Quality Evaluation of the Low-Resource Synthetically Generated Code-Mixed Hinglish Text

Vivek Srivastava
TCS Research
Pune, Maharashtra, India
srivastava.vivek2@tcs.com

Mayank Singh
IIT Gandhinagar
Gandhinagar, Gujarat, India
singh.mayank@iitgn.ac.in

Abstract

In this shared task, we seek the participating teams to investigate the factors influencing the quality of the code-mixed text generation systems. We synthetically generate code-mixed Hinglish sentences using two distinct approaches and employ human annotators to rate the generation quality. We propose two subtasks, *quality rating prediction* and *annotators' disagreement prediction* of the synthetic Hinglish dataset. The proposed subtasks will put forward the reasoning and explanation of the factors influencing the quality and human perception of the code-mixed text.

1 Introduction

Code-mixing is the phenomenon of mixing words and phrases from multiple languages in a single utterance of a text or speech. Figure 1 shows the example code-mixed Hinglish sentences generated from the corresponding parallel Hindi and English sentences. Code-mixed languages are prevalent amongst multilingual communities such as Spain, India, and China. With the inflation of social-media platforms in these communities, the availability of code-mixed data is seeking a boom. It has lead to several interesting research avenues for problems in computational linguistics such as language identification (Singh et al., 2018; Shekhar et al., 2020), machine translation (Dhar et al., 2018; Srivastava and Singh, 2020), language modeling (Pratapa et al., 2018), etc.

Over the years, we observe various computational linguistic conferences and workshops organizing the shared tasks involving the code-mixed languages. Diverse set of problems have been hosted such as sentiment analysis (Chakravarthi et al., 2021; Patwa et al., 2020), offensive language identification (Chakravarthi et al., 2021), word-level language identification (Solorio et al., 2014; Molina et al., 2016), information retrieval (Banerjee et al., 2016), etc.

Example I

HINGLISH: ye ek code mixed sentence ka example hai
ENGLISH : this is an example code-mixed sentence

Example II

HINGLISH : kal me movie dekhne ja raha hu. How are the reviews?
ENGLISH: I am going to watch the movie tomorrow. How are the reviews?

Figure 1: Example parallel Hinglish and English sentences. The code-mixed Hinglish sentences contain words from Hindi and English languages.

Despite these overwhelming attempts, the natural language generation (NLG) and evaluation of the code-mixed data remain understudied. The noisy and informal nature of the code-mixed text adds to the complexity of solving and evaluating the various NLG tasks such as summarization and machine translation. These inherent challenges (Srivastava and Singh, 2020) with the code-mixed data makes the widely popular evaluation metrics like BLEU and WER obsolete. Various metrics (e.g., CMI (Das and Gambäck, 2014; Gambäck and Das, 2016), M-index (Barnett et al., 2000), I-index (Guzmán et al., 2017), Burstiness (Goh and Barabási, 2008), Memory (Goh and Barabási, 2008), etc.) have been proposed to measure the complexity of code-mixed data, but they fail to capture the linguistic diversity which leads to poorly estimating the quality of code-mixed text (Srivastava and Singh, 2021a).

With this shared task[1] (see Section 2 and 4 for the detailed description), we look forward to the

[1] https://sites.google.com/view/hinglisheval

Proceedings of the 14th International Conference on Natural Language Generation (INLG), pages 314–319,
Aberdeen, Scotland, UK, 20-24 September 2021.

new strategies that cater to the broad requirement of the quality evaluation of the generated code-mixed text. These methods will entail various linguistic features encompassing syntax and semantics and the perspectives of human cognition such as writing style, emotion, sentiment, language, and preference. We also put forward a subtask to understand the factors influencing the human disagreement on the quality rating of the generated code-mixed text. This could help design a more robust quality evaluation system for the code-mixed data.

2 Task Overview

In this shared task, we propose two subtasks evaluating the quality of the code-mixed Hinglish text. First, we propose to predict the quality of Hinglish text on a scale of 1–10. We aim to identify the factors influencing the text's quality, which will help build high-quality code-mixed text generation systems. We synthetically generate the Hinglish sentences using two different approaches (see Section 3) leveraging popular English-Hindi parallel corpus. Besides, we also have at least two human-generated Hinglish sentences corresponding to each parallel sentence. The second subtask aims to predict the disagreement on a scale of 0–9 between the two annotators who have annotated the synthetically generated Hinglish sentences. Various factors influence this human disagreement, and we seek to investigate the reasoning behind this behavior.

