BHASHA: Achieving Sarcasm Interpreted Translation

Chethan Reddy Chittireddy
University of Florida
chittireddy.c@ufl.edu

Suraj Kolla
University of Florida
n.kolla@ufl.edu

Amith Prem Nagandla
University of Florida
a.nagandla@ufl.edu

Sai Prakash Mushkara
University of Florida
s.mushkara@ufl.edu

Abdul Kalam Azad Shaik
University of Florida
shaik.abdulkalam@ufl.edu

Abstract

The performance of integrated social media translators is not reliable in translating texts that include sarcasm. This is partly because the language space used in most social media is informal, and translation models cannot understand the meaning behind sarcasm. In this research, we address the problem of sarcasm interpretation in machine translation specifically for Telugu. We present fine-tuned transformer models for English to English sarcasm interpretation and English to Telugu sarcasm translation tasks. we test whether a direct model of English-sarcastic to Telugu-honest or an indirect model of English-sarcastic to English-honest to Telugu-honest is better. This involves combining NLP techniques for sarcasm interpretation and translation, addressing a gap in current research, which lacks focus on sarcasm explanation. We also emphasize the significance of human annotation in addressing open-class Neural Machine Translation (NMT) challenges, like translating sarcasm and achieving accurate translations for low-resource languages. We are optimistic that this research offers opportunities to enhance the efficiency of models addressing other open-class NMT problems.

1 Introduction

Considering the current popularity of social media platforms which has led to a surge in informal communication, often the communication channel is on a single channel shared by people communicating in multiple languages and to top it off most of the language space is informal and riddled with sarcasm. However, current translation tools struggle to accurately convey the intended meaning of sarcastic messages Figure - 1. This is because they cannot understand the nuances of informal language and the layered meaning behind sarcasm. This research aims to bridge this gap in the translation experience by developing a novel approach to translating sarcastic English tweets into honest Telugu interpretations.

We decided on a two pipelines approach to achieve this task as demonstrated by Figure - 3. Our main focus is on pipeline A, where the pipeline tackles sarcasm translation in two steps. First, we employ Seq2Seq models to identify and interpret the sarcasm present in the English tweet and convert the sentence into its English honest interpretations. Here, we hypothesize that transformer models, such as Bidirectional Encoder Representations from Transformers (BERT), will outperform traditional RNN-based approaches(current sarcasm interpretation models (Peled and Reichart, 2017) use this) due to their inherent ability to capture contextual meaning. By fine-tuning these models specifically for sarcasm detection, we aim to achieve a deeper understanding of the underlying sentiment intended through the sarcastic English sentence.

Secondly, an honest interpretation devoid of sarcasm will be translated into Telugu using machine translation techniques. This two-pronged approach ensures that the translated message accurately conveys the true meaning behind the sarcastic English tweet. Pipeline B will aim to achieve the Telugu interpretation directly from the English sarcasm sentence. While we approach this using two pipelines, we feel the performance using pipeline A would perform significantly better because Telugu is a low-resource language and the contextual performance of English to English is higher than English to any low-resource language.

This research has the potential to significantly improve communication and understanding, particularly in informal online interactions Figure - 2. By successfully translating sarcasm, we can enhance the user experience and foster more meaningful cross-lingual communication on social media platforms. Understanding which of the pipelines provides a high-quality interpretation enables us to establish a baseline for similar problems in the open class NMT problems. It is our goal to achieve high-quality low-resource interpretations and accurate translations, which we think is possible when implemented either with a high-resource to high-resource Seq2Seq model as an intermediate or over direct high-resource to low-resource Seq2Seq interpretation. Furthermore, this project contributes to the advancement of NLP by integrating sarcasm interpretation into the machine translation process, potentially leading to improved model performance in handling complex language features.

