Shiksha: A Technical Domain focused Translation Dataset and Model for Indian Languages

Anonymous ACL submission

Abstract

Neural Machine Translation (NMT) models are typically trained on datasets with limited exposure to **Scientific, Technical and Educational domains**. Translation models thus, in general, struggle with tasks that involve scientific understanding or technical jargon. Their performance is found to be even worse for **low-resource Indian languages**. Finding a translation dataset that tends to these domains in particular, poses a difficult challenge. In this paper, we address this by creating a multilingual parallel corpus containing **more than 2.8 million rows** of *English-to-Indic* and *Indic-to-Indic* high-quality translation pairs across **8 Indian languages**. We achieve this by bitext mining **human-translated transcriptions of NPTEL[1] video lectures**. We also finetune and evaluate NMT models using this corpus and surpass all other publicly available models at in-domain tasks. We also manage to improve the baseline by over 2 BLEU for these Indian languages on average, thus demonstrating the potential for generalizing to out-of-domain translation tasks as well. We are pleased to release the corresponding models and dataset, accessible via this link: https://huggingface.co/anon-auth.

1 Introduction

NPTEL (*National Programme on Technology Enhanced Learning*) has long been a valuable resource for free on-demand higher-educational content across a diverse range of specialized disciplines. Over the past two decades since its inception, NPTEL has curated an extensive library of over 56,000 hours of video lectures, all made publicly available along with their audio transcriptions in an easily accessible manner. In response to the growing number of Indian students, NPTEL has taken steps to support Indian language transcriptions for more than 12,000 hours of video content. These captions are primarily translations of the original English transcriptions, carefully crafted by subject-matter experts. This multi-year endeavor has led to the creation of a high-quality parallel textual resource spanning multiple Indian languages, covering various fields in the Scientific, Engineering, and Humanities domains. Our research leverages this rich data resource to develop competitive Machine Translation (MT) models specifically tailored for Indian languages. Additionally, we investigate how models fine-tuned on this data can assist human translators and help accelerate the mission of providing accurate Indic subtitles for all NPTEL video lectures. This effort aims to benefit a large audience of Indian students struggling with the lack of university-level educational content in their native tongues.

[1] https://nptel.ac.in

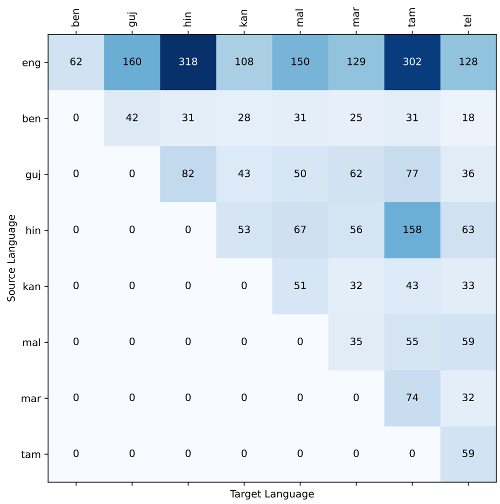


Figure 1: Translation Pair Counts (in thousands)

2 Where does Present-day MT fail?

Let's quickly look at how Machine Translation models in use widely today perform on Technical-domain tasks and instances in which they fail.

English source text	Our Model	Google Translate	IndicTrans2
What is the stress on this spring?†	इस स्प्रिंग पर क्या तनाव है?	इस वसंत पर तनाव क्या है?	इस वसंत पर क्या दबाव है?
I will go home this spring.‡	मैं इस वसंत में घर जाऊंगा।	मैं इस बसंत में घर जाऊंगा।	मैं इस वसंत में घर जाऊंगा।
I want to learn the rust language.†	मैं रस्ट भाषा सीखना चाहता हूं।	मैं जंग भाषा सीखना चाहता हूँ.	मैं जंग की भाषा सीखना चाहता हूँ।
My bicycle is covered with rust.‡	मेरी साइकिल जंग से ढकी हुई है।	मेरी साइकिल जंग से ढकी हुई है।	मेरी साइकिल जंग से ढकी हुई है।

Table 1: Example translations from English to Hindi in the Scientific/Technical domain.
Sentences marked as † are in-domain, while ‡ are out-of-domain.
The words in blue are terms with multiple meanings, that tend to get translated incorrectly.
The words in green represent the correct, expected translation by the model for the blue word in the given context.
The words in red represent incorrect translations.

