"A Little is Enough": Few-Shot Quality Estimation based Corpus Filtering improves Machine Translation

Akshay Batheja, Pushpak Bhattacharyya

CFILT, Indian Institute of Technology Bombay
{akshaybatheja, pb}@cse.iitb.ac.in

Abstract

Quality Estimation (QE) is the task of evaluating the quality of a translation when reference translation is not available. The goal of QE aligns with the task of corpus filtering, where we assign the quality score to the sentence pairs present in the pseudo-parallel corpus. We propose a Quality Estimation based Filtering approach to extract high-quality parallel data from the pseudo-parallel corpus. To the best of our knowledge, this is a novel adaptation of the QE framework to extract quality parallel corpus from the pseudo-parallel corpus. By training with this filtered corpus, we observe an improvement in the Machine Translation (MT) system's performance by up to **1.8** BLEU points, for English-Marathi, Chinese-English, and Hindi-Bengali language pairs, over the baseline model. The baseline model is the one that is trained on the whole pseudo-parallel corpus. Our Few-shot QE model transfer learned from the English-Marathi QE model and fine-tuned on only 500 Hindi-Bengali training instances, shows an improvement of up to **0.6** BLEU points for Hindi-Bengali language pair, compared to the baseline model. This demonstrates the promise of transfer learning in the setting under discussion. QE systems typically require in the order of (7K-25K) of training data. Our Hindi-Bengali QE is trained on only 500 instances of training that is $1/40^{th}$ of the normal requirement and achieves comparable performance. All the scripts and datasets utilized in this study will be publicly available.

1 Introduction

In recent times, Neural MT has shown excellent performance, having been trained on a large amount of parallel corpora (Dabre et al., 2020). However, not all language pairs have a substantial amount of parallel data. Hence, we have to rely on the noisy web-crawled corpora for low-resource languages. The task of **Parallel Corpus Filtering** aims to provide a scoring mechanism that helps extract good-quality parallel corpus from a noisy pseudo-parallel corpus. The task of **Quality Estimation** (QE) aims to provide a quality score for a translation when the reference translation is unavailable. We use Quality Estimation to assign the quality scores to the sentence pairs present in pseudo-parallel corpora and extract good-quality parallel sentences. We aim to improve the quality of Machine Translation for English(En)-Marathi(Mr), Hindi(Hi)-Bengali(Bn) and Chinese(Zh)-English(En) language pairs by using sentence-level QE-based corpus filtering. We observe that QE-based corpus filtering performs better than previously proposed methods.

Our contributions are:

1. Adaptation of the QE framework, which is normally used for MT evaluation, to extract high-quality parallel corpus from pseudo-parallel corpus; to the best of our knowledge, this is a novel adaptation of the QE framework to extracting quality parallel corpus from the pseudo-parallel corpus.

2. Demonstrating the promise of Few-Shot QE technique to generate training data for MT; a Hindi-Bengali QE model is trained with only 500 training instances transfer learned from an English-Marathi trained QE model; the filtered parallel data using this Hindi-Bengali QE system gives **0.6** BLEU point improvement over Hi-Bn MT system trained on the pseudo-parallel corpus.

3. Demonstrating performance improvement of the Machine Translation systems by up to **1.8** BLEU points for English-Marathi, Hindi-Bengali and Chinese-English language pairs, over the model trained on the whole pseudo-parallel corpus.