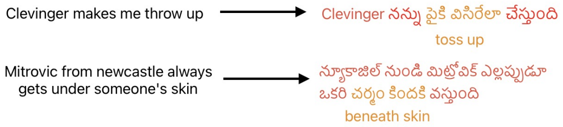
Clevinger makes me throw up
Clevinger నన్ను పైకి విసిరేలా చేస్తుంది
toss up
Mitrovic from newcastle always gets under someone's skin
న్యూకాజిల్ నుండి మిట్రోవిక్ ఎల్లప్పుడూ ఒకరి చర్మం కిందకి వస్తుంది
beneath skin

Figure 1: Current Translation Experience

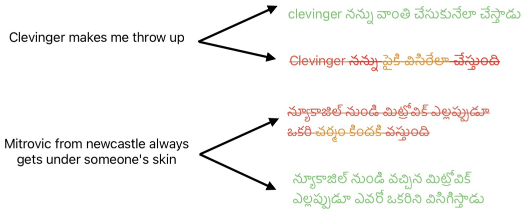
Clevinger makes me throw up
clevinger నన్ను వాంతి చేసుకునేలా చేస్తాడు
~~Clevinger నన్ను పైకి విసిరేలా చేస్తుంది~~
Mitrovic from newcastle always gets under someone's skin
~~న్యూకాజిల్ నుండి మిట్రోవిక్ ఎల్లప్పుడూ ఒకరి చర్మం కిందకి వస్తుంది~~
న్యూకాజిల్ నుండి వచ్చిన మిట్రోవిక్ ఎల్లప్పుడూ ఎవరో ఒకరిని విసిగిస్తాడు

Figure 2: Target Translation Experience

2 Background & Related Work

There has been a significant amount of research around sarcasm interpretation, but limited work on sarcasm translation, particularly in the context of text-to-text translation of memes. While notable progress has been made in sarcasm interpretation using multi-modal models (Desai et al., 2022), translating sarcastic content remains a challenge. Sarcasm translation is a part of open-class Neural Machine Translation (NMT) problems and it is from the fact that the meaning of these expressions is not composed by the meaning of their constituent words, but the Models tend to directly translate them leading to a translation which doesn't stand true to the intended meaning. There are several such language plays in English, and one such problem where there is significant progress in the translation experience is related to Idioms. (Baziotis et al., 2022) deals with evaluation and analysis of idioms, and gives us a good insight into how to approach these open class problems. However, it is to be noted that this paper does not deal with low-resource language adaptations, which we target to perform as a part of our project.

For sarcasm interpretation, our work builds build upon the approach proposed by (Peled and Reichart, 2017), who formulated it as a monolingual machine translation task. Their method, called SIGN (Sarcasm Sentiment Interpretation Generator), specifically targets sentiment words that convey the opposite of their literal meaning in sarcastic statements. SIGN first clusters sentiment words into positive and negative categories based on semantic relatedness. It then replaces these sentiment words with their corresponding cluster IDs in both the sarcastic source text and the non-sarcastic reference data. A phrase-based machine translation model is trained on this transformed data to map between the "sarcastic" source and "honest" target sides. At test time, SIGN generates outputs with cluster IDs replacing sentiment words, which then go through a de-clustering step to recover the final non-sarcastic interpretation. The paper explores three different de-clustering strategies as part of SIGN: centroid-based, context-based, and an oracle upper bound. In SIGN-centroid, each cluster-ID is replaced with the sentiment word closest to that cluster's centroid in the word embedding space. SIGN-context takes context into account, replacing each ID with the cluster word having the highest point-wise mutual information with neighboring words in the output. The oracle SIGN-oracle uses human judgments to choose the optimal sentiment word for each cluster-ID. While SIGN did not outperform baselines on automatic metrics, human evaluation showed its interpretations better captured the intended sentiment, especially with the context-based de-clustering approach.

For translating English into Telugu, the challenge is multifaceted, primarily due to the complexity inherent in accurately conveying nuanced expressions in a low-resource language like Telugu. As demonstrated in the efforts described in (Prasad and Muthukumaran, 2013) and related literature (Ramesh et al., 2023), while significant advancements have been made in machine translation for Indian languages, Telugu still presents unique challenges due to its rich morphological structure and syntactic diversity. To effectively tackle these challenges advanced translation models are essential. (Raffel et al., 2020) and (Tang et al., 2020) had presented machine translation models that have demonstrated exceptional performance in translation tasks across multiple language pairs. These models, based on the transformer architecture, can effectively capture long-range dependencies and contextual information, making them well-suited for the nuanced task of translating honest interpretations while preserving the intended sentiment and meaning. The studies (Desai et al., 2022) and (Baziotis et al., 2022) took advantage of transformer architectures like BART and mBART for achieving results on related tasks. We also need special custom tokenizers for evaluating the Telugu translation performance (Gala et al., 2023). These tokenizers are essential for dissecting and reconstructing the layered linguistic features unique to Telugu, such as sandhi (morphophonemic changes) and samasa (compound formations). This approach is essential not just for evaluating the literal meaning but also for considering the subtleties and cultural context embedded within the language, ensuring that the translations maintain the integrity and depth of the original honest interpretations.