Consider the text "I want to learn the rust language." Here we are talking about the programming language Rust and not the chemical phenomena. From Table 1 we can see that both Google Translate [2] and IndicTrans2 get their Hindi translations wrong. Not only that, if we backtranslate their results to English we get the sentence "I want to learn the language of war," which is very far from what we originally meant. This happens in this case because the Hind word for "rust" has two meanings, the phenomena of rust and also war. Thus, in such situations it is very important for the translation model to understand the context well as the meaning of the sentence can completely change with the wrong choice of a word. So, we can see that current models are prone to making such mistakes for tasks in the technical and scientific domains. Our paper hopes to help alleviate these shortcomings.

3 The Dataset

3.1 First, the Source

The initial step in creating any dataset involves obtaining the raw data. We chose against scraping subtitles from YouTube videos, and instead obtained the raw data by contacting NPTEL. They provided us with a document containing a list of 10,669 videos along with links to their corresponding transcriptions and related metadata. These transcriptions were bilingual documents, featuring alternating paragraphs of English and Indic text, interspersed with reference timestamps and video snapshots. Refer to Appendix A for a sample page.

[2] https://translate.google.com/

3.2 Data Cleaning and Extraction

Given the unusual format of these documents, we wrote a Python script to extract the meaningful text data from it while avoiding timestamp references. In this script, we first pull out the text from these documents and then use regex patterns to filter the timestamps. We then used simple paragraph segmenting tools from *nltk* (Bird and Klein, 2009) and *indic-nlp* (Kunchukuttan, 2020) libraries to identify and separate English and Indic sentences. With this, the lectures are now decomposed into parallel documents of text stored in separate files to create a massive bitext corpora.

3.3 Bitext Mining

With this parallel corpora in place, we begin the most crucial part: Bitext mining. Our objective is to find as many sentence-pairs as we can from the source data while still having high confidence in their translation accuracy. Luckily, Sentence-Alignment is a well studied problem dating as far back as 1991 (Brown et al.).

Recent work like Vecalign (Thompson and Koehn, 2019) has focused on using multi-lingual embedding models to find pairs based on vector similarity of sentence embeddings. These have been shown to achieve state-of-the-art performance, significantly surpassing previous approaches. In our work, we use SentAlign (Steingrimsson et al., 2023) which employs LABSE (Feng et al., 2022) along with optimized alignment algorithms to mine parallel documents with high confidence and efficiency. With this we are also able to find 1-n and n-1 sentence matches.

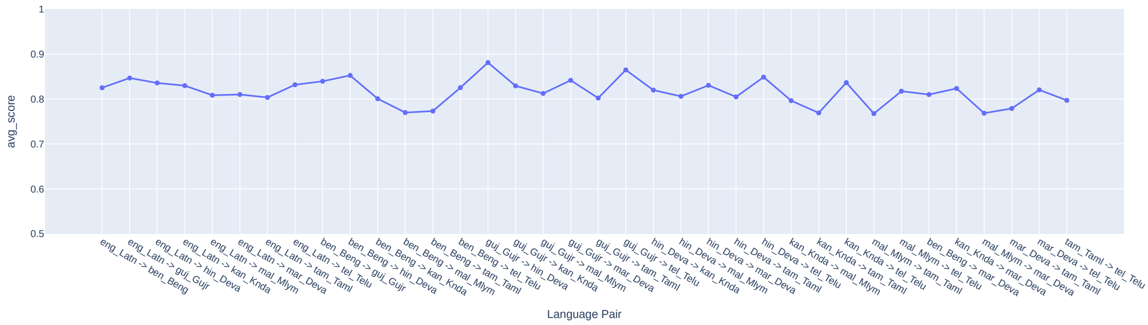


Figure 2: Average LABSE score across language pairs

3.4 Data Collation

Each lecture now has its own documents with all its sentences aligned into bilingual sentence-pairs. We collect these pairs and combine them, along with their lecture metadata, into a massive translation dataset. After post-processing with deduplication, we arrive at a corpus of roughly 2.8 million sentence-pairs.

