PHINC: A Parallel Hinglish Social Media Code-Mixed Corpus for Machine Translation

Vivek Srivastava [1] Mayank Singh [2]

[1]TCS Research and Innovation Pune, India [2]IIT Gandhinagar Gujarat, India

Motivation

Hinglish sentence	Google Translate	Bing Translate	Correct translation
kitna achha din hai, I could not have asked for isse achha	What a great day, Mr. Kud Note has asked for better than this	How much acha is the day, e-coused not how many asked for it achha	What a great day, I could not have asked for better than this
par if its possible and any other guest needs a room , mera room de de kisi ko bhi	par if its possible and any other guest needs a room , mera room de de kisi ko bhi	On if its possible egg any other guest needs coming room , my room day to anyone	but if it is possible and any other guest needs a room , give my room to anyone

Table 1:Hinglish (code-mixed Hindi-English) to English translation. The red marked phrases are incorrectly translated in addition to other grammatical errors.

Challenges with Code-Mixed Machine Translation

Challenge	Example(s)
C1: Ambiguity in language identification	'*is*', '*me*', '*to*'
C2: Spelling variations	'*jaldi*', '*jldi*', 'jaldiii'
C3: Named entity recognition	'*Bhartiya Janta Party*'
C4: Informal style of writing	'*Sad kabhi dekha h usko.. me never*'
C5: Misplaced/skipped punctuation	'*Aap kb se cricket khelne lage..never saw u bfr*'
C6: Missing context	'*Note kr lijiye.. Bandi chal rahi h*'

Table 2:We identify six major reasons for the failure of machine translation systems on the code-mixed data.

Previous Work

	Previous Work (Dhar et al., 2018)	PHINC
English sentences	✔	✔
Spelling variations	✔	✔
Short sentences	✔	✔
Ambiguous sentences	✔	✔
Abusive sentences	✔	✔
Total sentences	6,096	13,738

Table 3:Comparison with the parallel corpus proposed in the previous work [2].

Dataset

Dataset Source	Task	Platform	Dataset Size	Topics/Focus areas
Singh et al.	Named-entity recognition	Twitter	3,638	Politics, social events, sports, etc.
Swami et al.	Sarcasm detection	Twitter	5,250	Bollywood, cricket, and politics
Joshi et al.	Sentiment analysis	Facebook	3,879	Bollywood and politics
Barman et al.	Language identification	Facebook	771	Not available
Shete et al.	Sentiment analysis	Facebook	7,663	Politics, news articles, etc.
Khandelwal	Humor detection	Twitter	31,033	Not available
PHINC	Machine translation	Twitter & Facebook	13,738	Sports, politics, Bollywood, etc.

Table 4:We use six source datasets [5, 6, 3, 1, 7, 4] to create PHINC. We preprocess these datasets to remove short, monolingual, and low code-mixed sentences. Annotators do not provide translations for English, abusive, or ambiguous sentences.

Translation Augmentation Pipeline

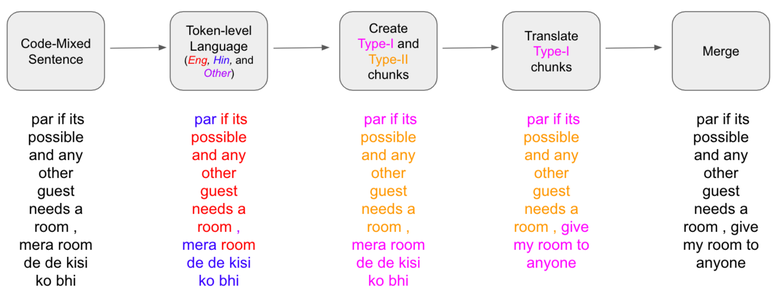

Figure 1:An example translation using the proposed pipeline build on top of Google Translate (PPGT).

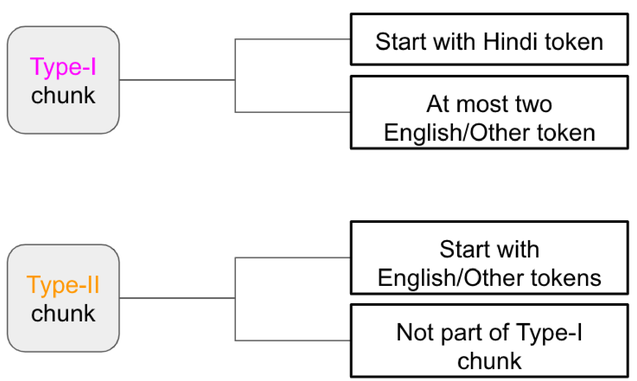

Figure 2:The criteria to create Type-I and Type-II chunks for the PPGT based code-mixed machine translation.

Results

CODE-MIXED SENTENCE: ab voh bola jisne kisi bhi party ko support karne se mana kardiya tha . . a flop show annaji
TYPE-I CHUNKS: [ab voh bola jisne kisi bhi party ko support karne se mana kardiya tha]
TYPE-II CHUNKS: [. . a flop show annaji]
ENGLISH TRANSLATION USING BT: Now Woh spoke , which was considered to support any party . . Come Flop Show Annaji
ENGLISH TRANSLATION USING GT: Now say that he had a desire to support any party. . A flop show Anna
ENGLISH TRANSLATION USING PPGT: Now speak that who had refused to support any party . . a flop show annaji

Figure 3:Example translation of the Hinglish sentence using Google Translate (GT), Bing Translate (BT), and PPGT.

	BLEU-1	WER	TER
BT	0.146	0.751	0.885
GT	0.151	0.600	0.718
PPGT	**0.153**	**0.566**	**0.685**

Table 5:Comparison of performance evaluation of various machine translation systems on the PHINC dataset.

References

[1] Utsab Barman, Amitava Das, Joachim Wagner, and Jennifer Foster.
Code mixing: A challenge for language identification in the language of social media.
In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 13--23, Doha, Qatar, October 2014. Association for Computational Linguistics.

[2] Mrinal Dhar, Vaibhav Kumar, and Manish Shrivastava.
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, 2018.

[3] Aditya Joshi, Ameya Prabhu, Manish Shrivastava, and Vasudeva Varma.
Towards sub-word level compositions for sentiment analysis of hindi-english code mixed text.
In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2482--2491, 2016.

[4] Ankush Khandelwal.
Humor detection corpus, 2018.
[Online; accessed 08-Jan-2020].

[5] Vinay Singh, Deepanshu Vijay, Syed Sarfaraz Akhtar, and Manish Shrivastava.
Named entity recognition for hindi-english code-mixed social media text.
In *Proceedings of the Seventh Named Entities Workshop*, pages 27--35, 2018.

[6] Sahil Swami, Ankush Khandelwal, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava.
A corpus of english-hindi code-mixed tweets for sarcasm detection.
arXiv preprint arXiv:1805.11869, 2018.

[7] GaganDeep Singh Chhabra Vrishank Shete and Lokesh Mittal.
Sentiment analysis on hindi-english code mixed data using svm, 2016.
[Online; accessed 08-Jan-2020].