TranQuality: Improving Translation Quality Through Ensemble Translators and Paraphrase Detection

Brittany Dougall[a] and **Jaishankar Raju**[b]
[a]Master of Information and Data Science, University of California Berkeley, bdougall@ischool.berkeley.edu
[b]Master of Information and Data Science, University of California Berkeley, jaishankar.raju@ischool.berkeley.edu

Abstract

Translation quality estimation (QE) is a valuable component of expanding available data, particularly for low-resource languages. Prior studies have addressed QE using both Siamese neural networks and multilingual pre-trained language models (mPLMs), with high accuracy. The resultant quality measures have been used to improve overall corpus quality through ensemble methods to choose the most accurate translation for each record. In this paper, we expand upon previous Siamese neural network and mPLM approaches for QE and ensemble translator record selection, through the inclusion of IndicBERT, an Indian-language ALBERT model. Our proposed systems, a Siamese neural network using IndicBERT embeddings, and single-language trained IndicBERT models, find improvements in record-level cosine similarity scores and lower edit distances, despite decreased SacreBleu scores.

Keywords: Ensemble methods, Siamese neural networks, IndicBERT, IndicTrans, MBART.

1 Introduction

Although India comprises more than 17% of the world's population [19], with more than 1*bn* speakers [10], NLP research in this domain has lagged that of English and European languages. This is due in part to poor quality or little available training data [10] and a lack of labeled testing data, meaning that language models designed for these languages cannot be evaluated [9, 18]. Indian-language grammatical features such as agglutination [7], SOV structure, and relatively free word order also pose a challenge for traditional NLP tasks such as translation.

Recent research has made significant strides in addressing these challenges, but is not without flaws. IndicBERT, an Indian-language ALBERT trained on 12 major Indian languages, was recently released [10], with high quality embeddings for previously low-resource languages. Within the last 2 years, transformer models trained for Indian language translation have been developed, with accuracy as measured by SacreBleu scores surpassing those from prior state-of-the-art models [14]. These new models are trained on multiple languages, exploiting the grammatical similarity between Indian languages to boost performance of those considered low-resource. However, as multilingual models, these Indian-language trained transformers also possess performance flaws associated with a multilingual design. Translation quality has been found to be inconsistent, with lower performance for low-resource languages compared to higher resource ones and lower performance for high resource languages than that of bilingual models [21, 8].

In our study, we seek to remedy the inconsistent quality of two Indian language transformers through paraphrase evaluation, which has previously been used as an extrinsic measure of translation quality [20, 1, 17]. After translating English into Hindi, Tamil, and Malayalam with IndicTrans and MBART, two transformer models pre-trained on Indian languages, we use paraphrase evaluation for QE. We chose to use IndicTrans and MBART since we wanted to assess our models' performance on high quality translations and since these transformers are both multilingual. IndicTrans is an MT5 transformer that is pre-trained for translation tasks on the Samanantar dataset, the largest publicly available Indic language corpus [14]. For MBART, we used the MBart-large-50 model from the checkpoint 'facebook/mbart-large-50-many-to-many-mmt', since it has already been fine-tuned for Indian language translation tasks. However, since MBART has only been trained on several Indian languages, we chose to limit our evaluation of translation quality to Hindi, Tamil, and

Malayalam, which are covered by both IndicTrans and MBART. We hypothesize that using paraphrase scoring models to choose the best translation for each record will result in corpora with higher SacreBleu scores than those obtained from the translations of a single model. We chose SacreBleu scores due to their easy implementation, widespread use in prior Indian language studies, and since these scores are a language-independent metric that can be computed for all of our 3 languages.

2 Background

Prior studies using ensemble methods to improve translation corpus quality have demonstrated the benefit of such an approach. A 2017 study using a neural network to choose the best English to Chinese translation between statistical machine translations (SMT) and neural machine translations (NMT) found improvements in SacreBleu scores over 5 points higher than the best single translator output [23]. A 2016 study using ensemble NMT translators with context-dependent weighting to compensate for individual translator errors resulted in an improvement up to 2.2 Bleu points higher than an individual model baseline [5]. A more recent study in 2021 found that averaging HTER scores produced by 3 mBERT models fine-tuned for multiple language translation tasks resulted in higher Pearson's correlation scores with the ground truth HTER score than any of the 3 individual models, across all 6 translation tasks [2].

Siamese networks have been shown to be effective and highly accurate in evaluating sentence similarity for both single language and multilingual tasks. As such, this structure is well-suited for an ensemble method to choose the best translation quality from multiple translators. In the 2019 study introducing SentenceBERT, the authors found that a single Siamese BERT network outperformed InferSent and the Universal Sentence Encoder across both paraphrase detection and multiple English language sentiment analysis tasks [16]. A 2020 study found that a single Siamese LSTM network outperformed a word overlap baseline linear regression model on predicting sentence similarity for both Portuguese and English across multiple domains [3]. Although Siamese BERT models have not yet been applied for evaluating Indian language translation, a prior study evaluating translation quality with Siamese CNNs found higher Pearson's correlation scores with human evaluators than an SVR baseline for an Indian language parallel corpora [8].

