Lost in Translation No More: Fine-tuned transformer-based models for CodeMix to English Machine Translation

Arindam Chatterjee[1,2], **Chhavi Sharma**[1], **Yashwanth V.P.**[1], **Niraj Kumar**[1], **Ayush Raj**[1], **Asif Ekbal**[2]
[1]Wipro Research, Lab45, Bangalore, India
[2]Indian Institute of Technology, Patna, India
{arindam.chatterjee4, chhavi.sharma5, yashwanth.p54, niraj.kumar63, ayush.raj3}@wipro.com,
asif@iitp.ac.in

Abstract

Codemixing, the linguistic phenomenon where a speaker alternates between two or more languages within a conversation or even a single utterance, presents a significant challenge for machine translation systems due to its syntactic complexity and contextual nuances. This paper introduces a set of advanced transformer-based models fine-tuned specifically for translating codemixed text to English, more specifically, Hindi-English (colloquially referred to as "Hinglish") codemixed text into English. Unlike standard bilingual corpora, codemixed data requires an understanding of the intricacies of grammatical structures and cultural contexts embedded within the language blend. Existing machine translation efforts in codemixed languages have largely been constrained by the paucity of robust datasets and models that can capture the nuanced semantic and syntactic interplay characteristic of such languages. We present a novel dataset $\text{PACMAN}_{\text{trans}}$ for Hinglish to English machine translation, based on the PACMAN strategy, meticulously curated to represent natural codemixing patterns. Our generic fine-tuned translation models trained on the novel data outperforms current state-of-the-art Large Language Models (LLMs) by **38%** in terms of BLEU score. Further, when fine-tuned on custom benchmark datasets, our focused dual fine-tuned models surpass the PHINC dataset BLEU score benchmark by **22%**. Our comparative analysis illustrates significant improvements in translation quality, showcasing the potential of fine-tuning transformer models in bridging the linguistic divide in codemixed language translation. The success of our models reflects a promising step forward in the quest to provide seamless translation services for the ever-growing multilingual population and the complex linguistic phenomena they generate.

1 Introduction

Codemixing, the grammatical fusion of two or more languages within a single utterance or discourse, is a pervasive linguistic practice among bilingual and multilingual communities (Myers-Scotton, 1995). It is a natural outcome of language contact, often observed in societies where speakers are fluent in both a local and a global language. The Hindi-English codemixed language, widely known as "Hinglish," is one such instance where the syntactic, morphological, and lexical elements of Hindi and English are interwoven, giving rise to a rich tapestry of linguistic expression (Bhatia and Ritchie, 2013).

Machine translation (MT) systems have traditionally been developed for well-defined language pairs with substantial parallel corpora. However, the translation of codemixed text poses unique challenges due to the absence of consistent grammatical rules and the complexities introduced by the informal and spontaneous nature of codemixing (Singh, 2018). The machine translation of such codemixed languages is relatively nascent and has been gaining traction with the rise of neural machine translation (NMT) models that are better at handling linguistic ambiguities (Johnson et al., 2017).

Despite the advances in NMT, the translation quality for codemixed languages, particularly for Hinglish to English, remains suboptimal. The existing models struggle with capturing the nuanced interplay of linguistic features from both languages and often fail to maintain the semantic integrity of the source text (Pratapa et al., 2018). Recent efforts in this domain have focused on creating more adept systems through innovations in model architecture and data resources (Aguilar et al., 2018).

In this work, we extend the existing body of research by introducing a suite of fine-tuned transformer-based models tailored for the Hinglish to English translation task. Our approach bene-

fits from a novel dataset specifically curated for Hinglish, using the PACMAN strategy (Chatterjee et al., 2022), reflecting various codemixing patterns that are representative of authentic speech and writing in naturally observed contexts. Our proposed fine-tuned models outperforms GPT4 by **38%** and the SOTA benchmark in Hinglish to English translation task, defined by the PHINC dataset (Khanuja et al., 2020) by **22%** in terms of BLEU Score, to set a new standard for the field.

