machine translation system must accurately disambiguate between these meanings depending on the context. Another issue is the cross-lingual variation; different languages have distinct grammatical rules, word orders, and idiomatic expressions. When code-mixing occurs, these variations must be accounted for. For example, English typically follows a subject-verb-object (SVO) word order, while Hindi can use the subject-object-verb (SOV) word order. A translation system needs to adapt to these variations for accurate translation. We plan on building custom contextual embeddings for the code-mixed texts to address the above issues, followed by a large language model to help us in machine translation. This way, our work goes well beyond the course curriculum and the existing work. Yet another technical issue is the amount of resources needed to train the model. Large language models are resource-hungry and need several gigabytes of memory for training/inferencing. This project also aims to address this problem by employing methods like pruning and quantization while maintaining the desired metric.

Different metrics can used to evaluate the model's performance. We assess the performance of our system using standard metrics such as BLEU (Bilingual Evaluation Understudy) and human evaluation. Additionally, we evaluate its effectiveness on real-world hand-crafted and web-scraped Spanglish/Hingligh text samples to measure its practical utility.

For any Deep-learning task, a critical step is getting the appropriate data and making it model-friendly; this may include actions such as data cleaning, performing exploratory data analysis (EDA), and some feature engineering. We have decided that Pavle, Xingyu, and Kishansinh will focus on getting this task done. Parallely, Nikhil, Tejas, and Steven will work on getting an end-to-end pipeline ready for this problem statement. It includes designing data generators, training, validating, and testing/inferencing modules. We intend to finish this task (data gathering and processing along with end-to-end pipeline design) by the end of October. Upon reaching a satisfactory level, all of us will work together to ingest data into our model, which should happen in the first week of November. After this, we will work on writing reports and continuously improving the model's performance on a rotational basis (starting the second week of November). This work division is tentative and subject to change. The author's naming in this work follows the alphabetical order.

References

Gustavo Aguilar, Sudipta Kar, and Thamar Solorio. 2020. LinCE: A Centralized Benchmark for Linguistic Code-switching Evaluation. In *Proceedings of The 12th Language Resources and Evaluation Con-*

ference, pages 1803–1813, Marseille, France. European Language Resources Association.

Mrinal Dhar, Vaibhav Kumar, and Manish Shrivastava. 2018. Enabling code-mixed translation: Parallel corpus creation and mt augmentation approach. In *Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing*, pages 131–140.

Devansh Gautam, Prashant Kodali, Kshitij Gupta, Anmol Goel, Manish Shrivastava, and Ponnurangam Kumaraguru. 2021. Comet: Towards code-mixed translation using parallel monolingual sentences. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 47–55.

Christian Huber, Enes Yavuz Ugan, and Alexander Waibel. 2022. Code-switching without switching: Language agnostic end-to-end speech translation. *arXiv preprint arXiv:2210.01512*.

Adithya Pratapa, Monojit Choudhury, and Sunayana Sitaram. 2018. Word embeddings for code-mixed language processing. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 3067–3072.

Bidisha Samanta, Sharmila Reddy, Hussain Jagirdar, Niloy Ganguly, and Soumen Chakrabarti. 2019. A deep generative model for code-switched text. *arXiv preprint arXiv:1906.08972*.

Sunayana Sitaram, Khyathi Raghavi Chandu, Sai Krishna Rallabandi, and Alan W Black. 2019. A survey of code-switched speech and language processing. *arXiv preprint arXiv:1904.00784*.

Vivek Srivastava and Mayank Singh. 2020. Phinc: A parallel hinglish social media code-mixed corpus for machine translation. *arXiv preprint arXiv:2004.09447*.