3.5 Data Analysis

To understand the quality and quantity of this data, we must first thoroughly analyse it. Our dataset has 8 English-Indic and 28 Indic-Indic language pairs. This means that there exists at least one common set of lectures among each language-pair, providing us with inter-Indic alignments for all languages covered in this dataset. We find 48.6% of these to be English-to-Indic language pairs. This is the direct result of English lectures having been translated into multiple languages, albeit with arbitrary combinations, giving us a robust Indic-to-Indic data subset.

For assessing the alignment quality of our translation pairs, we look at the average LABSE similarity scores as the primary measure. The plot of this metric (Figure 2) demonstrates a strong consistency in scores across all languages despite differences in the quantity of the mined sentence-pairs. These data points are also seen to be tending towards 0.8 and are never below 0.75, validating our confidence in the quality of the source data and the accuracy of our alignments.

4 The Model

The value of a dataset can only be best quantified when a model has been trained with it, evaluated against others and the results analyzed. We wish to fine-tune and evaluate a powerful MT base model to test the hypothesis that our dataset can help improve the performance of translation tasks in the Technical domain. We will test if existing models can be improved meaningfully by being fine-tuned on our dataset.

4.1 Baseline Model selection

When it comes to choosing a strong multi-lingual model that is, or at least close to, the state-of-the-art, we find that our options are limited. IndicTrans2 (IT2) could be a good choice, except that it provides different models for English-Indic, Indic-English and Indic-Indic directions. We eliminate this option since we hope to leverage transfer learning by training just one model in all the 36 language-pair combinations that our dataset supports. That leaves us with only two possible candidates: NLLB-200 (Team et al., 2022) and MADLAD-400 (Kudugunta et al., 2023). These are both massively multilingual Transformer models that promise to be the right baseline for our study. We choose NLLB on the basis of its superior evaluation scores for Indian languages on the Flores-200 (Goyal et al., 2022) benchmark, according to the results published in the MADLAD paper.

4.2 Training

NLLB-200 models are available in a wide variety of sizes ranging from a 600M parameter distilled model to a massive 54B parameter Mixture-of-Experts model. For our experiments, we decide to choose the 3.3B parameter version as a sweet spot between performance quality and compute requirements.

Still, even with this relatively smaller NLLB 3.3B model, running a Full Fine-Tuning (FFT) setup can turn out to be a very compute intensive en-

Our testset	
Models	**en-in**
NLLB	30.73 / 57.62
LoRA FT	**48.98 / 71.99**
IT2	39.66 / 66.49

Flores+	
Models	**en-in**
NLLB	19.73 / 54.27
LoRA FT	22.04 / 57.33
IT2	**24.08 / 59.45**

Table 2: Results are in the form <bleu>/<chrf++>.
These scores represent the average of all 8 languages covered by the dataset.
All models were evaluated without using beam-search or sampling.

deavour. The amount of time required to effectively train our model will also be significant. In our case we wish to execute a number of experiments with different approaches in hopes to achieve the best results. FFT thus would not be a feasible approach. Instead, we decide to utilize a Parameter-Efficient Fine Tuning (PEFT) method known as Low-Rank Adaptation (LoRA) (Hu et al., 2022) to train our model.

We primarily trained three models using three different approaches. All of them were done using LoRA with NLLB 3.3B. These approaches included: 1) training a model purely on our dataset in one direction, 2) training using Curriculum Learning (CL) (Bengio et al., 2009) with a cleaned subset of the BPCC corpus (Gala et al., 2023) with our 8 Indian languages, comprising of 4 million rows, before introducing our dataset, 3) training on a massive 12 million samples which included the cleaned BPCC corpus and our dataset in both directions. All our models were trained on a node of 8 NVIDIA A100 40GB GPUs. Evaluation results for all the three models were found to be similar, with our 3rd approach performing slightly better. The hyperparameters and detailed results for all three are available in Appendix B.