2 Related Work

The translation of codemixed text is an emerging field of study within the domain of natural language processing. Initial attempts to address the translation of mixed-language text primarily focused on rule-based systems, which quickly proved to be insufficient due to the unpredictable nature of code-switching and mixing (Dhar et al., 2018). With the advent of statistical machine translation (SMT), researchers began exploring data-driven approaches, although the scarcity of parallel corpora for codemixed languages remained a hindrance (Solorio and Liu, 2008).

The paradigm shift towards neural machine translation (NMT) has opened up new avenues for handling the complexities of codemixed language translation. The flexibility of neural networks, particularly the sequence-to-sequence models, has shown promise in capturing the nuances of mixed-language syntax (Singh and Shrivastava, 2018). Transformer-based architectures, introduced by Vaswani et al. (2017), have revolutionized NMT by enabling models to consider the entire context of the input sequence, which is particularly beneficial for the disambiguation of codemixed text (Pratapa et al., 2018).

In the context of Hinglish to English translation, Khanuja et al. (2020) introduced the PHINC dataset, a benchmark for codemixed machine translation. While several works have utilized this dataset, they have often fallen short in adequately handling the linguistic subtleties of Hinglish (Srivastava et al., 2020). Our work builds on these foundations and introduces improvements both in terms of the dataset and the transformer model fine-tuning, leading to significant advancements over the current state-of-the-art.

Another line of work that intersects with our research is the exploration of pre-trained language models for codemixed language processing. Models such as mBERT (Devlin et al., 2019a) and XLM-R (Conneau et al., 2020) have been fine-tuned for various codemixed NLP tasks with encouraging results (Aguilar et al., 2020). However, their direct application to machine translation for codemixed languages is still an under-researched area, which our study aims to address.

2.1 Transformer Models in NMT

Transformers have now become the de-facto standard in NMT due to their superior performance in comparison to previous RNN and CNN based models (Vaswani et al., 2017). The self-attention mechanism inherent to transformers allows for a more nuanced understanding of the source language, a feature that is incredibly beneficial when dealing with the complexities of codemixing (Winata et al., 2019).

2.2 Datasets for Codemixed Translation

One of the primary challenges in machine translation for codemixed languages is the lack of high-quality, large-scale datasets (Bali et al., 2014). While synthetic datasets have been proposed to augment the available data (Rijhwani et al., 2020), they often fail to capture the authentic use of language in natural settings. Our novel dataset contributes to filling this gap, providing a diverse and representative corpus for Hinglish to English translation.

In sum, our work not only extends the current literature in codemixed machine translation but also addresses the limitations of existing datasets and models. By leveraging the transformer architecture's strengths and introducing a novel, more representative dataset, we aim to push the boundaries of what is currently possible in the translation of Hinglish text.

3 Dataset

As already discussed in the preceding sections, our dataset lies at the core of this work. We created a novel custom dataset generated using the PACMAN strategy, as originally proposed by (Chatterjee et al., 2022). Unlike the authors of the PACMAN dataset, we used the Samanantar Parallel Corpus (Ramesh et al., 2021), which consists of a substantially larger *94 million* data samples, for generating our novel English-Hinglish Parallel Corpus. We named the dataset $\text{PACMAN}_{\text{trans}}$.

The PACMAN strategy uses English-Hindi parallel sentences to generate Hinglish codemix sentences following the Matrix Language Theory

(Joshi, 1982). We apply the same strategy and pick the English source sentence and the resulting Hinglish sentence as parallel sentences for our translation task. To ensure data quality and linguistic fidelity for PACMAN$_{trans}$, we performed several pre-processing steps, as outlined below:

Dataset	PACMAN$_{trans}$	PHINC
# samples	5866702	13738
Average sentence length	13.94	12.31
# samples acc. to sentence length		
0-5	404669	1858
6-10	1937302	4713
11-15	1538731	3245
16-20	1010481	2076
21-25	731336	1203
26 and above	244183	643

Table 1: Comparison of statistics between PACMAN$_{trans}$ and PHINC codemixed datasets. The key parameter to note here is the average sample length.

1. **Deduplication:** To eliminate redundancy and ensure diversity in the dataset, we conducted a deduplication process to remove duplicate sentences.

2. **Language Consistency:** Given the bilingual nature of the corpus (comprising English and Hindi sentences), we eliminated parallel sentence pairs where English words appeared within Hindi sentences. This step ensured that each sentence pair maintained the integrity of its respective language.