Multilingual pre-trained language models (mPLMs) have also been used directly for QE, suggesting a fine-tuned IndicBERT model may be a viable candidate for translation quality assessment. Besides the previously referenced study using 3 mBERT models, other researchers have found mPLMs to achieve high accuracy on QE tasks. In a 2020 study, both a Siamese XLM-R network and a single XLM-R model were trained for QE. While both models showed higher Pearson correlation scores with human evaluations than the baseline LSTM, the single XLM-R model received a higher average score than the Siamese network across the languages studied [15]. A 2*nd* study found that training an XLM-R model directly with direct assessment (DA) quality scores and relative ranking scores (RR) resulted in higher Pearson's correlation scores with human evaluators than the winning systems from the WMT 2019 QE task [22]. IndicBERT has previously been fine-tuned for paraphrase detection and scored highly - 93.75% accuracy on exact paraphrase detection and 84.33% accuracy on rough paraphrase detection, using the Amrita paraphrase dataset that we used in this study [10]. However, an IndicBERT checkpoint for paraphrase detection has not been publicly released.

Based upon these prior studies, we chose to test our hypothesis using both IndicBERT models fine tuned for paraphrase evaluation and a Siamese neural network using IndicBERT embeddings as model input.

3 Data

Language	Training	Validation	Test
Hi	2.5k	448	448
Ta	2.5k	425	426
Ml	2.5k	450	450
Pb	1.7k	250	250

Table 1: Amrita Paraphrase Sentence Pairs by Language and Train/Dev/Test Split

To train our paraphrase classification models, we used the Amrita paraphrase dataset [11], composed of sentence pairs from news articles in Hindi (Hi), Malayalam (Ml), Tamil (Ta), and Punjabi (Pb). In this dataset, each sentence pair is classified as paraphrases of each other (P) or not paraphrases of each other (NP). Classification was first performed by students at Amrita University, then verified by a language expert, then a linguistic expert in a sequential process. In the event of a conflicting label, the linguistic expert's label overrides the language expert label, which overrides the student provided label (inter-annotator agreement scores were not available). Each language's sentence pair counts after removing duplicate records in the original test set and splitting into validation and test sets is shown in Table 1. Each language's training corpus is exactly 60% NP and 40% P, while the validation and test dataset compositions vary by language due to our random split

to balance the records (see Appendix A Table 8 for the validation and test composition, rounded to the nearest percentage). Overall validation composition is approximately 55% NP and 45% P; test composition is approximately 56% NP and 44%. For our approach directly training IndicBERT for paraphrase evaluation, we used Hi, Ml, and Ta, with a separate model for each language. Since we trained on fewer records, IndicBERT test accuracy and loss rates were computed across each language's original test set (its combined validation and test sets). For our SNN, all languages were used and each language's test set contained only the test records obtained after splitting.

For our translation task, we used the dev dataset from the WAT 2021 8th workshop on machine translation [4] since it was already pre-cleaned and the target text translations were of high quality. This dataset was curated from PMIndia, a parallel language corpus commonly used for Indian NLP tasks. It consists of 1k language pairs for each Indian language translation, extracted from the website of the Prime Minister of India, with the dev set utilizing the same English sentences for each translation task. We used 4k sentences (the 1k English original sentences and their target translations in Hindi, Malayalam, and Tamil) for our analysis.

4 Methods

4.1 Siamese Neural Network

4.1.1 Framework

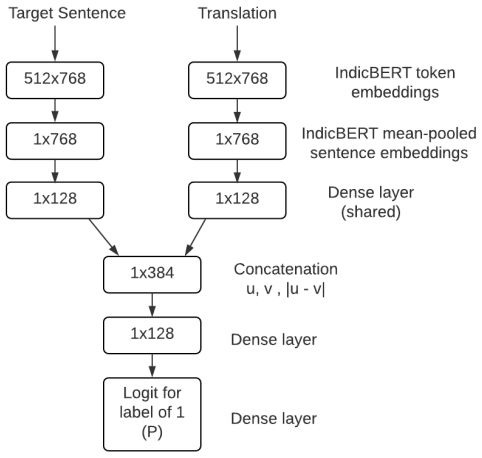


Figure 1: Diagram of Siamese network architecture.

