Hinglish to English Machine Translation using Multilingual Transformers

Vibhav Agarwal, Pooja Rao S B*, Dinesh Babu Jayagopi
International Institute of Information Technology Bangalore, India
* University of Lausanne, Switzerland
{vibhav.agarwal,jdinesh}@iiitb.ac.in, pooja.rao@unil.ch

Abstract

Code-Mixed language plays a very important role in communication in multilingual societies and with the recent increase in internet users especially in multilingual societies, the usage of such mixed language has also increased. However, the cross translation between the Hinglish Code-Mixed and English and vice-versa has not been explored very extensively. With the recent success of large pretrained language models, we explore the possibility of using multilingual pretrained transformers like mBART and mT5 for exploring one such task of code-mixed Hinglish to English machine translation. Further, we compare our approach with the only baseline over the PHINC dataset and report a significant jump from 15.3 to 29.5 in BLEU scores, a 92.8% improvement over the same dataset.

1 Introduction

Code-Mixing is the interleaving of tokens from different languages but in the same conversation. Code-Mixing is a common phenomenon in multilingual societies like India, China, Mexico, and in the last decade itself, there has been a massive surge of internet users and specifically from multilingual societies due to the popularity of various social media and messaging platforms. This has lead to a massive surge in mixed language data in the form of comments, conversations, etc. Unfortunately due to the informal nature of the code-mixed, it is hard to set a uniformly defined structure. However, linguists have formulated various hypotheses (Belazi et al., 1994; Pfaff, 1979; Poplack, 1981) and constraints (Sankoff and Poplack, 1981; Sciullo et al., 1986; Joshi, 1982) that can define a general rule for code-mixing.

Recent advances in attention-based mechanisms (Bahdanau et al., 2015) and transformers (Vaswani et al., 2017) have again shown significant performance improvement and shifted the communities' approach and interest in training larger neural models with deeper architecture. With the rise in large pretrained language models like (Devlin et al., 2019; Radford et al., 2019), there's been a lot of improvement in natural language processing problems. Prior work done in code-mixing like that of (Khanuja et al., 2020; Gupta et al., 2020) show the effectiveness of large multilingual pretrained language models like mBERT (Devlin et al., 2019) and XLM (Conneau and Lample, 2019) on code-mixed data. While GLUECoS attempts to set GLUE benchmark using finetuned mBERT, Gupta et al. (2020) shows the effectiveness of machine translation from a parallel corpus of English and Hindi to code-mixed sentence using XLM and Pointer-Generator model. Srivastava and Singh (2020); Dhar et al. (2018) proposes new task of Hinglish code-mixed to English translation and Srivastava and Singh (2020) collects a new social media code-mixed dataset called PHINC consisting of Hinglish code-mixed and parallel English sentences.

Our work attempts to utilize large seq2seq pretrained multilingual transformer based models like mT5 and mBART for the Hinglish to English machine translation task. We propose a dual curriculum learning method where first the models are trained for an English to code-mixed translation task and then finetuned again for a code-mixed to English task. There has been very little prior work involving this task and most of the work dealt with the English to code-mixed translation task for synthetic data generation. Our translation from code-mixed to English shows a significant improvement over the PHINC dataset and even beats the baseline presented by the PHINC dataset by a margin of 92.8%.

The rest of the paper is as follows - Section 2 talks about the prior work done in code-mixed machine translation and large multilingual pretrained

Proceedings of the Student Research Workshop associated with RANLP-2021, pages 16–21,
held online, Sep 1–3, 2021.
https://doi.org/10.26615/issn.2603-2821.2021_003

transformers. Section 3 discusses our model and the dataset used for the Hinglish to English translation. Section 4 addresses our experiments and our qualitative results for our approach. Finally, Section 5 addresses the conclusive remarks for this paper and presents a direction for future work.