3. **POS Annotation:** We employed the Stanza tool (Qi et al., 2020) to perform Part-of-Speech (POS) annotation on both the source and target sentences. This enhanced the linguistic information available for further analysis.

4. **Alignment Generation:** Building upon the POS-annotated sentences, we generated alignments using Fast-Aligner proposed in (Dyer et al., 2013) between the words in the matrix language (Hindi) and Embedded Language (English) sentences. These alignments facilitated subsequent transformations. In our research, we have observed that Hindi is used as the matrix language in an overwhelming majority of Hinglish sentences. Consequently, we nominated Hindi as the Matrix Language, for our data generation process.

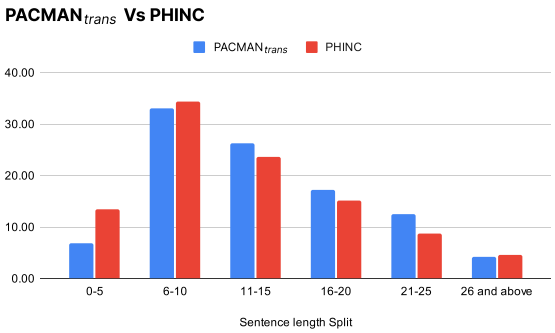

Figure 1: Percentage-wise sample distribution based on sentence length in PACMAN$_{trans}$ and PHINC datasets

.

Following the initial pre-processing steps, we adopted the PACMAN strategy to handle word replacement. Specifically, we focused on words in the dataset that exhibited one-to-one mapping between the Matrix Language and the Embedded Language, with both words falling into the categories of NOUN or ADJ. For such instances, we replaced these words to ensure the linguistic compatibility and cohesiveness of the dataset. This methodology also ensures coverage of almost all consistent codemixing patterns, making it authentic in terms of alignment to naturally observed codemixing.

The PACMAN$_{trans}$ dataset consisted of sizeable ***5.8 million*** entries, exclusively comprising codemix sentences *without* monolingual samples, with an average sentence length of *13.94* as shown in Table 1.

We conducted a comparative analysis of the data distribution between PACMAN$_{trans}$ and PHINC, specifically focusing on the sentence length. Figure 1 provides a graphical representation of how these two datasets differ with respect to the distribution of samples across various sentence lengths. It is evident from the graphical representation that the PACMAN$_{trans}$ dataset predominantly consists of samples falling within the 11 to 25 sentence length range. This observation emphasizes the suitability of PACMAN$_{trans}$ as a valuable dataset for real-life scenarios, given that a significant proportion of sentences in such scenarios tend to fall within this particular length range.

4 Experimental Setup

In several related works discussed in section 2, specifically with the closest body of research (Srivastava et al., 2020), the authors have fine-tuned pre-trained models like mBART, mT5, *etc.* on the

PHINC dataset. Since the $\text{PACMAN}_{\text{trans}}$ dataset (5.2M samples) is significantly larger (around 45 times) than the PHINC dataset (13K samples), we opted to first train domain-agnostic or generic models for our translation task. This stems from the fact that the Samanantar Dataset spans several domains (Ramesh et al., 2022).

Consequently, in order to compare against the PHINC benchmark we followed a *dual fine-tuning* strategy by fine-tuning our domain-agnostic translation models on the PHINC dataset. Through this strategy, we were able to generate two sets of transformer-based models, general-purpose and focused. We also observed the role of dataset size for the Codemix to English translation task, by compiling separate models for varied dataset sizes.

4.1 Models

As mentioned, we explored various transformer-based models for the translation task. We built different versions of these models as well based on the dataset size (train/validation/test) they were built on. The transformer-based models we fine-tuned are outlined below:

1. **T5**: The T5 model, introduced by Raffel et al. (2020), stands for "Text-to-Text Transfer Transformer" and represents a unifying framework that converts all NLP problems into a text-to-text format. The ingenuity behind T5 is the simplification of the NLP pipeline, where tasks like translation, question answering, and classification are all framed as generating text from text. This approach enables the model to use the same model, loss function, hyperparameters, *etc.*, across a diverse range of tasks, potentially simplifying the process of training and deploying NLP models (Raffel et al., 2020). T5's performance on benchmark datasets has set new standards, particularly on the GLUE and SuperGLUE benchmarks, which are designed to test the limits of NLP models' understanding capabilities (Wang et al., 2019b,a). This illustrates the model's generalization capabilities and its potential as a powerful tool in the field of natural language processing.