Figure 1 shows the design of our Siamese neural network (SNN) and input processes. Each sentence in the target and translation pair is first separately tokenized using IndicBERT's tokenizer to obtain the inputs. The tokenized inputs are then separately passed to our IndicBERT model and mean-pooled to obtain the sentence embeddings for each sentence in the pair. From there, each sentence in the pair separately passes through the same dense layer, allowing them to share the same weights and biases. In that layer, each sentence's dimensionality is reduced from 768 to 128, since most of the sentences are short in length and reducing dimensionality is not expected to result in significant information loss. Each batch of sentence pairs is then normalized using LayerNorm since the sentences in the batch are of differing lengths, followed by relu activation. After separately passing through the dense layer, the embeddings of each sentence in a given pair are concatenated, along with the absolute value of the element-wise difference between target and translation since this method of concatenation has been observed to result in greater model accuracy than simply concatenating the embeddings [16]. The concatenated embeddings are then normalized again with LayerNorm, passed through relu activation, and passed through a 2*nd* dense layer to reduce the dimensionality of the concatenated vectors back to 128. The batched inputs are then normalized again with LayerNorm, passed through relu activation, and passed through dropout. Lastly, the embeddings pass through a 3*rd* dense layer to reduce the dimensionality of each sentence pair concatenated embedding from 128 to 1, an output logit for the sentence pair relationship which is reshaped to have the same shape as the target label. During training, validation, and testing, these logits are converted to a label of 0 (NP) or 1 (P) and compared against the target label to assess accuracy.

4.1.2 Training

To train our SNN, we used the Amrita paraphrase training dataset, passing sentence 1 and sentence 2 into our model in Figure 1 in lieu of the target sentence and translation respectively. Since the class distribution wasn't balanced in the training data and to allow multiple languages to be selected in each training batch, we randomly shuffled the training records prior to training and used a weighted random sampler with replacement to ensure that the number of P and NP records in each batch was approximately equal (since 60% of training records were NP). During training, we used an Adam optimizer with BCEWithLogitLoss as our loss function and a learning rate of 2e-5. We chose to use BCEWithLogitLoss as our loss function since we are training our neural network for binary classification and since this loss function is more numerically stable than binary cross entropy following a sigmoid layer [13]. A learning rate of 2e-5 was chosen since it is a common learning rate for training neural networks and since higher initial learning rates produced higher validation loss, even when used with scheduler step.

4.1.3 Results

To choose the best SNN, we first trained the model with only Hi, Ta, and Ml paraphrase records for 10 epochs and used a dropout rate of 0.2. While the model was fairly accurate in its predictions on the training and test set (Table 2), test accuracy was lower than we anticipated, with the accuracy rate depressed by the model's performance on Tamil. Hi had an accuracy of 86%, Malayalam 78%, and Tamil 65%. The test f1-score was 0.71, with the majority of errors P pairs

labeled as NP (53% of which were Tamil sentence pairs).

We then decided to incorporate additional training data from Punjabi, since we reasoned that Pb P and NP pairs may have a different degree of separation than those currently seen in training. We believed that including Punjabi might benefit Hindi accuracy as well by better balancing language type representation in training (Hi and Pb are Aryan languages; Ta and Ml are Dravidian). In the trial with Punjabi, we increased both the number of training epochs to 15 and the training dropout from 0.2 to 0.5 to avoid over-fitting. This attempt resulted in improved training and test accuracy, with a small increase in validation accuracy (Table 2). Tamil accuracy rates increased to 70%, Hi to 87%, and the accuracy of Ml remained at 78% (Pb accuracy is 98%). While misclassification of P pairs decreased, it remained the predominant error type, with Ta sentences accounting for 47% of these errors. Correct labels were assigned with a high degree of confidence across all languages; incorrect labels were assigned with a lesser, but still high degree of confidence. Sentence error types vary by language, but include a failure to recognize the same entity when different names are used, misclassification after replacement with non-equivalent entity names, and mislabeling when newspaper headlines replace text (examples in Appendix B Figure 3).

Since we saw little improvement in Ta or Ml performance with additional training data, we reasoned that the P and NP pairs of these languages may be fundamentally different from that of the Aryan languages Hindi and Punjabi. We thought that we may see better model performance by creating separate models for the Aryan languages and Dravidian languages and reducing the number of training epochs back to 10 due to a decreased number of training data records. While this approach had a small benefit to Hi accuracy (increased to 88%), Ta accuracy and Ml accuracy declined (down to 64% and 74% respectively). We believe that this is due to Dravidian NP pairs being closer to P pairs than their Aryan counterparts and that the presence of some highly similar NP pairs in the Aryan language corpus aids the model in distinguishing between Dravidian language P and NP pairs. This theory is supported by an analysis of cosine scores between between correctly and incorrectly classified P/NP pairs from the best-performing SNN (Appendix C Figure 5), which shows that the median cosine similarity scores of Tamil P and NP pairs are close in value and that there are NP pairs possessing high cosine similarity within the Hindi and Punjabi corpora.