3 Dataset

As outlined in Section 1, the code-mixed NLG task observes a scarcity of high-quality datasets. Consequently, the quality evaluation of the generated code-mixed text remains unexplored. We propose a new dataset with Hinglish sentences generated synthetically and rated by human annotators to address this challenge. We create the dataset in two phases.

Phase 1: Human-generated Hinglish sentences: We select the English-Hindi parallel sentences from the IIT-B parallel corpus(Kunchukuttan et al., 2018) to generate the Hinglish sentences. The parallel corpus has 1,561,840 sentence pairs. We randomly select 5,000 sentence pairs, in which the number of tokens in both the sentences is more than five. We employ five human annotators and assign each 1,000 sentence pairs. Table 1 shows the annotation guidelines to generate the Hinglish sentences. Post annotation, we obtain 1,976 sentence pairs for which the annotators have generated at least two Hinglish sentences.

Phase 2: Synthetic Hinglish sentence generation and quality evaluation: We synthetically generate the Hinglish sentence corresponding to each of the parallel 1,976 English-Hindi sentence pairs. We employ two different code-mixed text generation (CMTG) techniques:

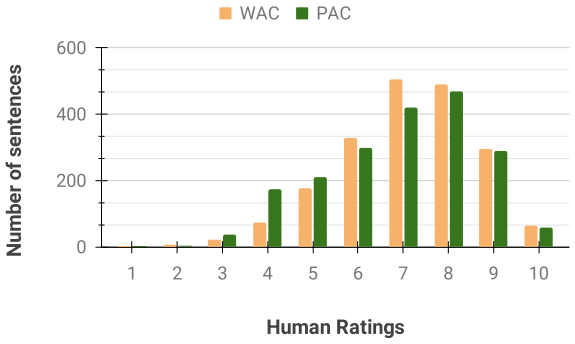
- Word-aligned CMTG (WAC): Here, we align the noun and adjective tokens between the parallel sentences. We replace the aligned Hindi token with the corresponding English token and transliterate the Hindi sentence to the Roman script.
- Phrase-aligned CMTG (PAC): Here, we align the key-phrases of length up to three tokens between the parallel sentences. We replace the aligned Hindi phrase with the corresponding English phrase and transliterate the Hindi sentence to the Roman script.

For the token alignment between parallel sentences, we use the online curated dictionaries, GIZA++ (Och and Ney, 2003) trained on the remaining IIT-B corpus, and cross-lingual word embedding trained on English and Hindi word vectors from FastText (Bojanowski et al., 2017). We employ eight human annotators[2] to provide a rating between 1 (low quality) to 10 (high quality) to the generated Hinglish sentences. Table 1 shows the annotation guidelines to rate the sentences. Figure 2a and 2b shows the distribution of the annotators' rating and their disagreement, respectively.

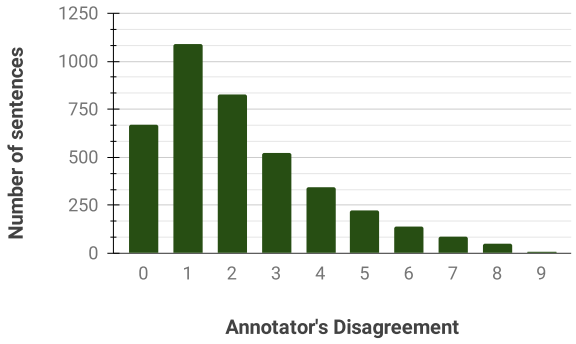
Data format: Table 2 shows an instance from the dataset. In total, we have 3,952 instances[3] in the dataset where each data instance i for subtask-1 (see Section 4.1) is represented as $\mathbf{X1}_i$={Eng$_i$, Hin$_i$, Synthetic_Hing$_i$} and $\mathbf{y1}_i$={Average_rating$_i$}. For subtask-2 (see Section 4.2), the instance j is represented as $\mathbf{X2}_j$={Eng$_j$, Hin$_j$, Synthetic_Hing$_j$} and $\mathbf{y2}_j$={Annotator_disagreement$_j$}. In addition, we provide at least two human generated Hinglish sentences corresponding to each data instance for both the subtasks. We shuffle and split the dataset in the ratio 70:10:20 with 2766, 395, and 791 data instances in train, validation, and test respectively. The more detailed description of the dataset is available in (Srivastava and Singh, 2021b).

[2]Different from the annotators in Phase 1. Each annotator gets 247 sentences generated by PAC and WAC, each corresponding to the same set of parallel sentences.