   In our experimental evaluation, as presented in Table2, we have leveraged two distinct configurations of the T5 model. These configurations are denoted as T5_1 and T5_2. T5_1 was trained on a dataset comprising *2M* samples, using a training duration of *20* epochs and an initial learning rate of *1e-3*. In contrast, T5_2 underwent training on a larger dataset consisting of *5.3M* samples. The training for T5_2 also spanned *20* epochs but employed a lower initial learning rate of *1e-4*. These configurations were instrumental in our investigations to analyze the model's performance under varying training conditions.

2. **mT5**: Building upon the success of the original T5 model, Google introduced mT5, a multilingual variant designed to handle tasks across multiple languages (Xue et al., 2021). The mT5 model extends the text-to-text framework of T5 to over 100 languages, pre-trained on a multi-lingual dataset derived from the Common Crawl corpus. This adaptation allows for the transfer of knowledge across languages and tasks, benefiting especially low-resource languages that typically do not have large dedicated datasets (Xue et al., 2021).

   In our experimental setup, we employed the mT5 small model, which was meticulously trained over a span of *20* epochs on our $\text{PACMAN}_{\text{trans}}$ dataset comprising 500K samples. The training process was carried out with a learning rate of *1e-4,* encapsulating the critical configuration aspects of our model training.

3. **NLLB**: META's No Language Left Behind (NLLB) project aims to develop a model that offers high-quality machine translation capabilities for the vast majority of the world's languages, including those that are low-resource and typically underrepresented in NLP research (Costa-jussà et al., 2022). The NLLB model has been trained on a dataset comprising text from 200 languages, with a focus on inclusivity and language equity. It represents a significant step towards breaking the language barriers in global communication and information access (Costa-jussà et al., 2022).

   In a parallel exploration akin to our investigation of T5 model variants, we also examined different configurations of NLLB as shown in Table 2. As previously mentioned, NLLB exhibits proficiency in understanding context across approximately 200 languages, with a focus on low-resource languages (Goyal et al., 2022). Our initial NLLB-based model, named

$NLLB_1$, involved fine-tuning the model with a modest dataset of *10k* samples over *15* epochs. Building on this, we expanded the dataset to *20k* samples and increased the training duration by *10* additional epochs, yielding $NLLB_2$. Notably, the observed efficiency improvements encouraged us to further extend our exploration. Thus, we augmented the dataset to *100k* samples and conducted *30* training epochs to create $NLLB_3$. This progressive approach allowed us to harness the potential of the NLLB model effectively on fewer samples than the T5 variants.

4.2 Evaluation Metrics

1. **Lexical Based Evaluation**: For lexical evaluation (word-based evaluation) we used the **BLEU Score** or Bilingual Evaluation Understudy score, introduced by (Papineni et al., 2002). It is a well-established metric for evaluating the quality of machine-generated text, such as translations. This metric measures the similarity between the generated text and reference text(s) using n-gram precision and incorporates a brevity penalty to account for the length of the generated output. BLEU has become a standard benchmark for assessing the effectiveness of machine translation systems and has been widely adopted in the natural language processing community. It provides an objective and quantitative measure of the quality of the generated text, allowing for rigorous and reproducible evaluation of language generation tasks.

2. **Context Based Evaluation:** In the context of machine translation evaluation, BLEU has long been regarded as a prominent metric for assessing the quality of translations. However, our investigation, focused on codemixed sentences, revealed an intriguing challenge. When dealing with CM text, a single sentence can have multiple correct monolingual translations. These translations are semantically accurate but may exhibit less similarity to the ground truth sentence, resulting in relatively low BLEU scores for otherwise proficient models. In light of this observation, we extended our evaluation beyond BLEU and incorporated an assessment of the semantic similarity or relevance of the translations generated by the model with the ground truth. To achieve this, we harnessed BERT-based embeddings (Devlin et al., 2019b), capitalizing on BERT's contextual understanding of words in English sentences. Unlike models such as Word2Vec, where each word possesses a fixed representation, BERT dynamically adjusts word representations based on their surrounding context. By computing the cosine similarity between these embeddings, we introduced a dual evaluation approach that allows for a more comprehensive assessment of translation quality in the intricate landscape of codemix translation evaluation. This metric draws resemblance with the $\textbf{BERT}_{\textbf{score}}$ metric introduced by (Zhang et al., 2019), but differs in the fact that our context-based metric is sentence/paragraph based, rather than word-based as in $\text{BERT}_{\text{score}}$. We named this metric $\textbf{BERT}_{\textbf{SemRel}}$.