4.3 Evaluation

For evaluation we compare our third model, trained on 12 million rows, with the baseline NLLB model and the 1B parameter IndicTrans2 model. For an in-domain test, we used the top one thousand rows (by LABSE score) of our held-out test set for each language. Our model outperforms the rest on our test set and demonstrates the efficacy of our model at translations involving the technical domain. We further test our models on the Flores+ [3] (Previously Flores-200) devtest set. We find that our model is also able to generalize well, as seen from the improvements on the baseline scores. Our results manage to come closer to IndicTrans2, which was trained on a corpus far larger than ours. These scores are depicted in Table 2 above. Language-wise comparison of evaluation scores are also available in Appendix B.

[3] https://github.com/openlanguagedata/flores

5 Translingua

This research goes beyond just experiments. Our models are now built into a tool called Translingua, that is being widely used by human annotators across India to translate NPTEL lecture transcripts into more languages than ever before, with far better speed and accuracy. A screenshot of this tool along with feedback of our respondents on translation quality is available in Appendix C.

Conclusion

In this paper, we introduced Shiksha, a novel translation dataset and model tailored for Indian languages, with a particular focus on the Scientific, Technical, and Educational domains. We created a robust multilingual parallel corpus consisting of over 2.8 million high-quality translation pairs across 8 Indian languages. Our approach involved meticulous data extraction, cleaning, and bitext mining to ensure the accuracy and relevance of the dataset. We also fine-tuned state-of-the-art baseline NMT models using this dataset and demonstrated significant performance improvements in not only in-domain, but also out-of-domain translation tasks.

With this paper, we wish to encourage the importance of domain-specific datasets in advancing NMT capabilities. We believe that our dataset and models will serve as valuable resources for the community and foster further research in multilingual NMT.

Limitations

Despite the promising results of our dataset and model, there are some limitations that need to be acknowledged:

- The dataset is heavily skewed towards scientific, technical, and educational domains, sourced primarily from NPTEL video lectures. This can lead to degradation in translation quality for general tasks in unexpected ways that standard benchmarks may not catch. We recommend supplementing our dataset with additional diverse and balanced sources covering a wide range of domains, including everyday conversational language, literature, social media, and news articles. This will help ensure a more stable training and evaluation process, ultimately enhancing the translation system's robustness and accuracy across different contexts.

- We have not meaningfully tested our model's performance on Indic-English or Indic-Indic directions as our research was focused primarily on translating out of English. Our models thus may not perform well on those language directions.

- The quality of our translation dataset and models is heavily dependent upon the accuracy of the original NPTEL transcriptions. Any errors or inconsistencies in them are propagated into our dataset, potentially affecting the training and evaluation of the translation models. Further human evaluation might be needed to verify the quality of these translations.

References

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 41–48, New York, NY, USA. Association for Computing Machinery.

Edward Loper Bird, Steven and Ewan Klein. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc.

Peter F. Brown, Jennifer C. Lai, and Robert L. Mercer. 1991. Aligning sentences in parallel corpora. In *29th Annual Meeting of the Association for Computational Linguistics*, pages 169–176, Berkeley, California, USA. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Jay Gala, Pranjal A Chitale, A K Raghavan, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar M, Janki Atul Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Transactions on Machine Learning Research*.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A. Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. Madlad-400: A multilingual and document-level large audited dataset. *Preprint*, arXiv:2309.04662.

Anoop Kunchukuttan. 2020. The IndicNLP Library. `https://github.com/anoopkunchukuttan/indic_nlp_library/blob/master/docs/indicnlp.pdf`.

Steinthor Steingrimsson, Hrafn Loftsson, and Andy Way. 2023. SentAlign: Accurate and scalable sentence alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 256–263, Singapore. Association for Computational Linguistics.