Metric	SNN w/o Pb	SNN w/ Pb
Train accuracy	91.15%	94.59%
Train loss	0.267	0.177
Validation accuracy	77.63%	82.04%
Validation loss	0.509	0.432
Test accuracy	76.96%	81.26%
Test loss	0.515	0.446

Table 2: Training, Validation, and Test Accuracy and Loss in SNN Trials.

4.2 Fine-Tuning IndicBERT

4.2.1 Training

For this model, we fine tuned IndicBERT [10] for a sequence classification task. We used the same Amrita paraphrase training dataset, that we used for the SNN. We did not try to weight for the difference in distribution of the classes. The inputs were pairs of sentences in the form [CLS] A [SEP] B [SEP] and the target was a class [P, NP] representing Paraphrase and Non Paraphrase. We created one model each for Tamil, Hindi and Malayalam. We grid searched four hyper-parameters listed in Table 3. The model outputs the logits for the two classes.The best model results are listed in Table 4. Models trained for 1 epoch with a max sequence length of 256, a learning rate of 2e-5, and a batch size of 16 achieved the highest test accuracy and lowest test loss.

Hyper parameter	Search range
Epochs	[1-5]
Max Seq length	[128, 256, 512]
Batch size	[8, 16, 32]
Learning rate	[2e-04, 2e-05, 2e-06]

Table 3: Hyper parameter search for IndicBERT fine-tuning.

Model	Accuracy	Loss
TA	86.13%	0.445
HI	90.21%	0.341
ML	87.55%	0.441

Table 4: Test Accuracy and Loss from the fine-tuned IndicBERT models using the best hyperparameters. The previously stated best-performing hyperparameters were used in all 3 models.

4.2.2 Results

In analyzing model misclassification, we noticed several themes - contextual replacement of entities, introduction of entities (proper nouns), and confusion between quotes and a report. Appendix B Figure 4 shows one example of each of these themes. The highlighted text shows what is different between the inputs. In these models, the most common classification error varied by language, with Tamil having the highest error rate (misclassification of NP pairs as P).

4.3 Translation

To produce our translations, we separately passed each record in the English dev dataset from the PMIndia corpus to IndicTrans and MBART and produced a translation of each record in Hi, Ta, and Ml. These translations along with their original sentences were passed to our models as inputs for scoring separately. This is due to the structure of the Amrita

training dataset, which contained only a paraphrase or a non-paraphrase in each pair (sentence 1 and sentence 2, which is either a P or NP of sentence 1), meaning that we were unable to train the network to evaluate 2 sentences against the same target simultaneously. Our models output a logit (SNN) or pair of logits (IndicBERT) for each translation created by MBART and IndicTrans. A sigmoid (SNN) or softmax (IndicBERT) function was then applied to the logits to find each translation's paraphrase score, as is depicted in Figure 2, and the translation for each record with the higher paraphrase probability score was selected. In the event of a tie (found only when both translators produced the same translation), IndicTrans' translation was used.

For each language, we then computed SacreBleu scores [12] [1] using the intl tokenizer, as it supports non-English languages, including non-ASCII characters. As seen in Table 5, translation quality varies between language tasks. Hindi achieves the highest SacreBleu score for both models, with IndicTrans translations achieving higher scores on the translated dev set for all 3 languages.

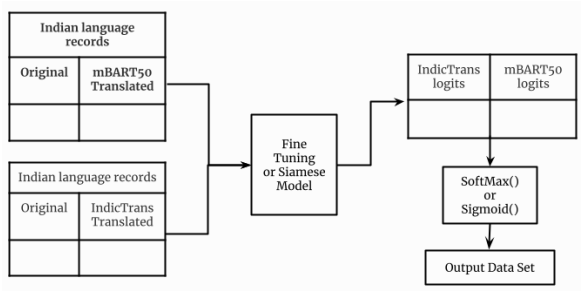


Figure 2: Pipeline for generating best translations output

Translation	Indic Trans	MBART
En-Hi	39.1	28.1
En-Ta	15.4	13.0
En-Ml	8.3	1.3

Table 5: SacreBleu Scores by Translation and Translator Used.

5 Translation Results and Discussion

5.1 Siamese Neural Network

Since the SNN trained with Pb achieved the best combined accuracy on our target languages and had relatively low test loss, we chose to use it to evaluate the quality of our translations. Of the 3k target records, our model scored IndicTrans' translation as closer to the target text than that of MBART for 64% of all records with differing MBART and IndicTrans translation probabilities (for 33 target sentences, 24 Hi and 9 Ta, MBART and IndicTrans produced the same sentence with equivalent P probabilities). This pattern held true for all 3 languages, with the closest split for Ta (MBART produced the better translation 46% of the time). For Hi translations, the SNN had high confidence that both MBART and IndicTrans' translations are paraphrases of the target text. For Ta, the SNN assigned the majority of each translators' translations a label of NP; however, when it classified either the IndicTrans or MBART translation as a paraphrase of the target text, it did so with confidence. For Ml, the SNN also classified most MBART translations as NP and close to 50% of the IndicTrans translations as NP (Appendix D Figure 6). However, for both Ta and Ml, when the translation was predicted to be a paraphrase, the probability score was high.