4.3 Data for Evaluation

1. $\textbf{PACMAN}_{\textbf{trans}}$ **unobserved data:** For evaluating our domain-agnostic translation models, we handpicked around 500 samples that constitute a distinct subset of the $\text{PACMAN}_{\text{trans}}$ dataset, deliberately chosen to be unfamiliar to the model. We took care to ensure that the training data did not overlap with this subset. These samples encompass a wide range of characteristics, including ones that are straightforward, intricate, and tied to specific domains. This diversity is crucial for a comprehensive evaluation of the model's ability to handle various types of data. We also selected these samples to pan across different sentence lengths, CMI values, and representative of natural codemixing, ensuring fairness in the experiments conducted. We call this dataset $\text{PACMAN}^{unobserved}_{trans}$. We conducted a comparative analysis of our generic translation models against state-of-the-art LLMs *viz.*, Google NMT (Wu et al., 2016), GPT3.5, and GPT4 (OpenAI, 2023) as baselines.

2. **PHINC:** Proposed by (Srivastava and Singh, 2020), PHINC is a valuable resource comprising Hinglish with English translation pairs. This dataset encompasses approximately 13K parallel pairs, all presented in a Romanized script. In addition, the authors have thoughtfully provided a set of 1.5K samples specifically for testing. This dataset serves as a

Models	PACMAN$^{unobserved}_{trans}$ data				
	BERT$_{\text{SemRel}}$	**BLEU-1**	**BLEU-2**	**BLEU-3**	**BLEU-4**
T5$_1$	0.871	42.009	28.93	20.918	15.644
T5$_2$	0.9534	**79.53**	**72.868**	**66.839**	**61.43**
mT5	**0.9723**	78.8508	70.4945	62.3976	55.2857
NLLB$_1$	0.9565	67.9977	57.1737	47.9628	40.3967
NLLB$_2$	0.9626	71.8366	61.6381	52.6104	45.1445
NLLB$_3$	0.9675	75.7942	66.321	57.5663	50.1878
Google NMT	0.8215	32.411	20.095	13.823	9.876
GPT3.5	0.9168	49.528	37.074	28.414	22.171
GPT4	0.933	57.347	45.06	35.928	28.847

Table 2: Performance of our cross-domain translation models against state-of-the-art Translation Engines and Large Language Models on based on PACMAN$^{unobserved}_{trans}$ data based on BLEU Scores and BERT$_{\text{SemRel}}$ as discussed in section 4.2. GPT4 exhibits the best performance among the LLMs. T5$_2$ stands out in terms of BLEU score. mT5 shows the highest BERT$_{\text{SemRel}}$ score. NLLB$_3$ although trained on significantly smaller data, performs at par with T5$_2$ and mT5 for both metrics.

benchmark, facilitating the evaluation of machine translation (MT) systems in the challenging domain of Hinglish to English translation. As already mentioned previously in this section, we use our *dual fine-tuning* strategy (fine-tune our domain-agnostic model) on the PHINC dataset, for comparative benchmarking of our translation models.

5 Results and Observations

The results of the performances of the models proposed in section 4.1 on the evaluation datasets discussed in section 4.3 are shown in Table 2 and 4.

5.1 Domain-agnostic Translation Models

In our study, we conducted an evaluation of several transformer-based models on our novel PACMAN$_{\text{trans}}$ dataset, as detailed in Table2. Notably, our analysis reveals intriguing insights into the performance of these models. Despite being trained on a relatively smaller dataset of 500K samples, mT5 demonstrates superior performance in terms of semantic relevance. This can be attributed to its multilingual proficiency, allowing it to effectively capture contextual nuances across languages. Conversely, T5$_2$ outshines in terms of BLEU scores, which rely on n-grams, measuring the alignment between generated text and ground truth based purely on word sequence similarity. It is also worth noting the comparative performance of T5 and mT5, wherein mT5 holds its ground in terms of BLEU scores despite its smaller training dataset.