[3]Two synthetic Hinglish sentences are generated for each parallel sentence pair.



(a)



(b)

Figure 2: Distribution of (a) human evaluation scores and (b) disagreement in human scores in the synthetically generated Hinglish sentences.

Task	Guidelines
Hinglish text generation	1. The Hinglish sentence should be written in Roman script. 2. The Hinglish sentence should have words from both the source languages. 3. Avoid using new words, wherever possible, that are not present in both sentences. 4. If the source sentences are not the translation of each other, mark the sentence pair as "#".
Quality rating	The rating depends on the following three factors: 1. The similarity between the generated Hinglish sentence and the source sentences. 2. The readability of the generated sentence. 3. The grammatical correctness of the generated sentence.

Table 1: Annotation guidelines to the annotators for the two different tasks.

4 The Two Tasks

4.1 Subtask 1: Quality rating prediction

The first subtask is predicting the quality rating of the code-mixed text. The participating teams can use the English, Hindi, and human-generated Hinglish sentences to predict the average rating[4] as provided by the human annotators to the synthetic Hinglish sentences. In addition, we seek the teams to answer the following research questions implicitly with their experiments (not an exhaustive list):

- **RQ1.1**: Do the quality of source English and Hindi sentences impact Hinglish sentences' quality?
- **RQ1.2**: Does the quality of Hinglish text generated by humans has any correlation with the quality of Hinglish text generated synthetically?
- **RQ1.3**: Does the dominance of a language (English or Hindi) present in the Hinglish sentence impact the rating provided by the humans?
- **RQ1.4**: How does the semantic and the syntactic correctness of the Hinglish sentence influence its quality?

[4]We take the greatest integer i $\leq$ average of the two rating scores.

4.2 Subtask 2: Annotators' disagreement prediction

The next subtask is predicting the disagreement between the ratings provided by the human annotators to the synthetic Hinglish sentences. We calculate the disagreement between the ratings as the absolute difference between the two rating scores. Additionally, we seek the participating teams to answer the following research questions implicitly with their experiments (not an exhaustive list):

- **RQ2.1**: Does the quality of sentences in the source languages (English and Hindi) have any influence on the quality of the synthetic Hinglish sentences as seen by different individuals?
- **RQ2.2**: Does the quality of human-generated Hinglish sentence has any correlation with the quality of synthetic Hinglish text as seen by different individuals?
- **RQ2.3**: Do humans have a language bias while rating the quality of the code-mixed text?
- **RQ2.4**: Do the similarity between human-generated and synthetic Hinglish sentences influence the annotators' disagreement?

English	Hindi	Human-generated Hinglish	Synthetic Hinglish 1	Synthetic Hinglish 2
The reward of goodness shall be nothing but goodness.	अच्छाई का बदला अच्छाई के सिवा और क्या हो सकता है?	The reward of achai shall be nothing but achai. Goodness ka badla goodness ke siva aur kya ho sakta hai. Achai ka badla shall be nothing but achai.	reward ka badla reward ke nothing aur kya ho sakta hai **Rating1**: 7 **Rating2**: 4	reward of goodness goodness ke siva aur kya ho sakta hai **Rating1**: 9 **Rating2**: 7

Table 2: Example human-generated and synthetic Hinglish sentences from the dataset along with the source English and Hindi sentences. Two different human annotators rate the synthetic Hinglish sentences on the scale 1-10 (low-high quality).

5 Evaluation

We use the following three evaluation metrics:

- **F1-score (FS)**: We use the weighted F1-score to evaluate the system performance. The score ranges from 0 (worst) to 1 (best).
- **Cohen's Kappa (CK)**: We use the Cohen's Kappa score to measure the agreement between the predicted and the actual rating. The score ranges from ≤ 0 (high disagreement) to 1 (high agreement).
- **Mean Squared Error (MSE)**: MSE suggests the difference between the actual and the predicted scores. A low MSE score is preferred, with zero being the lowest possible score.

For the first subtask, we use all three metrics, whereas we use FS and MSE to evaluate the second subtask.

6 Pilot Experiment

We conducted a simple pilot experiment with a SOTA multilingual contextual language model M-BERT (Devlin et al., 2019). We fine-tune the pre-trained M-BERT model by adding one hidden-layer neural network on the top. We use the Relu activation function, AdamW optimizer with 0.03 learning rate, cross-entropy loss, and a batch size of 32. We use the contextual word-embedding corresponding to the synthetic Hinglish sentences in the dataset as an input to the model. The architecture remains the same for both subtasks.