Despite choosing IndicTrans' translation in most cases where P probabilities differed, SacreBleu scores for Hi and Ml were lower than those from IndicTrans as the sole translator source (Table6). For these languages, IndicTrans' translation was chosen for 69% of target records and 73% of target records respectively where translations were non-equivalent. For Ta, IndicTrans' translations were selected for 53% of non-equivalent translations.

Translation	Indic Trans	MBART	SNN w/ Pb
En-Hi	39.1	28.1	38.7
En-Ta	15.4	13.0	15.5
En-Ml	8.3	1.3	7.2

Table 6: Translation SacreBleu scores by individual translator and ensemble corpora selected by the SNN trained with Pb.

Additional analysis suggests that the decrease in SacreBleu is not necessarily due to model failure. The sentences for which MBART is considered the better translator tend to have a cosine similarity score closer to that of the target sentence than do the translations not chosen from IndicTrans, with the reverse being true when IndicTrans is considered to have produced the better translation, across all languages (77% of cases). The answer to the decrease in scores appears to be revealed in an examination of Levenshtein character edit distances. Translations selected from IndicTrans tend to have lower edit distances than the non-selected MBART sentences (79% of the time). Conversely, when MBART's translation is selected, for 54% of target records, the edit distance is greater than that of the non-chosen IndicTrans translation. These results suggest that although Hi and Ml SacreBleu scores are lower than IndicTrans as the sole translator, overall translation quality across the translations improves or remains the same.

5.2 IndicBERT

Although the best-performing IndicBERT models achieved high accuracy on P/NP classification, SacreBleu scores on the selected translations were similar or lower than those from IndicTrans alone, as seen in Table 7. Of particular note is the Hi IndicBERT model, which had a test accuracy of 91%, yet its

[1] $Signature$: $nrefs$: 1|$case$: $mixed|eff$: $no|tok$: $intl|smooth$: $exp|version$: 2.0.0.

highest scored translations had a SacreBleu score 8.7 points lower than those of IndicTrans as the sole translator. For the Hi IndicBERT selected records, cosine similarity scores tended to be lower than the non-chosen translation regardless of translator selected. Since the MBART translation was selected more often than IndicTrans' translation, edit distances tended to be higher than the non-selected translations (in 63% of cases where the translations were non-equivalent).

Translation	Indic Trans	MBART	Best IndicBERT
En-Hi	39.1	28.1	30.4
En-Ta	15.4	13.0	15.9
En-Ml	8.3	1.3	7.5

Table 7: Translation SacreBleu scores by individual translator and ensemble corpora selected by the best-performing IndicBERT models.

For the Ta and Ml IndicBERT models, cosine similarity score and edit distance trends varied by language and translator selected. For Ta, the selected IndicTrans translations had a lower cosine similarity score to the target text in 53% of non-equivalent translations, but the edit distances tended to be closer to the target text (for 60% of non-equivalent records). Conversely, when MBART was selected, its records tended to have both a higher cosine similarity score and a lower edit distance than the non-selected IndicTrans record (for 68% of records and 59% of records respectively). Overall, edit distances of the selected translation were higher than of the non-chosen translation, in approximately 64% of cases. Cosine similarity scores overall were higher than the non-chosen translation for approximately 56% of selected records. For Ml, cosine similarity scores were higher than of the non-chosen translation for 58% of cases, with scores tending to increase regardless of translator source. Edit distances when an IndicTrans translation were chosen tended to be lower than those of MBART; when MBART's translations were selected, edit distances increased in 59% of cases. However, since IndicTrans' translations were selected for 83% of target records, edit distances overall tended to be lower for the chosen than the non-chosen record.

5.3 Model Comparison

During our analysis we found that unlike the SNN, a Hi IndicBERT model chose the MBART translation for approximately 60% of Hindi records vs 31% of the time for the SNN. While both models predicted that at least one of the translators produced a paraphrase for most records, Hi IndicBERT assigned a P label with a lower degree of confidence to both sets of translations (Appendix D Figure 6), with little difference between MBART and IndicTrans scores. This difference could potentially be due to dialect differences between the Hindi training records and the translation corpus - while reviewing the target translations, we discovered that the PMIndia corpus contains records in an older Hindi dialect than is present in the Amrita paraphrase training corpus. Having been trained on one dialect, a Hi IndicBERT model may have lower confidence when predicting P/NP labels in another, albeit similar, dialect. Conversely, since the SNN has learned to assign labels by the difference between the embeddings in a multilingual context and has seen a range of embedding differences, it is still confident when making predictions on an unseen dialect.