Furthermore, our exploration extends to different configurations of the NLLB model, performaning at par with the T5 variants across both metrics, even on significantly smaller datasets. It is evident that the performance of NLLB shows a notable improvement as the size of the training dataset increases. This observed trend could be attributed to the model's remarkable capacity to grasp context across a wide spectrum of languages and domains. Such adaptability potentially enhances its ability to comprehend context and patterns more effectively when adapting to new linguistic contexts. We intend to delve deeper into this model, as it exhibits proficiency in generating semantically enriched translations.

Furthermore, our comparative analysis extends to established models such as pretrained Google NMT, GPT3.5, and GPT4. Notably, our model mT5 surpasses the performance of GPT4 by **2.2%** in terms of semantic relevance and T5 outperformed GPT4 by **38.68%** in terms of BLEU score . This discrepancy may be attributed to their lack of explicit training on codemix text, rendering them less adept at generating coherent and contextually appropriate sentences in this domain. Our findings underscore the potential of our models and the dataset in addressing the unique challenges posed by codemix language translations. Table3 presents the performances of our models as well as the baseline LLMs on a PACMAN$^{unobserved}_{trans}$ sample. This table includes information on the generated translations, their corresponding BLEU scores,

Hinglish Sentence: Concerned rajya Electricity Boards ko ek specific form mein formal aavedan karnaa hoga.			
Ground Truth (English Sentence): A formal application needs to be made in a specific form to the concerned State Electricity Boards.			
Model	**Generated text**	**BLEU-1**	**BERT$_{\text{SemRel}}$**
T5$_1$	rajya Electricity Boards. have to make formal avedan in a specific form.	38.46	0.81
T5$_2$	A formal application in a specific form has to be made by the concerned State Electricity Boards.	**88.24**	0.99
NLLB$_1$	The concerned State Electricity Boards. have to make a formal application in a specific form.	58.34	0.95
NLLB2$_2$	The concerned State Electricity Boards have to submit a formal application in a specific form.	52.51	0.96
NLLB$_3$	The formal application in a specific form has to be made by the concerned State Electricity Boards.	**82.35**	0.99
mT5	A formal application in a specific form has to be made by the concerned State Electricity Boards.	**88.24**	0.99
Google NMT	Concerned State Electricity Boards will have to make a formal application in a specific form.	46.68	0.96
GPT3.5	Concerned State Electricity Boards will need to submit a formal application in a specific form.	46.68	0.97
GPT4	The concerned State Electricity Boards will have to make a formal application in a specific form.	52.84	0.97

Table 3: A sample comparison of translations generated by our domain-agnostic models v/s state-of-the-art LLMs.

Variants	**Models**	**BLEU-1 Score**	**BERT$_{\text{SemRel}}$**
Cross-Domain models (Trained on PACMAN$_{\text{trans}}$ only)	T5	25.33	0.85
	mT5	**27.87**	**0.87**
	NLLB	24.3	0.86
Custom Dual Fine-Tuned Models	**mT5**	**36.64**	**0.9**
	NLLB	27.47	0.88
Baseline Models	PHINCS	15.3	-
	mBART	25.3	-
	mT5	29.5	-

Table 4: Performance of our *dual fine-tuned* custom translation model on PHINC dataset over baselines PHINCS(Srivastava and Singh, 2020) mBART and mT5(Agarwal et al., 2021). Our mT5 dual fine-tuned model surpasses the existing benchmark by **22%**.

and BERT$_{\text{SemRel}}$ *w.r.to* the ground truth sentences.

Considering the commendable performance of our models, we proceeded to conduct a comprehensive evaluation of both the models and the dataset on the benchmark dataset, as presented in the following section.

5.2 Dual Fine-tuned Models for benchmarking on PHINC

In light of our models' robust performance on the PACMAN$^{unobserved}_{trans}$ dataset, we extended our evaluation to the well-established PHINC benchmark dataset, the results of which are presented in Table 4. Our initial assessments were carried out using models that had undergone explicit fine-tuning on PACMAN$_{\text{trans}}$ data. Remarkably, our models T5, mT5 and NLLB with **25.33**, **27.87** and **24.3** BLEU scores surpassed one of the PHINC baseline models in performance.