Table 3 shows the result of the baseline experiment. We observe that the fine-tuned version of M-BERT performs poorly on both the subtasks on all the evaluation metrics. These language models are not as effective for both the subtasks as compared to other code-mixed text classification tasks where they seem to perform better than other rule-based and neural approaches (Gupta et al., 2021; Winata et al., 2021). Overall, we observe the poor performance of M-BERT based classifier on the current two subtasks. Specifically, for subtask 1, the agreement (measured by CK score) between predicted rating and human rating is close to 0. These results present an excellent opportunity to propose a shared task that enhances the generated code-mixed text quality estimation.

	Subtask 1			Subtask 2	
	FS	**CK**	**MSE**	**FS**	**MSE**
Val	0.202	0.003	2.797	0.209	4.987
Test	0.256	0.092	2.628	0.242	4.317

Table 3: Evaluation of the pilot experiment.

7 Conclusion

In contrast to the non-code-mixed text, the noisy and informal nature (e.g., spelling variation, missing punctuation, and language switching) of the code-mixed text makes the quality evaluation more loosely defined. Consequently, we need to build models that can effectively gauge the human perception of the quality of the code-mixed text. This shared task will help to build efficient and robust code-mixed text generation and evaluation systems.

References

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) at fire-2016. In *Forum for Information Retrieval Evaluation*, pages 39–49. Springer.

Ruthanna Barnett, Eva Codó, Eva Eppler, Montse Forcadell, Penelope Gardner-Chloros, Roeland van Hout, Melissa Moyer, Maria Carme Torras, Maria Teresa Turell, Mark Sebba, Marianne Starren, and Sietse Wensing. 2000. The lides coding manual: A document for preparing and analyzing language interaction data version 1.1—july, 1999. *International Journal of Bilingualism*, 4(2):131–132.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with

subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Navya Jose, Anand Kumar M, Thomas Mandl, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Hariharan R L, John P. McCrae, and Elizabeth Sherly. 2021. Findings of the shared task on offensive language identification in Tamil, Malayalam, and Kannada. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 133–145, Kyiv. Association for Computational Linguistics.

Amitava Das and Björn Gambäck. 2014. Identifying languages at the word level in code-mixed indian social media text. In *Proceedings of the 11th International Conference on Natural Language Processing*, pages 378–387.

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.

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.

Björn Gambäck and Amitava Das. 2016. Comparing the level of code-switching in corpora. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1850–1855.

K-I Goh and A-L Barabási. 2008. Burstiness and memory in complex systems. *EPL (Europhysics Letters)*, 81(4):48002.

Akshat Gupta, Sai Krishna Rallabandi, and Alan W Black. 2021. Task-specific pre-training and cross lingual transfer for sentiment analysis in dravidian code-switched languages. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 73–79.

Gualberto A Guzmán, Joseph Ricard, Jacqueline Serigos, Barbara E Bullock, and Almeida Jacqueline Toribio. 2017. Metrics for modeling code-switching across corpora. In *INTERSPEECH*, pages 67–71.

Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2018. The iit bombay english-hindi parallel corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

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.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Parth Patwa, Gustavo Aguilar, Sudipta Kar, Suraj Pandey, PYKL Srinivas, Björn Gambäck, Tanmoy Chakraborty, Thamar Solorio, and Amitava Das. 2020. Semeval-2020 task 9: Overview of sentiment analysis of code-mixed tweets. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 774–790.

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.

Shashi Shekhar, Dilip Kumar Sharma, and MM Sufyan Beg. 2020. Language identification framework in code-mixed social media text based on quantum lstm—the word belongs to which language? *Modern Physics Letters B*, 34(06):2050086.

Kushagra Singh, Indira Sen, and Ponnurangam Kumaraguru. 2018. Language identification and named entity recognition in hinglish code mixed tweets. In *Proceedings of ACL 2018, Student Research Workshop*, pages 52–58.

Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Ghoneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang, et al. 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.

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.

Vivek Srivastava and Mayank Singh. 2021a. Challenges and limitations with the metrics measuring the complexity of code-mixed text. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 6–14.

Vivek Srivastava and Mayank Singh. 2021b. Hinge: A dataset for generation and evaluation of code-mixed hinglish text. *arXiv preprint arXiv:2107.03760*.

Genta Indra Winata, Samuel Cahyawijaya, Zihan Liu, Zhaojiang Lin, Andrea Madotto, and Pascale Fung. 2021. Are multilingual models effective in code-switching? In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 142–153.