The Ta and Ml IndicBERT models, however, assign high probability scores to the translations when considering them paraphrases of the target text (Appendix D Figure 6). In comparison, SNN classifies most translations, for both IndicTrans and MBART, as NP of the target text. This phenomena reflects SNN's classification errors for Ta and Ml, where the model's most common mistake was classifying P pairs as NP. Nevertheless, since the SNN classifies both MBART and IndicTrans translations as NP with approximately equivalent probabilities, this behavior has little impact upon SacreBleu scores. For Ml, both IndicBERT and the SNN network classify a high percentage of MBART translations as non-paraphrases (59% of translations and 79% respectively). These high NP rates reflect MBART translation errors, in which multiple sentences were either left in their original English form or only part of the sentence was translated.

6 Conclusion and Next Steps

The corpora created by selecting each target record's highest scored translation had lower SacreBleu scores than those of IndicTrans alone, save for Tamil. However, higher cosine similarity scores of selected than non-selected translations for the SNN, Ta IndicBERT, and Ml IndicBERT models indicate that measuring SacreBleu alone may not adequately capture whether or not ensemble translation quality is better or worse than a single translator and suggest avenues for future analysis. One potential approach would be to obtain a dataset with multiple translations for each target with pre-existing quality rankings, and then evaluate whether or not our models produce the same rankings.

Additional study observations pose questions for future courses of action. During our analysis, we noted that the Hi IndicBERT model predicted similarly low probabilities for both MBART and IndicTrans translations, possibly due to its training on a different dialect than that seen in translation evaluation. The model's performance would be better understood by classifying the dialect of all Hi pairs in both the PMIndia and Amrita paraphrase corpora, to verify the extent to which the dialect composition differs. If dialect composition is found to substantially differ, one could explore training an IndicBERT model with multiple dialects or a different dataset containing the older Hindi dialect seen in the PMIndia corpus. For both of the ensemble methods, if a dataset with examples of a paraphrase and a non-paraphrase were obtained and the model trained using triplet loss, this may result in fewer classification errors, particularly for Tamil, where NP and P pairs have very close cosine similarity values. Training with bidirectional hard-negatives ranking loss, which emphasizes hard negatives (the negative pair closest to a positive pair) [6] may also improve Tamil classification accuracy, particularly for the SNN.

References

[1] C. Callison-Burch, Paraphrasing and translation, Ph.D. thesis, University of Edinburgh Edinburgh (2007).

[2] S. Chowdhury, N. Baili, B. Vannah, Ensemble fine-tuned mbert for translation quality estimation, arXiv preprint arXiv:2109.03914.

[3] J. V. A. de Souza, L. E. S. E. Oliveira, Y. B. Gumiel, D. R. Carvalho, C. M. C. Moro, Exploiting siamese neural networks on short text similarity tasks for multiple domains and languages, in: International Conference on Computational Processing of the Portuguese Language, Springer, Cham, 2020, pp. 357–367.

[4] A. for Computational Linguistics, The 8th Workshop on Asian Translation, http://lotus.kuee.kyoto-u.ac.jp/WAT/WAT2021/, accessed 2021-11-14.

[5] E. Garmash, C. Monz, Ensemble learning for multi-source neural machine translation, in: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, 2016, pp. 1409–1418.

[6] X. Hao, Bi-directional hard-negatives ranking loss for cross-modal video-text retrieval, 2020.

[7] B. Harish, R. K. Rangan, A comprehensive survey on indian regional language processing, SN Applied Sciences 2 (7) (2020) 1–16.

[8] N. Jhaveri, M. Gupta, V. Varma, Translation quality estimation for indian languages, 2018.

[9] P. Joshi, S. Santy, A. Budhiraja, K. Bali, M. Choudhury, The state and fate of linguistic diversity and inclusion in the nlp world, arXiv preprint arXiv:2004.09095.

[10] D. Kakwani, A. Kunchukuttan, S. Golla, N. Gokul, A. Bhattacharyya, M. M. Khapra, P. Kumar, inlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, 2020, pp. 4948–4961.

[11] S. Kumar, Shared task on detecting paraphrases in indian languages (dpil), http://www.nlp.amrita.edu/dpil_cen/.

[12] M. Post, A call for clarity in reporting BLEU scores, in: Proceedings of the Third Conference on Machine Translation: Research Papers, Association for Computational Linguistics, Belgium, Brussels, 2018, pp. 186–191.
URL https://www.aclweb.org/anthology/W18-6319

[13] PyTorch, BCEWithLogitsLoss, https://pytorch.org/docs/stable/generated/torch.nn.BCEWithLogitsLoss.html, accessed 2021-11-14.