Further exploration involved fine-tuning our PACMAN-trained models on the PHINC training set (dual fine-tuning), comprising 13K samples. In this phase, mT5 emerged as the top performer, achieving a BLEU score of **0.36** and a BERT$_{\text{SemRel}}$ score of **0.90**. These results highlight the exceptional adaptability of PACMAN$_{\text{trans}}$ data. The dataset's consistency, error-free nature, and readability have proven invaluable in imparting the knowledge required for generating diverse sentence types, particularly for Hinglish to English translation.

In Table 4, our study includes baseline models, namely, PHINC(Srivastava and Singh, 2020), mBART and mT5(Agarwal et al., 2021). It is noteworthy that these models underwent fine-tuning using specific datasets. Google Translate was fine tuned on PHINC dataset, while both mBART and mT5 were intially fine-tuned on the dataset proposed by (Zhou et al., 2018) consisting of roughly 10K English and Hinglish codemixed sentences followed by fine-tuning on PHINC training dataset. These tailored fine-tuning approaches were applied to enhance the performance of these models in the context of codemix language.

Our analysis illuminates the remarkable efficacy of transformer-based models imbued with domain-specific linguistic knowledge. Specifically, when applied to the challenging task of translating code-mixed sentences in pretrained languages, these models demonstrate a marked superiority.

Our investigation underscores a fundamental truth: The performance of these models is intrin-

sically tied to the nature of the datasets on which they were meticulously trained. This relationship is vividly exemplified in Table 4, where our state-of-the-art (SOTA) model and the baseline model exhibit striking disparities in their BLEU Scores and semantic relevance. These discrepancies are primarily attributed to the contrasting foundational training datasets, accentuating the pivotal role of data quality and consistency. Our curated dataset for this task excels in these dimensions, boasting grammatical correctness and a unique capacity to facilitate optimal model learning, resulting in superior codemix to English translation.

6 Conclusion

In this work, we have presented a comprehensive approach to the challenge of translating codemixed text from Hinglish to English. Our contributions and achievements are summarized as follows:

- We introduced a novel dataset for Hinglish to English translation, which is more representative of the linguistic nuances found in codemixed social media text. This dataset is expected to serve as a valuable resource for future research in codemixed language processing.

- Our fine-tuned transformer-based models have demonstrated a marked improvement over the existing benchmarks, specifically outperforming the current state-of-the-art on the PHINC dataset. This highlights the effectiveness of our model architecture and training methodologies.

- Our findings also indicate that the adaptability of transformer models to the codemixing phenomenon can be further improved through targeted data augmentation strategies, suggesting a new direction for future research.

- Notably, our investigation reveals a pivotal strategy: Retraining a model initially trained on generic data with a domain-specific dataset. This approach yields outcomes that are not merely improved, but demonstrably superior.

- Lastly, our work contributes to the understanding of codemixing in machine translation, setting the stage for more linguistically-informed approaches to multilingual NLP.

In conclusion, the advancements presented in this paper not only push the boundaries of machine translation for codemixed languages but also provide insights that could inform the development of more robust NLP systems capable of handling the fluid dynamics of human language. We believe that our contributions will pave the way for more nuanced and effective translation systems, fostering better communication across language barriers.

7 Future Work

The promising results obtained from our current research lay a solid foundation for several exciting avenues of future work. We plan to extend our efforts in the following directions:

- **Expansion of the Hinglish Dataset:** To further enhance the robustness and applicability of our translation models, we aim to expand our novel dataset to capture a wider spectrum of the Hinglish language, including regional variations and different genres of communication.

- **Inclusion of Other Codemixed Language Pairs:** Building on the success of our current transformer-based models, we plan to explore their application to other codemixed language pairs such as Spanish-English, Marathi-English, Telugu-English *etc.*. By adapting our approach to these new language pairs, we hope to address the under representation of such languages in machine translation research.

- **Cross-Lingual Transfer Learning:** Investigating cross-lingual transfer learning techniques is another area we are keen to explore. By leveraging models trained on high-resource language pairs, we aim to improve translation quality for low-resource codemixed languages, which often suffer from data scarcity.