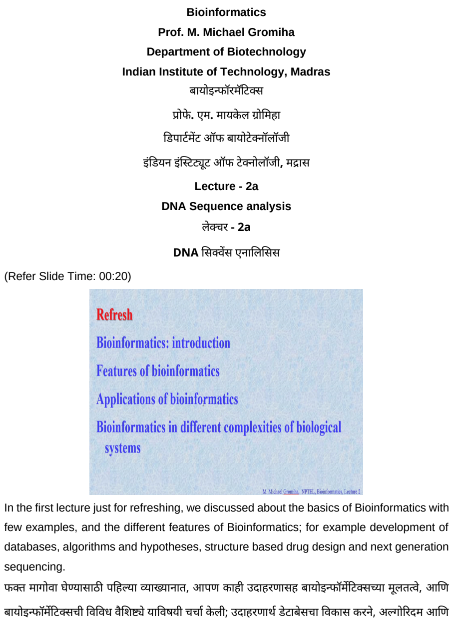
NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff

Wang. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Brian Thompson and Philipp Koehn. 2019. Vecalign: Improved sentence alignment in linear time and space. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1342–1348, Hong Kong, China. Association for Computational Linguistics.

Appendices

A Source Document



Bioinformatics
Prof. M. Michael Gromiha
Department of Biotechnology
Indian Institute of Technology, Madras
बायोइन्फॉर्मेटिक्स
प्रोफे. एम. मायकेल ग्रोमिहा
डिपार्टमेंट ऑफ बायोटेक्नॉलॉजी
इंडियन इंस्टिट्यूट ऑफ टेक्नोलॉजी, मद्रास
Lecture - 2a
DNA Sequence analysis
लेक्चर - 2a
DNA सिक्वेंस एनालिसिस

(Refer Slide Time: 00:20)

In the first lecture just for refreshing, we discussed about the basics of Bioinformatics with few examples, and the different features of Bioinformatics; for example development of databases, algorithms and hypotheses, structure based drug design and next generation sequencing.
फक्त मागोवा घेण्यासाठी पहिल्या व्याख्यानात, आपण काही उदाहरणासह बायोइन्फॉर्मेटिक्सच्या मूलतत्वे, आणि बायोइन्फॉर्मेटिक्सची विविध वैशिष्ट्ये याविषयी चर्चा केली; उदाहरणार्थ डेटाबेसचा विकास करने, अल्गोरिदम आणि

Figure 3: A sample page from a bilingual document

B Model Hyperparameters and Results

Parameter	Setting
peft-type	LORA
rank	256
lora alpha	256
lora dropout	0.1
rslora	True
target modules	all-linear
learning rate	4e-5
optimizer	adafactor
data-type	BF-16
epochs	1

Table 3: Hyperparameters for our 3rd approach.
First approach was trained for 10 epochs and second for 4 epochs seperately

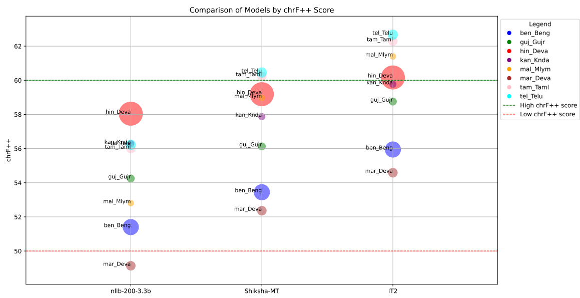


Figure 4: Chrf++ comparison between NLLB, IT2 and our model across all Indian languages.
The size of the bubble represents the population of the speakers.