2 Related Work

Code-Mixing refers to the interleaving of words belonging to different languages. This happens predominately in multilingual societies and is increasing rapidly with the increase in internet users on social media and messaging platforms. This has lead to a rapid increase in research interest in recent years and several tasks have been conducted as part of Code-Switching workshops (Diab et al., 2014, 2016). Most of the work in these workshops was single NLP problem-specific which were solved using specifically tailored models like Language Identification (Solorio et al., 2014; Molina et al., 2016), Named Entity Recognition (Rao and Devi, 2016; Aguilar et al., 2018), Question Answering (Chandu et al., 2018), Parts-of-Speech tagging (Jamatia et al., 2018), and Information Retrieval (banerjee et al., 2016). This was changed by Khanuja et al. (2020) which introduced a common benchmark for all the tasks using a single finetuned mBERT (Devlin et al., 2019) model downstreamed for all the benchmark tasks.

2.1 Code-Mixed Machine Translation

Machine Translation on code-mixed language is a relatively less explored area. There are only a few studies on English-Hinglish code-mixed language including the work of Dhar et al. (2018); Gupta et al. (2020); Srivastava and Singh (2020) despite a large populous of code-mixed speakers in South Asian countries. Dhar et al. (2018) collects a dataset of 6,096 Hinglish-English bitexts and propose a pipeline where they identify the languages involved in the code-mixed sentence, compute the matrix language and then translate the resulting sentence into the target language. Srivastava and Singh (2020) collects a large parallel corpus called PHINC, consisting of 13,738 Hinglish-English bitexts that they claimed was better than those of Dhar et al. (2018) in terms of diversity, quality, and generality. They propose a pipeline where they selectively identify token languages and then translate Hindi phrases to English using a monolingual translation system while keeping the rest of the phrases intact. This is the only work that addresses a Hinglish to English machine translation task. Gupta et al. (2020) propose a code-mixed text generator built upon the encoder-decoder framework, where the linguistic features obtained from a transformer based language model are encoded using the encoder. They proposed using features from a pretrained cross-lingual transformer based model XLM (Conneau and Lample, 2019) along with Pointer-Generator (See et al., 2017) model as its decoder for the code-mixed text generation.

2.2 Multilingual Pretrained Models

Transformer-based neural models have increasingly become a go-to solution for any NLP problem using models trained with a self-supervised objective like BERT (Devlin et al., 2019), BART (Lewis et al., 2020), and T5 (Raffel et al., 2020). In cross-lingual or multilingual domains, there has been a rapid increase in the number of models built upon BERT, BART, or T5 architecture or they use a similar architecture. Works like mBERT, mBART (Liu et al., 2020), mT5 (Xue et al., 2021), XLM (Conneau and Lample, 2019) and XLM-R (Conneau et al., 2020) are a few examples of such models. While mBERT and XLM-R are encoder only architectures that could be used for downstream classification tasks, XLM is a seq2seq encoder-decoder task-specific model, built mostly for translation tasks. Models like mBART and mT5 are seq2seq encoder-decoder architecture that can solve multiple downstream tasks like summarization, translation, or any other language generational task without any additional modeling head. These models are trained on multiple languages at once with a self-supervised objective like span corruption, permutation, etc. While mBART is pretrained on 25 languages on the same BART objective, mT5 is pretrained on 101 languages on the T5 model's objective. None of these multilingual models are trained on any code-mixed language or at least are not aware of their pretraining data consisting of code-mixed data. Therefore, it becomes an important question to verify and validate these model performances on code-mixed languages.

3 System Overview

In this Section, we propose our mBART and mT5 based machine translation model and the dataset used for finetuning the same.

Hinglish Code-Mixed to English Translation		
Mujhe lagta hai wo captured humanoid amphibian creature play karta hai	→	I think he plays a captured humanoid amphibian creature.
main character ek orphan hai jo ek river mein paya gaya	→	The main character is an orphan who is found in a river.
kya aapko lagtha hein ki ache movie ko high-profile actors hi chahiye?	→	Do you think a good movie should have high-profile actors?
Main suprised hu ye itna low hai	→	I'm suprised how low it is.

Table 1: Samples from the Hinglish to English translation task. Red tokens refer to the Hindi tokens in the Roman script.