By addressing these goals, we hope not only to refine the translation mechanisms for Hinglish but also to extend our methodologies to a broader array of codemixed languages, promoting inclusivity and linguistic diversity in the NLP community.

References

Vibhav Agarwal, Pooja Rao, and Dinesh Jayagopi. 2021. Hinglish to english machine translation using multilingual transformers. pages 16–21.

Gustavo Aguilar, Fahad AlGhamdi, Victor Soto, Mona Diab, Julia Hirschberg, and Thamar Solorio. 2018. Named entity recognition on code-switched data: Overview of the calcs 2018 shared task. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 138–147.

Gustavo Aguilar, Sudipta Kar, and Thamar Solorio. 2020. Lince: A centralized benchmark for linguistic code-switching evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Kalika Bali, Jatin Sharma, Monojit Choudhury, and Yogarshi Vyas. 2014. Borrowing in hindi-english bilingual corpora. In *Proceedings of the Workshop on Code-Switching at the 52nd Annual Meeting of the Association for Computational Linguistics*.

Tej K. Bhatia and William C. Ritchie. 2013. *Hinglish: Code-mixing in Indian English*. Cambridge University Press.

Arindam Chatterjee, Chhavi Sharma, Ayush Raj, and Asif Ekbal. 2022. PACMAN:PArallel CodeMixed dAta generatioN for POS tagging. In *Proceedings of the 19th International Conference on Natural Language Processing (ICON)*, pages 234–244, New Delhi, India. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Marta R Costa-jussà, Hendrik Baan, Yulia Tsvetkov, Isaac Caswell, Sebastian Ruder Paulius de la Fuente, Sami Virpioja, Pavel Tsvetkov, and Radu Soricut. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019a. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1:4171–4186.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Abhisek Dhar et al. 2018. Handling code-mixed data using machine translation. *Language in India*, 18(4).

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *NAACL*.

Naman Goyal, Vishrav Goyal, Graham Neubig, Florian Metze, Michael Seltzer, Lori Levin, and Alan W Black. 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6102–6122. Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Aravind K. Joshi. 1982. Processing of sentences with intra-sentential code-switching. In *Proceedings of the 9th Conference on Computational Linguistics - Volume 1*, COLING '82, page 145–150, CZE. Academia Praha.

Simran Khanuja, Sandipan Dandapat, Kaushal Sankaranarayanan, and Kalika Bali. 2020. A new dataset and method for automatically grading esol texts. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2795–2806.

Carol Myers-Scotton. 1995. Social motivations for code-switching: Evidence from africa. *Clarendon Press/Oxford University Press*.

OpenAI. 2023. Gpt-4 technical report.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation.

Adithya Pratapa, Gayatri Bhat, Monojit Choudhury, Sunayana Sitaram, Sandipan Dandapat, and Kalika Bali. 2018. Language modeling for code-mixing: The role of linguistic theory based synthetic data. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1543–1553, Melbourne, Australia. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. *CoRR*, abs/2003.07082.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Deepak Kumar, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2021. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages. *CoRR*, abs/2104.05596.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 Indic languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

Shruti Rijhwani, Royal Sequiera, Monojit Choudhury, Kalika Bali, and Thamar Solorio. 2020. Estimating code-switching on twitter with a novel generalized word-level language detection technique. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Anoop Kumar Singh. 2018. Challenges in building neural machine translation systems for code-mixed language. In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*.

Anoop Kumar Singh and Manish Shrivastava. 2018. A twitter corpus for hindi-english code mixed pos tagging. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*.

Thamar Solorio and Yang Liu. 2008. Learning to translate mixed-language text. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*.

Mayank Srivastava et al. 2020. Phinc: A parallel hinglish social media code-mixed corpus for machine translation. *Language Resources and Evaluation*.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019a. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32. NeurIPS.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019b. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019*.

Genta Indra Winata, Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. 2019. Hierarchical attention network for code-switching named entity recognition. *Proceedings of the 3rd Workshop on Computational Approaches to Linguistic Code-Switching*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. 2018. A dataset for document grounded conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 708–713, Brussels, Belgium. Association for Computational Linguistics.