Findings of the Association for Computational Linguistics: ACL 2023, pages 14175–14185
July 9-14, 2023

2 Related work

2.1 Parallel Corpus Filtering

Neural Machine Translation (NMT) is extremely *data hungry* (Sutskever et al., 2014; Bahdanau et al., 2015; Vaswani et al., 2017). Recently, there has been a growing interest in the process of filtering noisy parallel corpora to enhance the data used for training machine translation systems. The Conference on Machine Translation (WMT) has organized annual Shared Tasks on Parallel Corpus Filtering (WMT 2018, WMT 2019, WMT 2020). Lu et al. (2020) proposed an approach that uses the Dual Bilingual GPT-2 model and the Dual Conditional CrossEntropy Model to evaluate the quality of the parallel corpus. Feng et al. (2020) proposed the LaBSE model, which is a multilingual sentence embedding model trained on 109 languages, including some Indic languages. Herold et al. (2022) mentioned different types of noise that can be injected in a parallel corpus and investigated whether state-of-the-art filtering models are capable of removing all the noise types proposed by Khayrallah and Koehn (2018).

Most recently, Batheja and Bhattacharyya (2022) used a combination of Phrase Pair Injection and LaBSE (Feng et al., 2020) based Corpus Filtering to extract high-quality parallel data from a noisy parallel corpus. In contrast, we use QE-based filtering to extract high-quality data from noisy pseudo-parallel data. We observe that QE quality scores are superior to the LaBSE quality scores.

2.2 Quality Estimation

Quality Estimation (QE) is the task of evaluating the quality of a translation when reference translation is not available. The state-of-the-art MonoTransQuest architecture, proposed by Ranasinghe et al. (2020), builds upon XLM-R, a widely-used pretrained cross-lingual language model known for its ability to generalize to low-resource languages (Conneau et al., 2020). (Kocyigit et al., 2022) proposed a combination of multitask training, data augmentation and contrastive learning to achieve better and more robust QE in a Parallel Corpus Mining setting. The Parallel Corpus Mining task aims to detect the most similar texts in a large multilingual collection and perform sentence alignment. This motivates us to use QE in the Parallel Corpus Filtering task.

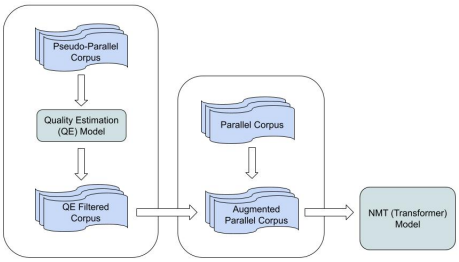

Figure 1: Quality Estimation based Filtering Pipeline

3 Approaches

We first discuss methods to extract good-quality parallel sentences from the pseudo-parallel corpus. Then we discuss a transfer learning-based filtering approach in few-shot settings.

3.1 LaBSE based Filtering

Language Agnostic BERT Sentence Embedding model (Feng et al., 2020) is a multilingual embedding model that supports 109 languages, including some Indic languages. We generate the sentence embeddings for the source and target sides of the pseudo-parallel corpora using the LaBSE [1] model. Then, we compute the cosine similarity between the source and target sentence embeddings. After that, we extract good-quality parallel sentences based on a threshold value of the similarity scores.

3.2 Phrase Pair Injection (PPI) with LaBSE-based Filtering

Batheja and Bhattacharyya (2022) proposed a combination of Phrase Pair Injection (Sen et al., 2021) and LaBSE-based Corpus Filtering to extract high-quality parallel data from a noisy parallel corpus. We train a PBSMT model on the noisy pseudo-parallel corpus using the *Moses* [2] decoder. Then, we extract phrase pairs with the highest translation probability. Finally, we perform LaBSE-based filtering on these phrase pairs to remove poor-quality phrase pairs. We augment these high-quality phrase pairs with LaBSE-filtered parallel sentences.

3.3 Quality Estimation based Filtering

In this approach, we train the MonoTransQuest[3] (Ranasinghe et al., 2020) model and use

[1] https://huggingface.co/sentence-transformers/LaBSE

[2] http://www2.statmt.org/moses/?n=Development.GetStarted

[3] https://github.com/TharinduDR/TransQuest

it to generate the quality scores for the pseudo-parallel corpus of the corresponding language pair. Then, we extract high-quality parallel sentences from the pseudo-parallel corpus using a threshold quality score value.

3.4 Few-shot Quality Estimation

The Quality Estimation task requires human-annotated Direct Assessment scores for the corresponding language pairs. In few-shot settings, we fine-tune a pre-trained QE model for a high-resource language pair on QE data for the corresponding low-resource language pair to obtain a QE model for the low-resource language pair.

4 Mathematical Preliminaries

LaBSE scoring Let $D = \{(x_i, y_i)\}_{i=1}^{N}$ be a pseudo-parallel corpus with N examples, where x_i and y_i represents i^{th} source and target sentence respectively. We first feed all the source sentences present in the pseudo parallel corpus as input to the LaBSE[4] (Feng et al., 2020) model, which is a Dual encoder model with BERT-based encoding modules to obtain source sentence embeddings (S_i). The sentence embeddings are extracted as the l2 normalized **[CLS]** token representations from the last transformer block. Then, we feed all the target sentences as input to the LaBSE model to obtain target sentence embeddings (T_i). We then compute cosine similarity ($score_i$) between the source and the corresponding target sentence embeddings.

$$S_i = LaBSE\,(x_i) \quad (1)$$

$$T_i = LaBSE\,(y_i) \quad (2)$$

$$score_i = cosine\_similarity\,(S_i, T_i) \quad (3)$$

QE scoring We feed "$x_i[SEP]y_i$" as an input to the MonoTransQuest (Ranasinghe et al., 2020) architecture which uses a single XLM-R model. The output of the [CLS] token is used as the input of a softmax layer that predicts the quality score ($score_i$) of the i^{th} sentence pair $< x_i, y_i >$.

$$score_i = MonoTransQuest\,(x_i, y_i) \quad (4)$$

5 Experimental Setup

5.1 Dataset

In all NMT experiments, we use two sets of corpus, namely, Parallel and Pseudo-Parallel corpus. The

Corpus Name	Language Pairs	Sentence Pairs
Parallel Corpus	Hindi-Bengali	3.6M
	English-Marathi	248K
	Chinese-English	62K
Pseudo-Parallel Corpus	Hindi-Bengali	6.3M
	English-Marathi	3.28M
	Chinese-English	24.7M

Table 1: Dataset Statistics of Parallel and Pseudo-Parallel Corpus for the task of Neural Machine Translation

Language	train	dev	test
English-Marathi	26,000	1,000	1,000
Chinese-English	7,000	1,000	1,000
Hindi-Bengali	440	50	10

Table 2: Dataset Statistics of human-annotated z-standardized Domain Adaptation (DA) scores for the task of Quality Estimation

Parallel corpus consists of high-quality sentence pairs, while the **Pseudo-Parallel** corpus contains sentence pairs of varying quality.

The En-Mr Parallel Corpus consists of the ILCI phase 1, Bible, PIB and PM-India corpus (Jha, 2010; Christos Christodouloupoulos, 2015; Haddow and Kirefu, 2020). The Zh-En Parallel corpus consists of ParaMed[5] corpus. The Hi-Bn Parallel corpus is obtained from the OPUS[6] corpus repository. The En-Mr and Zh-En Pseudo-Parallel Corpus consist of the Samanantar (Ramesh et al., 2021) and WMT18 Zh-En[7] corpus, respectively. The Hi-Bn Pseudo-Parallel Corpus consists of Samanantar and Tatoeba (Tiedemann, 2020) corpus. The detailed data statistics are mentioned in table 1.

In QE experiments, we create a small corpus (500 instances) for Hindi-Bengali language pair that consists of human-annotated Domain Adaptation scores for each sentence pair annotated by three annotators. The pairwise Pearson Correlation between the three annotators of Hindi-Bengali QE is **0.68**, **0.61** and **0.67**. This indicates a good agreement between the three annotators. Please refer to **Appendix** A.3 for further annotation details. We use the QE data provided by Ranasinghe et al. (2020) and Deoghare and Bhattacharyya (2022) for the Zh-En and En-Mr language pairs, respectively. The detailed QE data statistics are mentioned in

[4] https://huggingface.co/sentence-transformers/LaBSE

[5] https://github.com/boxiangliu/ParaMed

[6] https://opus.nlpl.eu/

[7] http://data.statmt.org/wmt18/translation-task/preprocessed/zh-en/

Technique	# Sentence Pairs	En→Mr	Mr→En
QE based Filtering	2.61M	9.4	**17.7**
LaBSE + PPI-LaBSE based Filtering (Batheja and Bhattacharyya, 2022)	4.09M	**9.9**	17.0
LaBSE based Filtering	2.85M	8.8	16.7
Baseline	3.53M	8.8	15.9

Table 3: BLEU scores of En→Mr and Mr→En NMT models on FLORES101 test data. Here, we establish the efficacy of QE-based filtering in extracting a high-quality parallel corpus from En-Mr pseudo-parallel corpus. For actual instances of translations please refer to Appendix A.1.

Technique	# Sentence Pairs	Zh→En
QE based Filtering	15.09M	8.7
LaBSE + PPI-LaBSE based Filtering (Batheja and Bhattacharyya, 2022)	15.59M	8.47
LaBSE based Filtering	15.57M	8.29
Baseline	24.8M	7.85

Table 4: BLEU scores of Zh→En NMT models on FLORES101 test data. Here, we establish the efficacy of QE-based filtering in extracting a high-quality parallel corpus from Zh→En pseudo-parallel corpus. For actual instances of translations please refer to Appendix A.1

table 2.

For evaluation, we use the FLORES 101 test set which contains 1,012 sentence pairs for each language pair.

5.2 Models

We use MonoTransQuest model architecture to train the QE models. We use the Indic NLP library for preprocessing the Indic language data and Moses for preprocessing the English language data. For Indic languages, we normalize and tokenize the data. For English, we lowercase and tokenize the data. We use a Transformer based architecture provided by OpenNMT-py library to train the NMT models for all our experiments. The optimizer used was adam with betas (0.9, 0.98). The initial learning rate used was 5e-4 with the inverse square root learning rate scheduler. We use 8000 warmup updates. The dropout probability value used was 0.1 and the criterion used was label smoothed cross entropy with label smoothing of 0.1. We use a batch size of 4096 tokens. All the models were trained for 200,000 training steps. We use MonoTransquest[8] model to train the sentence-level QE model. We start with a learning rate of 2e-5 and use 5% of training data for warm-up. We use early patience over ten steps. We use a batch size of eight. The model architecture is mentioned in **Appendix** A.2.

Baseline We train the baseline NMT models on the whole pseudo-parallel corpus augmented with the parallel corpus for the corresponding language pairs.

LaBSE based Filtering In this model, we use the LaBSE filtering with threshold 0.8 to extract good quality parallel sentences from the En-Mr, Hi-Bn and Zh-En pseudo-parallel corpus. Then we augment the parallel corpus with the LaBSE-filtered parallel sentences and train the respective NMT models.

LaBSE + PPI-LaBSE based Filtering We extract LaBSE Filtered parallel sentences and phrases from the pseudo-parallel corpus and augment them with the parallel corpora to train the respective NMT models.

Our Model, QE based Filtering We train the sentence-level QE model from scratch for En-Mr and Zh-En language pairs using their respective training data, Table 2. We use the English-Marathi pre-trained QE model for the Hi-Bn language pair and finetune it on Hi-Bn training data, Table 2. We compute quality scores for the noisy pseudo-parallel corpora using the trained QE models. Then, we extract high-quality sentence pairs from the pseudo-parallel corpus using the threshold values of -0.5, -0.4, and 0 for En-Mr, Zh-En, and Hi-Bn language pairs, respectively. We augment the extracted high-quality sentence pairs with the parallel corpus and train the respective NMT models.

[8] https://github.com/TharinduDR/TransQuest

Technique	# Sentence Pairs	Hi→Bn	Bn→Hi
QE based Filtering	7.77M	13.28	21.06
LaBSE + PPI-LaBSE based Filtering (Batheja and Bhattacharyya, 2022)	8.73M	13.24	20.51
LaBSE based Filtering	7.77M	13.23	20.48
Baseline	10M	12.91	20.43

Table 5: BLEU scores of Hi→Bn and Bn→En NMT models on FLORES101 test data. Here, we establish the efficacy of few-shot QE-based filtering using a pre-trained En-Mr model fine-tuned on Hi-Bn QE data to extract a high-quality parallel corpus from the Hi-Bn pseudo-parallel corpus. For actual instances of translations please refer to Appendix A.1

6 Results and Analysis

We evaluate our NMT models using BLEU (Papineni et al., 2002). We use *sacrebleu* (Post, 2018) python library to calculate the BLEU scores. Table 5 shows that QE based filtering model outperforms all other models for Hi-Bn, En-Mr and Zh-En language pairs. The **QE based Filtering** model improves the MT system's performance by **0.85**, **0.6**, **1.8**, **0.37** and **0.63** BLEU points over the **baseline** model for Zh→En, En→Mr, Mr→En,Hi→Bn and Bn→Hi, respectively. It also outperforms **LaBSE + PPI-LaBSE based Filtering** model by up to **0.7** BLEU points for Zh-En, En-Mr and Hi-Bn language pairs. The LaBSE + PPI-LaBSE based Filtering model performs better than QE based Filtering model for En→Mr language direction. The LaBSE + PPI-LaBSE model, which is trained on nearly twice the amount of training data compared to the QE-based filtering model, can be a contributing factor to its better performance in En→Mr.

The improvement in the performance of the Bn→Hi QE-based filtered MT system is comparable to the En→Mr and Zh→En QE-based filtered MT model. The Hi-Bn QE model is trained with only 500 training instances transfer learned from En-Mr trained QE models. This demonstrates the promise of the few-shot QE technique to generate training data for MT.

Technique	En-Mr	Hi-Bn	Zh-En
LaBSE	0.44	0.51	0.2
QE	**0.52**	**0.53**	**0.4**

Table 6: Pearson Correlation between human annotated quality scores and quality scores computed using LaBSE and QE

We compute Pearson Correlation between human annotated quality scores and quality scores computed using LaBSE and QE, shown in Table 6. The QE quality scores have a higher correlation with human annotated quality scores, compared to LaBSE quality scores for all 3 language pairs. Table 7 shows the Pearson Correlation between LaBSE and QE quality scores for all 3 language pairs. We observe that the LaBSE quality score has a low correlation with the QE quality score and the QE quality score has a high correlation with the human annotated quality score. This establishes the superiority of QE over the LaBSE quality score.

En-Mr	Hi-Bn	Zh-En
0.5	0.37	0.28

Table 7: Pearson Correlation between LaBSE and QE quality scores computed on the pseudo-parallel corpus for En-Mr, Hi-Bn and Zh-En language pairs respectively

7 Conclusion and Future Work

We introduced a simple Quality Estimation based corpus filtering approach to extract high-quality parallel data from the noisy pseudo-parallel corpora. The takeaway from our work is that sentence-level QE-based filtering performs better than LaBSE-based filtering and helps improve the performance of NMT systems. We also show that few-shot QE models trained using a transfer learning-based approach can be used to extract good-quality parallel corpus from the pseudo-parallel corpus. Only $1/40^{th}$ of the normal data requirement (7K-25K) of QE training data achieves comparable performance for the Hindi-Bengali language pair. We also show that the QE quality score is superior to the LaBSE quality score.

In the future, we plan to use the proposed corpus filtering technique for other language pairs. This will provide us with a general overview of how this filtering technique performs for multiple languages.

Acknowledgements

We would like to thank the anonymous reviewers for their insightful feedback. We also express our gratitude towards Shivam Mhaskar, Sourabh Deoghare and other members of the Machine Translation group at CFILT, IIT Bombay, for their interesting and insightful comments.

Limitations

Although our primary effort in this work was to extract as much parallel corpora as possible, the improvement in the performance has been found to be only marginal. The LaBSE and QE-based filtering experiments involve a hyper-parameter called "threshold quality score." To achieve optimal results, we conduct experiments with different values of this hyper-parameter. The proposed few-shot transfer learning technique requires a small amount of data that needs to be annotated by multiple annotators.

Ethics Statement

The aim of our work is to extract high-quality parallel corpus from a noisy pseudo-parallel corpus. The datasets that we used in this work are publicly available and we have cited the sources of all the datasets that we have used. Publicly available datasets can contain biased sentences. We have also created a dataset for Hindi-Bengali few-shot QE. We briefly discuss the annotation guideline given to the annotators for the task in the **Appendix** A.3.

References

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. 3rd International Conference on Learning Representations, ICLR 2015 ; Conference date: 07-05-2015 Through 09-05-2015.

Akshay Batheja and Pushpak Bhattacharyya. 2022. Improving machine translation with phrase pair injection and corpus filtering.

Mark Steedman. Christos Christodouloupoulos. 2015. A massively parallel corpus: the bible in 100 languages. Language resources and evalua tion.

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.

Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A survey of multilingual neural machine translation. *ACM Comput. Surv.*, 53(5).

Sourabh Deoghare and Pushpak Bhattacharyya. 2022. Iit bombay's wmt22 automatic post-editing shared task submission. In *Proceedings of the Seventh Conference on Machine Translation*, pages 682–688, Abu Dhabi. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding.

Barry Haddow and Faheem Kirefu. 2020. Pmindia – a collection of parallel corpora of languages of india.

Christian Herold, Jan Rosendahl, Joris Vanvinckenroye, and Hermann Ney. 2022. Detecting various types of noise for neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2542–2551, Dublin, Ireland. Association for Computational Linguistics.

Girish Nath Jha. 2010. The TDIL program and the Indian langauge corpora intitiative (ILCI). In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA).

Huda Khayrallah and Philipp Koehn. 2018. On the impact of various types of noise on neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 74–83, Melbourne, Australia. Association for Computational Linguistics.

Muhammed Kocyigit, Jiho Lee, and Derry Wijaya. 2022. Better quality estimation for low resource corpus mining. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 533–543, Dublin, Ireland. Association for Computational Linguistics.

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)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jun Lu, Xin Ge, Yangbin Shi, and Yuqi Zhang. 2020. Alibaba submission to the WMT20 parallel corpus filtering task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 979–984, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2021. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages.

Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. 2020. TransQuest: Translation quality estimation with cross-lingual transformers. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5070–5081, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Sukanta Sen, Mohammed Hasanuzzaman, Asif Ekbal, Pushpak Bhattacharyya, and Andy Way. 2021. Neural machine translation of low-resource languages using smt phrase pair injection. *Natural Language Engineering*, 27(3):271–292.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Jörg Tiedemann. 2020. The Tatoeba Translation Challenge – Realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

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 *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

A Appendix

A.1 Instances of Translations (Referred from Table 5)

The instances of translations for all 3 language pairs are provided in Table 6, 7, 8, 9 and 10.

A.2 Model Architecture (Referred from section 5.2)

We use a Transformer based architecture to train English-Marathi, Hindi-Bengali and Chinese-English NMT models for all our experiments. The encoder of the Transformer consists of 6 encoder layers and 8 encoder attention heads. The encoder uses embeddings of dimension 512. The decoder of the Transformer also consists of 6 decoder layers and 8 decoder attention heads. We use MonoTransQuest architecture to train English-Marathi, Hindi-Bengali and Chinese-English QE models for all our experiments. We use a single Nvidia A100 GPU with 40 GB memory to train our NMT and QE models

A.3 Annotation Details (Referred from section 5)

A.3.1 Annotator Demographic

For the Direct Assessment score annotation, we requested three native language speakers of Bengali who are well-versed in Hindi and have attended their graduate degrees in the Hindi language. They were aged between 25 to 42 and were paid for the time they spent on annotations.

A.3.2 Guidelines

The guidelines provided to the annotators for the Quality Estimation task are shown in Figure 2.

A.3.3 Dataset

We create Hi-Bn QE data for our few-shot settings. We use 500 high-quality Hindi sentences from IIT Bombay English-Hindi parallel corpus (Kunchukuttan et al., 2018). We use the Hindi-Bengali NMT model to generate translations for the 500 Hindi sentences. We provide this Hindi-Bengali parallel data to the annotators for the Direct Assessment Task. The Direct Assessment tasks require the annotators to score the MT translations as per the guidelines provided, Figure A.3.

English (Source)	though it is widely used, especially among non-romani, the word "gypsy" is often considered offensive because of its associations with negative stereotypes and inaccurate perceptions of romani people.
Marathi (Reference)	याचा वापर विशेषतः रोमानी नसलेल्यांमध्ये व्यापकपणे केला जात असला तरी, ""जिप्सी"" हा शब्द बरेचदा नकारत्मक रूढी आणि रोमानी लोकांबद्दलच्या चुकीच्या समजुतींशी संबंधित असल्यामुळे आक्षेपार्ह समजला जातो.ल.
Marathi (Baseline)	याचा मोठ्या प्रमाणात वापर केला जातो, विशेषतः नॉन-रोमानीमध्ये "जिप्सी" हा शब्द सहसा त्याच्या नकारात्मक क्रिएटिक्समुळे आणि रोमानी लोकांच्या असत्य विचारांमुळे होतो. आहे.
Marathi (QE based Filtering)	याचा मोठ्या प्रमाणात वापर केला जातो, विशेषतः नॉन-रोमनीमध्ये, "जिप्सी" हा शब्द सहसा त्यांच्या नात्यांमुळे आणि रोमी लोकांच्या अअचूक शब्दांमुळे घातक मानला जातो.

Table 6: Examples of NMT model output for En→Mr

Marathi (Source)	परंतु, हे सत्य नाही. या दस्तऐवजाच्या पाठीमागे काहीतरी लिहिले असले तरी, तो एक खजिन्याचा नकाशा नाही.
English (Reference)	however, that is not true. although there is something written on the back of the document, it is not a treasure map.
English (Baseline)	but this is not true.
English (QE based Filtering)	but, that is not true, though something is written behind this document, it is not a map of a treasure.

Table 7: Examples of NMT model output for Mr→En

Hindi (Source)	प्राचीन चीन में अलग - अलग समय अवधि दिखाने का एक अनूठा तरीका था ; चीन का प्रत्येक चरण या प्रत्येक परिवार जो सत्ता में था , एक ख़ास राजवंश था .
Bengali (Reference)	প্রাচীন চীনে বিভিন্ন সময়কাল দেখানোর একটি অনন্য উপায় ছিল ; চীন বা ক্ষমতায় থাকা প্রতিটি পরিবারের প্রতিটি পর্যায় ছিল একটি স্বতন্ত্র রাজবংশ ।
Bengali (Baseline)	প্রাচীন চীন বিভিন্ন সময় একটি অনন্য উপায় ছিল; চীন প্রতিটি পর্যায়ে বা প্রতিটি পরিবার যারা ক্ষমতায় ছিল, একটি বিশেষ রাজবংশ ছিল.
Bengali (QE based Filtering)	প্রাচীন চীনে বিভিন্ন সময়কাল দেখানোর একটি অনন্য উপায় ছিল; চীন প্রতিটি পর্যায়ে বা প্রতিটি পরিবার যারা ক্ষমতায় ছিল, একটি বিশেষ রাজবংশ ছিল.

Table 8: Examples of NMT model output for Hi→Bn

Bengali (Source)	আন্তর্জাতিক নিষেধাজ্ঞার মানে হলো নতুন বিমান ক্রয় করা যাবেনা ।
Hindi (Reference)	अंतर्राष्ट्रीय प्रतिबंधों का मतलब है कि नए विमान नहीं खरीदे जा सकते .
Hindi (Baseline)	अंतर्राष्ट्रीय प्रतिबंधों का मतलब है कि एक नया विमान नहीं खरीदा जा सकता है।
Hindi (QE based Filtering)	अंतरराष्ट्रीय प्रतिबंधों का मतलब है कि नए विमान नहीं खरीदे जा सकते हैं।

Table 9: Examples of NMT model output for Bn→Hi

Chinese (Source)	科学家表示，这次碰撞引起的爆炸规模巨大。
English (Reference)	scientists say the explosion caused by the collision was massive.
English (Baseline)	the science said that the explosion had caused a great deal .
English (QE based Filtering)	scientists say the explosion caused by the crash is huge .

Table 10: Examples of NMT model output for Zh→En

Overall Score	Translation conveys source meaning?	How much translation conveys to source?
1 - 10	Completely inaccurate.	The MT output is unintelligible. Studying the meaning of the sentence is hopeless; even allowing for context, one feels that guessing would be too unreliable. • The translation is incomprehensible, and machine translated output contains a mix of languages/dialects [which are not Bengali] (Adequacy) • None of the keywords are translated in the target language. (Adequacy) • There are major grammatical errors and typos (Fluency)
11 - 30	Inaccurate but contains some keywords.	• Incomprehensible translation (Fluency) • Translation contains some keywords but not all (Adequacy) • There are numerous grammatical errors and typos.
31 -50	Partially. Target reflects partially the source.	The general idea of the MT output is intelligible only after considerable study. • Translation is only partially understandable, and the overall meaning is not conveyed. (Adequacy and Fluency) • Translation contains some keywords (Adequacy) • There are many grammatical errors and typos (Fluency)
51 - 70	Yes. Target reflects the overall meaning.	The MT output is generally clear and intelligible. Despite some inaccuracies or infelicities of the sentence, one can understand (almost) immediately what it means. • Translation is understandable and reflects the source meaning. (Fluency) • Translation contains most keywords (Adequacy) • Only minor grammar errors (Fluency)
71 - 90	Yes. Target reflects the source meaning without errors.	The MT output is perfectly clear and intelligible. It is grammatical and reads like ordinary text. • Translation is very closed to the source meaning (Fluency) • Translation contains all keywords (Adequacy) • No errors but there are better word choices in the target language. (Adequacy)
91 - 100	Yes. Target reflects source meaning without errors.	• Perfect translation (Adequacy and Fluency) • Accurately reflects the meaning of the source (Fluency) • No errors

Figure 2: Evaluation Guidelines for the Hindi-Bengali Direct Assessment (for Quality Estimation) Task.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section: Limitations

☒ A2. Did you discuss any potential risks of your work?
Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section: Abstract and Section: 1 (Introduction)

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 5.2: Models; Section 5.1 Dataset

☑ B1. Did you cite the creators of artifacts you used?
Section 2: Related Work; Section 5.1: Dataset

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Left blank.

☒ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Left blank.

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 5: Experimental Setup; Appendix

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 5.2: Models

C ☑ Did you run computational experiments?

Section 5: Experimental Setup; Appendix A.2: Model Architecture; A.3: Training Details

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix A.2: Model Architecture; A.3: Training Details

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 5: Experimental Setup

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 6: Results and Analysis

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 6: Results and Analysis

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Section 5.1: Dataset

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Section: Appendix

☒ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Left blank.

☒ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Left blank.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Left blank.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Section: Appendix