3.1 Machine Translation model

We use mBART and mT5 models finetuned on Hinglish code-mixed to English data as described in Section 3.2. mBART is a multilingual seq2seq denoising bidirectional auto-encoder pretrained using the same BART (Lewis et al., 2020) objective but on large-scale monolingual corpora of 25 languages. It is based on the same transformer (Vaswani et al., 2017) architecture and consists of 12 encoder and decoder layers each with 16 attention heads and model dimensions being 1024 resulting in roughly 680 million parameters. mT5 is the multilingual variant of the T5 model pretrained on 101 languages. It has a similar transformer architecture with 2 encoder and decoder layers each, model dimensions being 1024 and 12 attention heads resulting in approximately 770 million parameters.

3.2 Dataset

We use the following datasets to finetune and test our models for Hinglish code-mixed to English translation task:

- **CMU Hinglish** is an extended code-mixed form of the Document Grounded Conversation (Zhou et al., 2018) dataset. It consists of roughly 10,000 English and Hinglish code-mixed sentences.
- **PHINC** (Srivastava and Singh, 2020) consists of Hinglish code-mixed to English translation pairs. It contains roughly 13,000 parallel pairs.

For both of the datasets, we transliterate the Hindi Devanagari script into its Roman script form using CSNLI (Bhat et al., 2017, 2018) and Microsoft Translator.

4 Experiments & Results

In this section, we describe our experimental setup and our finetuning results for our models described in Section 3.

4.1 Experimental Setup

Our proposed approach is written in Pytorch (Paszke et al., 2019) and all the transformer based models and their associated weights are from the HuggingFace's Transformer (Wolf et al., 2020) package. We use *mbart-cc-25* model weights for mBART and *mt5-base* model weight for mT5 in all our modeling. Both the mBART and mT5 models were trained using the AdamW optimizer with weight decay. We used all the default hyperparameters except the number of training epochs, mBART was trained for 5 epochs while the mT5 took larger epochs (50) to converge the training.

We train both of our models in a dual curriculum learning method where we first finetune it on a English to code-mixed data so that model identifies what code-mixed language looks like and then we finetune it again on the Hinglish code-mixed gold dataset for our final task. We show some of the example translations of our best performing model in Table 1. We show the performance of both mT5 and mBART models on the datasets described in Section 3.2 in Table 2. For the BLEU evaluation, we use the *sacrebleu* metric from HuggingFace's Dataset package.

Baseline: Our only comparative baseline is defined by Srivastava and Singh (2020) where they collect a code-mixed dataset from social media and conversations called PHINC and propose a machine translation pipeline built on top of Google Translate with a BLEU score of 15.3.

Datasets	BLEU score
Original Baseline	
PHINC	15.3
mBART model (ours)	
CMU Hinglish → Reverse CMU Hinglish	24.0
CMU Hinglish → PHINC	25.3
mT5 model (ours)	
CMU Hinglish → Reverse CMU Hinglish	28.6
CMU Hinglish → PHINC	**29.5**

Table 2: BLEU score for code-mixed Hinglish to English Translation

4.2 Results

As shown in Table 2, we first finetune our mBART and mT5 models on the CMU Hinglish dataset. These finetuned models have a BLEU score of 11.53 and 11.23 respectively. These are then further finetuned for our final task of Hinglish to English translation. We perform and analyse the second finetune on both the flipped CMU Hinglish dataset and the PHINC dataset.

All four variations of our dual finetuned mBART and mT5 models outperform the original baseline. While the best performing mBART model improves over the baseline by 65.3%, mT5 model beats the PHINC baseline by a margin of 92.8%. This shows the prowess of the large pretrained multilingual transformers in code-mixed translation tasks.

5 Conclusion

Our work proposes using large multilingual transformers (mBART and mT5) and demonstrates how finetuning them in a dual curriculum learning method improves the performance of code-mixed Hinglish to English machine translation tasks. We show a very significant improvement over the PHINC baseline by a margin of 92.8% over the BLEU scores.

As part of the future work, we would like to further improve our machine translation model by using a large amount of synthetic code-mixed data to improve our English to code-mixed translation model, which would further improve our performance on the code-mixed to English translation task. We would also like to extend this work to other low resource and code-mixed languages.