3 Methodology

3.1 Datasets

To evaluate the sarcasm interpretation pipelines we have in our current project, we need a dataset where we have English sarcasm sentence collections and possible sarcastic interpreted Telugu translations for each of those English sentences. Since there are no such high-quality datasets that have the above-mentioned features that are currently available, it is an important part of the project to generate a high-quality, thoroughly vetted dataset that serves our needs. To build this dataset, we are currently extending upon a base dataset called Sarcasm SIGN dataset from (Peled and Reichart, 2017)[1], which contains 5 possible sarcasm interpretations for each of the 2993 unique English sarcastic tweets. This gives us a solid foundation to generate a similar dataset, where we have five possible Telugu sarcasm interpretations given the English sarcasm sentence. To speed up the process of creating the dataset, we first produced an uncorrected translation version, where we utilized the existing high-quality translation performance of the Google translation models using the Google Translate API. We translated all the 14965 English interpretations into Telugu interpretations.

What separates our project from just utilizing the existing Google Translate is the fact that we produced a final version of interpretations, which we now call the corrected translated version. We achieved this corrected translation by manually correcting all those 14965 translations. The task of correction is split among all our teammates who are well-versed in the Telugu language, which is our primary language. The correction procedure includes correcting the sentences in Telugu to better represent the original meaning of the English interpretations. Since Telugu is a low-resource language, it is very possible that the current translators do not have enough language context to directly translate most of the words into English. In the process of correction, we also note down all the repeated patterns in the mis-translation cases.

After going through our notes, these are some of our main observations. Translator models sometimes fail to transliterate non-alphabetic symbols into Telugu. Some return unicode characters instead of just returning the symbol without translation. We also have several cases of mis-translations, where we do not have native terminology for a given English word. Let us take the scenario where the translator is translating the name of a company or a football team or a translation which is also part of the open class NMT problems such as idiom translations etc. It is not given that the translator models have the information of their native Telugu counterparts. We utilized top Telugu newspapers such as Eenadu[2] and Sakshi[3] to search for the given entity and to get its Telugu counterpart. We update it with the Telugu counterpart if present. If a native usage is not present, we either leave the text in English, since it is a common practice to have some English nouns and all English numbers used directly in Telugu, or we express the phonetics of the English word in Telugu (which is also a common practice). Moving forward, we use these manually labeled Telugu interpretation data as our ground truth for experimenting and scoring the model results.

3.2 Experiment Design

We present two schemes to interpret and translate English sarcasm. For the first one, we interpret and translate the model in two phases. First, we fine-tune a seq2seq model on English sarcasm to English interpretation. In the second phase, we fine-tune a machine translation model from English interpretation to Telugu interpretation. For

[1] SarcasmSIGN Dataset

[2] Eenadu.net Newspaper

[3] Sakshi.com Newspaper

the second one, we fine-tune different pre-trained machine translation models on our English sarcasm to Telugu interpretation data.

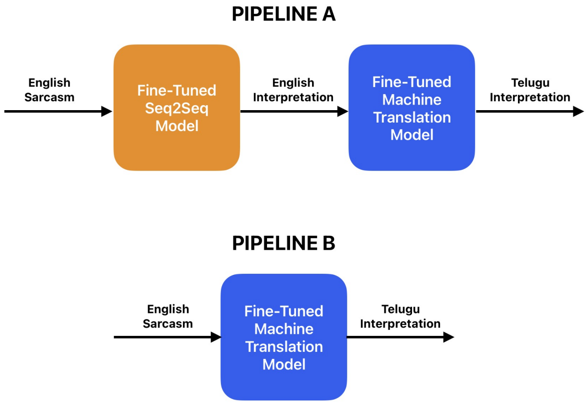

Figure 3: Target Translation Experience

We leveraged state-of-the-art seq2seq and machine translation models that are readily available in the HuggingFace library. Specifically, we employed the powerful google-t5/t5-*[4], and `facebook/bart-*`[5] for our English to English sarcasm interpretation task and `google/mt5-*`[6], and `facebook/mbart-*`[7] models for our translation tasks.