[14] G. Ramesh, S. Doddapaneni, A. Bheemaraj, M. Jobanputra, R. AK, A. Sharma, S. Sahoo, H. Diddee, D. Kakwani, N. Kumar, et al., Samanantar: The largest publicly available parallel corpora collection for 11 indic languages, arXiv preprint arXiv:2104.05596.

[15] T. Ranasinghe, C. Orasan, R. Mitkov, Transquest at wmt2020: Sentence-level direct assessment, arXiv preprint arXiv:2010.05318.

[16] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, arXiv preprint arXiv:1908.10084.

[17] P. Resnik, O. Buzek, C. Hu, Y. Kronrod, A. Quinn, B. B. Bederson, Improving translation via targeted paraphrasing, in: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, 2010, pp. 127–137.

[18] A. Srinivasan, S. Sitaram, T. Ganu, S. Dandapat, K. Bali, M. Choudhury, Predicting the performance of multilingual nlp models, arXiv preprint arXiv:2110.08875.

[19] StatisticsTimes, India population 2021, https://statisticstimes.com/demographics/country/india-population.php#:~:text=India%20accounts%20for%20a%20meager,its%20global%20share%20is%20decreasing,accessed 2021-11-14.

[20] B. Thompson, M. Post, Automatic machine translation evaluation in many languages via zero-shot paraphrasing, arXiv preprint arXiv:2004.14564.

[21] S. Wu, M. Dredze, Are all languages created equal in multilingual bert?, arXiv preprint arXiv:2005.09093.

[22] J. Zhang, J. van Genabith, Translation quality estimation by jointly learning to score and rank, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 2592–2598.

[23] L. Zhou, W. Hu, J. Zhang, C. Zong, Neural system combination for machine translation, arXiv preprint arXiv:1704.06393.

Appendix A P/NP Composition

Table 8: P/NP Composition of Paraphrase Validation and Test Sets

Table 9: Validation

Lang	NP	P
Hi	55%	45%
Ta	58%	42%
Ml	54%	46%
Pb	52%	48%

Table 10: Test

Lang	NP	P
Hi	55%	45%
Ta	58%	42%
Ml	54%	46%
Pb	52%	48%

Appendix B Mislabeled Paraphrase Test Sentence Pairs

Sentence1	Sentence2	Predicted Class
भारत ने बुधवार को परमाणु क्षमता संपन्न और स्वदेश में विकसित पृथ्वी-दो' मिसाइल का सफल प्रक्षेपण किया। India on Wednesday successfully launched the nuclear-capable and indigenously developed Prithvi-II missile.	डीआरडीओ ने बुधवार को स्वदेशी निर्मित पृथ्वी-दो' मिसाइल का सफल परीक्षण किया है। DRDO on Wednesday successfully test-fired indigenously built Prithvi-II missile.	P
முதல்-அமைச்சர் ஜெயலலிதா இன்று சட்டமன்றப் பேரவையில், கச்சத்தீவு பிரச்சினை குறித்து எதிர்க்கட்சித் தலைவர் மு.க. ஸ்டாலின் பேசியதற்கு பதிலளித்து பேசினார். Chief Minister Jayalalithaa today addressed the Assembly on the issue of Kachchativu as a response to th e opposition leader M K. Stalin	எதிர்க்கட்சித் தலைவர் ஸ்டாலின் பேசும் போது தமிழக அரசை மத்திய அரசு கலந்தாலோசிக்கவில்லை எனத் தெரிவித்தார். Opposition leader Stalin said the central government had not consulted the Tamil Nadu government.	NP
மாலத்தீவு அதிபரை கொலை செய்ய முயன்றதாக கூறப்பட்ட வழக்கில், முன்னாள் துணை அதிபர் அகமது அதீப்புக்கு 15 ஆண்டு சிறை தண்டனை விதிக்கப்பட்டுள்ளது. Former Vice President Ahmed Adeeb has been sentenced to 15 years in prison for allegedly trying to assassinate the Maldivian president.	மாலத்தீவு அதிபரை கொல்ல முயற்சி: முன்னாள்துணைஅதிபருக்கு15 ஆண்டு சிறை தண்டனை Attempt to assassinate Maldivian president: Former Vice President sentenced to 15 years in prison.	NP

Figure 3: Examples of SNN Mislabeled Paraphrase Test Sentence Pairs.

Sentence1	Sentence2	Predicted Class
इसमें लिक्विड प्रोपल्शन ट्विन इंजन लगे हैं। This has Liquid Propulsion Twin Engines.	एडवांस टेक्नोलॉजी वाली पृथ्वी-दो मिसाइल में दो इंजन लगाए गए हैं। Advance tech Pritvi-II missile has two engines	NP
इसके साथ ही इसी हफ्ते के अन्त में सीबीआइ एसपी त्यागी के तीनों भाइयों से भी पूछताछ करेगी। Along with this, at the end of this week, CBI will also interrogate the three brothers of SP Tyagi.	इस सप्ताह के अंत में वह त्यागी के रिश्तेदारों - संजीव, राजीव और संदीप से पूछताछ करने जा रही है. Later this week she is going to question Tyagi's relatives - Sanjeev, Rajeev and Sandeep.	NP
'குளச்சல் அருகே துறைமுகம் கட்டுவது சரியல்ல. இந்த பிரச்சி னை குறித்து பிரதமர் மோடியை சந்தித்து முறையிடுவேன்'' என்று கேரள முதல்வர் பினராயி விஜயன் கூறியுள்ளார் 'It is not right to build a port near Kulachal. I will meet Prime Minister Modi and appeal on this issue, "said Kerala Chief Minister Binarayi Vijayan	குளச்சலில் துறைமுகம் ஏற்படுத்த பிரதமர் மோடியிடம் எதிர்ப்பு தெரிவிப்பேன்: சட்டப்பேரவையில் கேரள முதல்வர் பதில் I will oppose Prime Minister Modi to build a port in Kulachal: Kerala Chief Minister's reply in the Assembly	NP

Figure 4: Examples of IndicBERT Mislabeled Paraphrase Test Sentence Pairs.

Appendix C Paraphrase Test Set Cosine Similarity

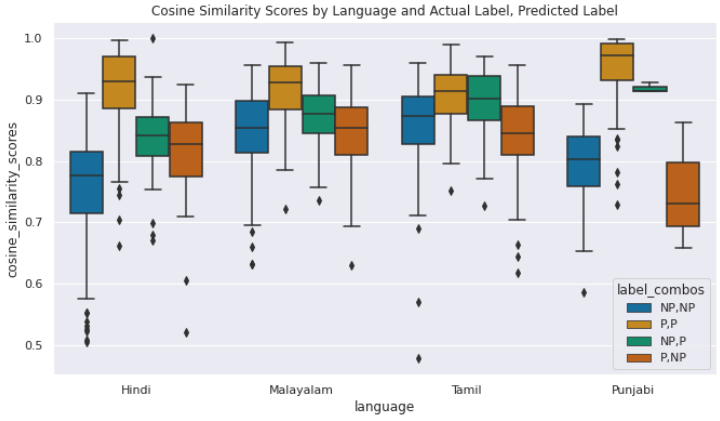
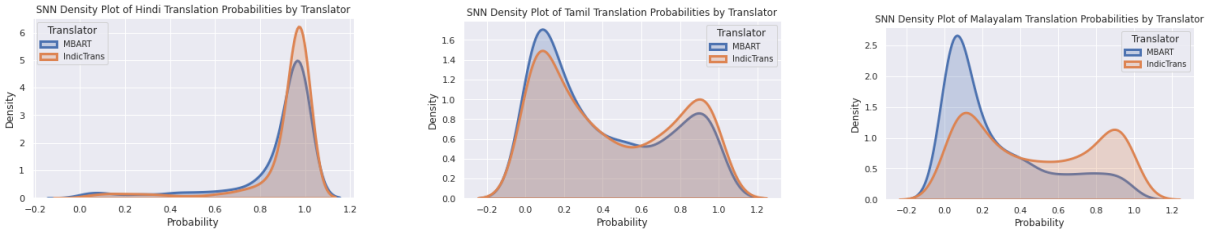
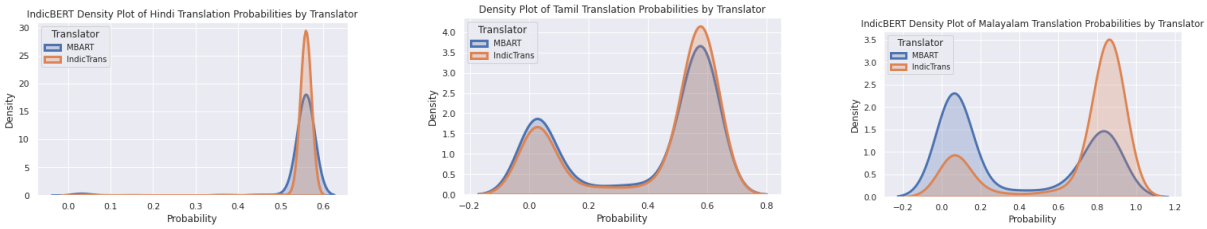


Figure 5: SNN Trained w/ Pb Paraphrase Test Cosine Similarity Scores by Language, Actual, and Predicted Label.

Appendix D Comparison of Translator Translation Probabilities



(a) SNN



(b) IndicBERT

Figure 6: IndicTrans and MBART Translation Probabilities by Language and Model. MBART predictions are shown in blue; IndicTrans in orange.