References

Gustavo Aguilar, Fahad AlGhamdi, Victor Soto, Thamar Solorio, Mona Diab, and Julia Hirschberg, editors. 2018. *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*. Association for Computational Linguistics, Melbourne, Australia.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Somnath banerjee, Kunal Chakma, Sudip Kumar Naskar, Amitava Das, Paolo Rosso, Sivaji Bandyopadhyay, and Monojit Choudhury. 2016. Overview of the Mixed Script Information Retrieval (MSIR). In *Proceedings of FIRE 2016*. FIRE.

Hedi M. Belazi, Edward J. Rubin, and Almeida Jacqueline Toribio. 1994. Code switching and x-bar theory: The functional head constraint. *Linguistic Inquiry*, 25(2):221–237.

Irshad Bhat, Riyaz A. Bhat, Manish Shrivastava, and Dipti Sharma. 2017. Joining hands: Exploiting monolingual treebanks for parsing of code-mixing data. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 324–330, Valencia, Spain. Association for Computational Linguistics.

Irshad Bhat, Riyaz A. Bhat, Manish Shrivastava, and Dipti Sharma. 2018. Universal Dependency parsing for Hindi-English code-switching. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics:*

Human Language Technologies, Volume 1 (Long Papers), pages 987–998, New Orleans, Louisiana. Association for Computational Linguistics.

Khyathi Chandu, Ekaterina Loginova, Vishal Gupta, Josef van Genabith, Günter Neumann, Manoj Chinnakotla, Eric Nyberg, and Alan W. Black. 2018. Code-mixed question answering challenge: Crowdsourcing data and techniques. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 29–38, Melbourne, Australia. 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. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 7057–7067.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Mona Diab, Pascale Fung, Mahmoud Ghoneim, Julia Hirschberg, and Thamar Solorio, editors. 2016. *Proceedings of the Second Workshop on Computational Approaches to Code Switching*. Association for Computational Linguistics, Austin, Texas.

Mona Diab, Julia Hirschberg, Pascale Fung, and Thamar Solorio, editors. 2014. *Proceedings of the First Workshop on Computational Approaches to Code Switching*. Association for Computational Linguistics, Doha, Qatar.

Deepak Gupta, Asif Ekbal, and Pushpak Bhattacharyya. 2020. A semi-supervised approach to generate the code-mixed text using pre-trained encoder and transfer learning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2267–2280, Online. Association for Computational Linguistics.

Anupam Jamatia, Björn Gambäck, and Amitava Das. 2018. Collecting and Annotating Indian Social Media Code-Mixed Corpora. In *Computational Linguistics and Intelligent Text Processing*, pages 406–417, Cham. Springer International Publishing.

Aravind K. Joshi. 1982. Processing of sentences with intra-sentential code-switching. In *Coling 1982: Proceedings of the Ninth International Conference on Computational Linguistics*.

Simran Khanuja, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, and Monojit Choudhury. 2020. GLUECoS: An evaluation benchmark for code-switched NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3575–3585, Online. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Giovanni Molina, Fahad AlGhamdi, Mahmoud Ghoneim, Abdelati Hawwari, Nicolas Rey-Villamizar, Mona Diab, and Thamar Solorio. 2016. Overview for the second shared task on language identification in code-switched data. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 40–49, Austin, Texas. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 8024–8035.

Carol W. Pfaff. 1979. Constraints on language mixing: Intrasentential code-switching and borrowing in spanish/english. *Language*, 55(2):291–318.

Shana Poplack. 1981. *Syntactic structure and social function of code-switching*, pages 169–184.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

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.

Pattabhi R. K. Rao and S. Devi. 2016. Cmee-il: Code mix entity extraction in indian languages from social media text @ fire 2016 - an overview. In *FIRE*.

David Sankoff and Shana Poplack. 1981. A formal grammar for code-switching. *Papers in Linguistics - International Journal of Human Communication*, 14:3–46.

Anne-Marie Di Sciullo, Pieter Muysken, and Rajendra Singh. 1986. Government and code-mixing. *Journal of Linguistics*, 22(1):1–24.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Ghoneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang, and Pascale Fung. 2014. 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*, pages 62–72, Doha, Qatar. Association for Computational Linguistics.

Vivek Srivastava and Mayank Singh. 2020. PHINC: A parallel Hinglish social media code-mixed corpus for machine translation. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 41–49, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

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.