For all the models, we use validation loss as an early stopping criterion and use the model with the best validation loss scores to evaluate the training. We kept the early stopping patience to 5, and the models were fine-tuned for around 15 epochs. We trained the models on three `NVIDIA A100-SXM4-80GB` GPUs to leverage hardware acceleration. We used a batch size of 32. To evaluate the performance of our models, we used 20% of the data each for validation and testing. We report the automatic and human-evaluated scores on the test set.

3.3 Test Design & Metrics

To evaluate the performance of our proposed models we employ three different metrics BLEU, ROUGE, and PINC. PINC (Chen and Dolan, 2011) is a score based on n-gram dissimilarity between the source and the prediction texts, originally used to check the performance of a paraphrasing task. We use it to only compare our English to English sarcasm interpretation performance with (Peled and Reichart, 2017).

We calculated all the metrics on our test dataset. For Telugu evaluation, we used our manually translated Telugu interpretations as references for the end (Telugu) results in both pipelines. Following (Ramesh et al., 2023), we used the pre-trained tokenizer of the respective machine translation models to tokenize the Telugu predictions and references before calculating the metrics.

Human Evaluation: We also employ a human evaluation to assess the quality of the sarcasm translation. We designed our human evaluation approach based on (Desai et al., 2022)'s human evaluation. For this, we sample 25 random examples from the test set and ask our 7 evaluators (Evaluators are linguistic experts in the 20-30Y age bracket) to rate the translation outputs from the best model (model with the highest BLEU score) of each pipeline. Evaluators rate the interpreted translations based on two metrics Adequacy and Fluency. The former metric measures the correctness of interpretations of the underlying sarcasm of the English sentences, whereas the latter metric measures the coherency of the output Telugu sentence.

The evaluators are provided with 4 rating options: `Excellent`, `Good`, `Fair`, and `Poor` to rate each translation on the two metrics. We first perform majority voting for each question based on the ratings by the evaluators, for both fluency and adequacy. After question-wise majority voting, we average the evaluation metrics on the whole dataset, to get a final metric that we feel can provide us better insight as to which model produced a better real-life interpreted translation.

For tie breaks in majority voting, in the case of a two-way tie, we take the lowest rating as the winner, and for ties of greater length, we chose the median value as the winner for the tie-breaking scheme.

4 Results and Analysis

We discuss the results of our experiment here and compare its performance with the previous works in the literature.

In Table 1 we present the results we obtained by fine-tuning seq2seq models using the English Sarcasm to English Interpretation data as part of our pipeline A. For the SIGN model we present the best results reported by (Peled and Reichart, 2017)

[4] T5 HuggingFace
[5] BART HuggingFace
[6] mT5 HuggingFace
[7] mBART HuggingFace

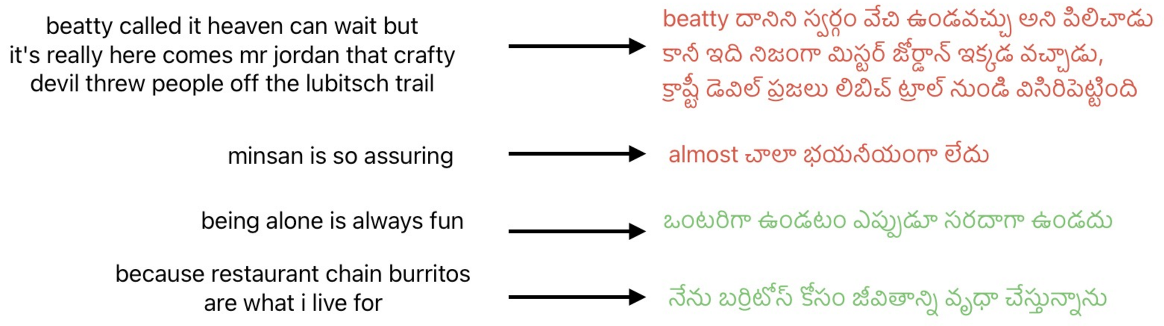
beatty called it heaven can wait but it's really here comes mr jordan that crafty devil threw people off the lubitsch trail
beatty దానిని స్వర్గం వేచి ఉండవచ్చు అని పిలిచాడు కానీ ