CNLP-NITS-PP at MixMT 2022: Hinglish–English Code-Mixed Machine Translation

Sahinur Rahman Laskar[1], **Rahul Singh**[1], **Shyambabu Pandey**[1], **Riyanka Manna**[2]
Partha Pakray[1], **Sivaji Bandyopadhyay**[1]
[1]Department of Computer Science and Engineering, National Institute of Technology, Silchar, India
[2]Department of Computer Science and Engineering, Adamas University, Kolkata, India
{sahinurlaskar.nits, rahuljan, babushyampandey, riyankamanna16}@gmail.com
{parthapakray,sivaji.cse.ju}@gmail.com

Abstract

The mixing of two or more languages in speech or text is known as code-mixing. In this form of communication, users mix words and phrases from multiple languages. Code-mixing is very common in the context of Indian languages due to the presence of multilingual societies. The probability of the existence of code-mixed sentences in almost all Indian languages since in India English is the dominant language for social media textual communication platforms. We have participated in the WMT22 shared task of code-mixed machine translation with the team name: CNLP-NITS-PP. In this task, we have prepared a synthetic Hinglish–English parallel corpus using transliteration of original Hindi sentences to tackle the limitation of the parallel corpus, where, we mainly considered sentences that have named-entity (proper noun) from the available English-Hindi parallel corpus. With the addition of synthetic bi-text data to the original parallel corpus (train set), our transformer-based neural machine translation models have attained recall-oriented understudy for gisting evaluation (ROUGE-L) scores of 0.23815, 0.33729, and word error rate (WER) scores of 0.95458, 0.88451 at Sub-Task-1 (English-to-Hinglish) and Sub-Task-2 (Hinglish-to-English) for test set results respectively.

1 Introduction

The mixing of alternating words from two different language vocabulary without misinterpreting the context of the sentence is known as code-switching or code-mixing (Poulisse, 1998). This style of communication is one of the most frequent in multilingual communities, such as India. English is extensively mixed with local languages, such as Hindi, and Bengali, which causes code-mixed English-Hindi: Hinglish and English-Bengali: Binglish languages (Sailaja, 2011). Code-mixing is not observed in formal literature such as books but is commonly used on social media platforms such as Facebook and Twitter. The WMT22 organizes shared task code-mixed machine translation for English-to-Hinglish and Hinglish-to-English, where the main challenge is low-resource availability of parallel corpus. We have participated in the same task and to mitigate the issue of data scarcity, a synthetic Hinglish-English parallel corpus is prepared (as discussed in Section 3.1). In this work, the transformer-based neural machine translation (NMT) technique (Vaswani et al., 2017; Laskar et al., 2022) is utilized to build NMT models for both directions (English-to-Hinglish, Hinglish-to-English) of code-mixed MT.

2 Related Work

In recent times, many significant NLP studies have included the study of code-mixed languages. The EMNLP 2022 seventh conference on machine translation (WMT22) has put forward several tasks directed to meet new challenges in the field of NLP for code-mixed Indian languages. The competition has attracted many researchers to follow up with these tasks, which have eventually led to new directions and problems in this domain. The task of machine translation for code-mixed languages has not been an active area of research due to the scarcity of manually annotated datasets. Recently, researchers have been developing datasets for code-mixed MT that includes Hinglish-English parallel corpus, namely, HinGe (Srivastava and Singh, 2021) and PHINC (Srivastava and Singh, 2020) to overcome the datasets scarcity issue to build code-mix MT that is associated with the code-mixed text from various social media platforms. In this work, we addressed the issue of data scarcity by using synthetic Hinglish–English parallel corpus to increase the training data for code-mixed MT shared tasks at WMT22.

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 1158 - 1161
December 7-8, 2022 ©2022 Association for Computational Linguistics

3 System Description

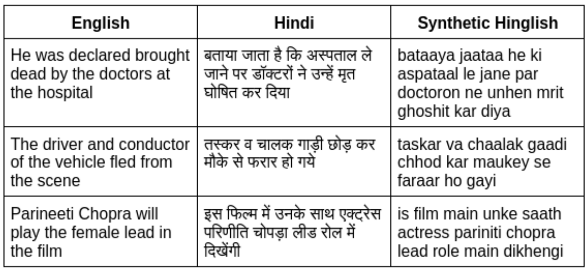
The experiments are carried out in four phases, namely, synthetic data preparation and augmentation to the train set, data preprocessing, model training, and testing. The OpenNMT-py (Klein et al., 2017) tool is utilized to build the NMT models independently for English-to-Hinglish (subtask-1) and Hinglish-to-English (subtask-2).

3.1 Dataset Description

We have used the dataset provided by the WMT22 organizer[1] and the statistics are presented in Table 1. Moreover, the synthetic English-Hinglish parallel dataset is prepared and directly augmented with the train set to expand the training amount of data. For synthetic data preparation, the English-Hindi parallel sentences are collected from Samanantar dataset (Ramesh et al., 2022) and selected $100k$ sentences (maximum length of 15 words). To select parallel sentences, the following steps are considered:

- Step-1: Extract proper nouns (named-entity) from the English side using NLTK[2] toolkit.
- Step-2: Extract English sentences that have extracted proper nouns in Step-1.
- Step-3: Select corresponding Hindi sentences of English that are extracted in Step-2.

Then, Hindi side sentences are transliterated into English script using Indic-trans[3] (Bhat et al., 2014) and prepared synthetic Hinglish sentences. Thus, we have prepared $100k$ Hinglish–English synthetic parallel corpus. The sample sentences of synthetic Hinglish-English are presented in Figure 1. The data statistics of the train set, before and after augmentation of synthetic Hinglish–English corpus is presented in Table 2.



English	Hindi	Synthetic Hinglish
He was declared brought dead by the doctors at the hospital	बताया जाता है कि अस्पताल ले जाने पर डॉक्टरों ने उन्हें मृत घोषित कर दिया	bataaya jaataa he ki aspataal le jane par doctoron ne unhen mrit ghoshit kar diya
The driver and conductor of the vehicle fled from the scene	तस्कर व चालक गाड़ी छोड़ कर मौके से फरार हो गये	taskar va chaalak gaadi chhod kar maukey se faraar ho gayi
Parineeti Chopra will play the female lead in the film	इस फिल्म में उनके साथ एक्ट्रेस परिणीति चोपड़ा लीड रोल में दिखेंगी	is film main unke saath actress pariniti chopra lead role main dikhengi

Figure 1: Sample sentences of synthetic Hinglish-English.

3.2 Experimental Setup

We have performed byte pair encoding jointly (subword level) (Sennrich et al., 2016) on the Hinglish-English with $32k$ merge operations. The subword level source-target vocabulary is shared during the training process of the NMT model. The OpenNMT-py toolkit has been used for text data tokenization, preprocessing, and conducting the NMT model training. We have followed the default settings of the 6 layer transformer model (Vaswani et al., 2017) in the training process. We have used a batch size of 32, 0.1 drop-outs, and an Adam optimizer with a 0.001 learning rate during the training process. The NMT model is trained on a single GPU with early stopping criteria, i.e., the model training is halted if it does not converge on the validation set for more than 10 epochs. The obtained trained model is used to translate the test data provided by the WMT22 organizers.

4 Results

The WMT22 shared task organizer published the evaluation result[4] of the code-mixed machine translation (MixMT) task for English–Hinglish language pair. We participated with the team name CNLP-NITS-PP in the monolingual to code-mixed machine translation: English-to-Hinglish (Sub-Task-1) and code-mixed to a monolingual machine translation: Hinglish-to-English (Sub-Task-2) submission tracks of the same task where ten teams participated. The automatic evaluation metrics, namely, ROUGE (Recall-Oriented Understudy for Gisting Evaluation) (Lin, 2004), WER (Word Error Rate) (Morris et al., 2004) and human evaluation (HE) are used for the evaluation of results. Table 3, 4 reported the official results of our systems in terms of automatic and HE evaluation metrics. We

[1] https://www.statmt.org/wmt22/code-mixed-translation-task.html

[2] https://github.com/nltk/nltk

[3] https://github.com/libindic/indic-trans

[4] https://codalab.lisn.upsaclay.fr/competitions/2861#results

Task	Data Set	No. of Sentences	Tokens	
			English	Hinglish
Sub-Task-1	Train Set	2766	47347	52074
	Validation Set	500	5847	5565
	Test Set	1500	17694	17049
Sub-Task-2	Train Set	13738	169158	176410
	Validation Set	500	5847	10263
	Test Set	1500	27659	29335

Table 1: Data Statistics of English-Hinglish (provided by the organizer).

Train Set	Number of Parallel Sentence/Segments
Before Augmentation	2766 (Sub-Task-1) 13738 (Sub-Task-2)
After Augmentation	102,766 (Sub-Task-1) 113,738 (Sub-Task-2)

Table 2: Data Statistics of train set (before and after augmentation).

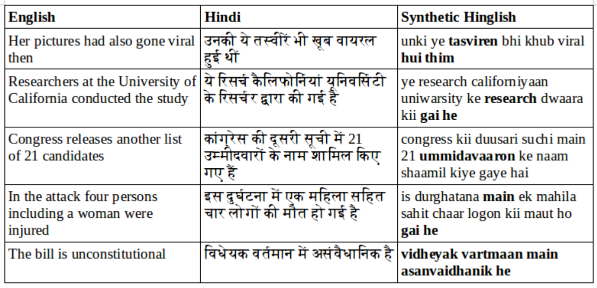
have attained better automatic evaluation scores and positions in Sub-Task-2 as compared to Sub-Task-1 for the validation and test set, whereas, in the case of human evaluation, we have achieved a higher score and position in Sub-Task-1 than Sub-Task-2. It is observed that due to the presence of a high amount of transliteration errors in synthetic code-mixed sentences, i.e., Hinglish, the predicted sentences suffer lower translation accuracy. A few examples of transliteration errors are presented in Figure 2.

Task	Set	ROUGE-L	WER
Sub-Task-1	Validation	0.23359 ($8th$)	0.97136 ($7th$)
	Test	0.23815 ($7th$)	0.95458 ($7th$)
Sub-Task-2	Validation	0.33835 ($4th$)	0.88002 ($3rd$)
	Test	0.33729 ($6th$)	0.88451 ($6th$)

Table 3: Our system's results (official) at MixMT shared task (WMT22).

Task	HE
Sub-Task-1	2.10 ($4th$)
Sub-Task-2	1.35 ($7th$)

Table 4: Our system's human evaluation results (official) at MixMT shared task (WMT22).



English	Hindi	Synthetic Hinglish
Her pictures had also gone viral then	उनकी ये तस्वीरें भी खूब वायरल हुई थीं	unki ye **tasviren** bhi khub viral **hui thim**
Researchers at the University of California conducted the study	ये रिसर्च कैलिफोर्नियां यूनिवर्सिटी के रिसर्चर द्वारा की गई है	ye research californiyaan uniwarsity ke **research** dwaara kii **gai he**
Congress releases another list of 21 candidates	कांग्रेस की दूसरी सूची में 21 उम्मीदवारों के नाम शामिल किए गए हैं	congress kii duusari suchi main 21 **ummidavaaron** ke naam shaamil kiye gaye hai
In the attack four persons including a woman were injured	इस दुर्घटना में एक महिला सहित चार लोगों की मौत हो गई है	is durghatana **main** ek mahila sahit chaar logon kii maut ho **gai he**
The bill is unconstitutional	विधेयक वर्तमान में असंवैधानिक है	**vidheyak vartmaan main asanvaidhanik he**

Figure 2: Sample examples of transliteration errors.

5 Conclusion and Future Work

In this work, we have investigated a transformer-based model for Hinglish–English language pair in the WMT22 code-mixed MT task. We have addressed the data scarcity issue by the augmentation of synthetic Hinglish–English parallel sentences to the train set for both English-to-Hinglish and Hinglish-to-English translation tasks (Sub-Task-1 and Sub-Task-2). Furthermore, synthetic parallel data will be corrected in the future to improve translational performance.

Acknowledgements

The authors would like to thank the Center for Natural Language Processing (CNLP), Artificial Intelligence (AI), and the Department of Computer Science and Engineering at National Institute of Technology, Silchar for providing the requisite support and infrastructure to execute this work.

References

Irshad Ahmad Bhat, Vandan Mujadia, Aniruddha Tammewar, Riyaz Ahmad Bhat, and Manish Shrivastava. 2014. Iiit-h system submission for fire2014 shared task on transliterated search. In *Proceedings of the Forum for Information Retrieval Evaluation*, FIRE '14, page 48–53, New York, NY, USA. Association for Computing Machinery.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Sahinur Rahman Laskar, Abdullah Faiz Ur Rahman Khilji, Partha Pakray, and Sivaji Bandyopadhyay. 2022. Improved neural machine translation for low-resource english–assamese pair. *Journal of Intelligent and Fuzzy Systems*, 42(5):4727–4738.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Andrew Cameron Morris, Viktoria Maier, and Phil D. Green. 2004. From wer and ril to mer and wil: improved evaluation measures for connected speech recognition. In *INTERSPEECH*.

Nanda Poulisse. 1998. Duelling languages: Grammatical structure in codeswitching. *International Journal of Bilingualism*, 2(3):377–380.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Divyanshu Kakwani, Navneet Kumar, et al. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

Pingali Sailaja. 2011. Hinglish: code-switching in Indian English. *ELT Journal*, 65(4):473–480.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. 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@EMNLP 2020 Online, November 19, 2020*, pages 41–49. Association for Computational Linguistics.

Vivek Srivastava and Mayank Singh. 2021. Hinge: A dataset for generation and evaluation of code-mixed hinglish text. *CoRR*, abs/2107.03760.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.