C Translingua

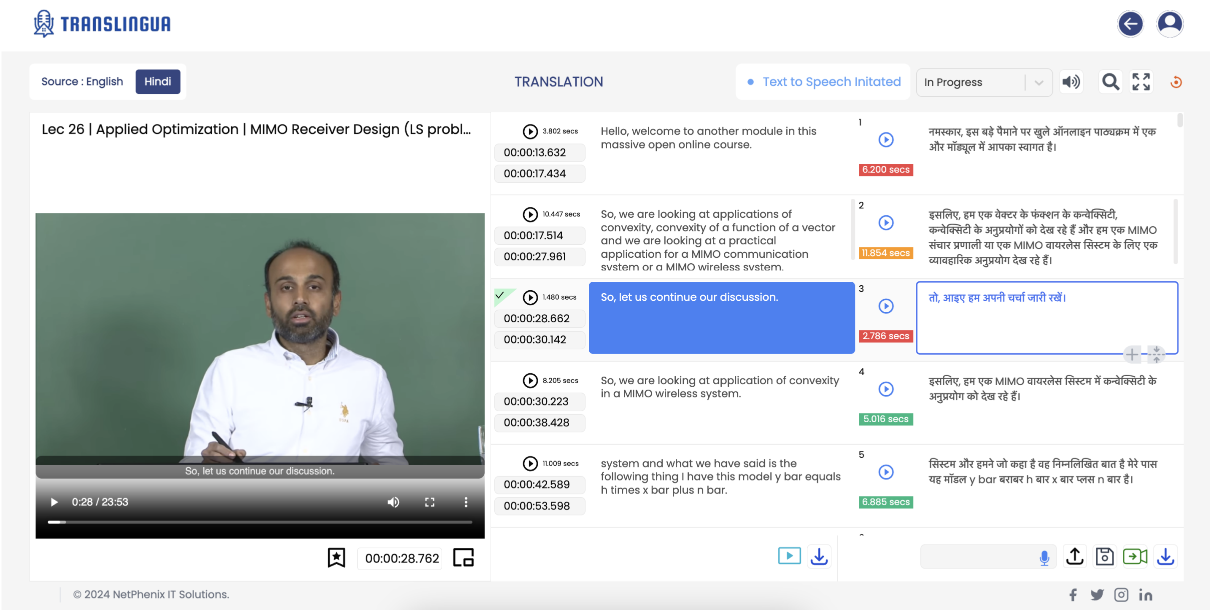


Figure 5: A screenshot from the Translingua tool

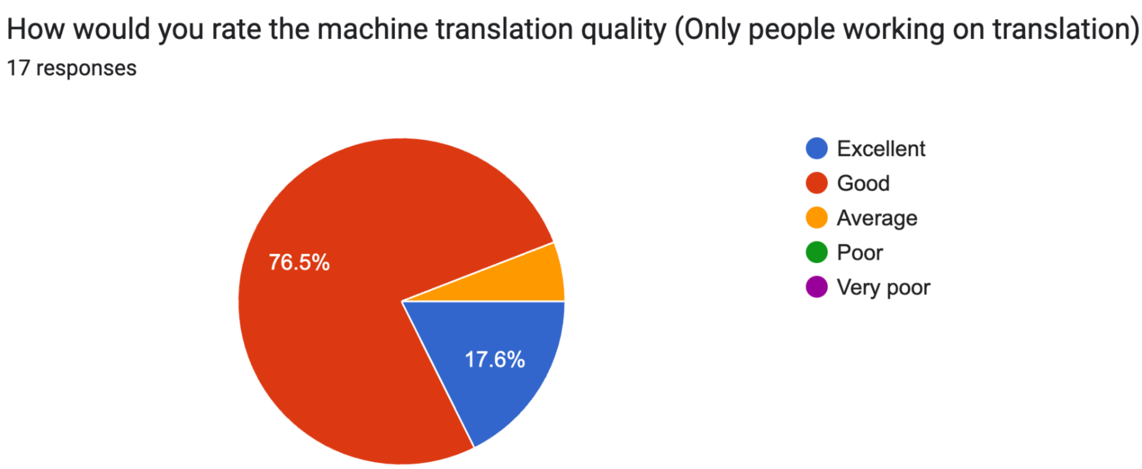


Figure 6: Feedback on Translation Quality from a subset of Users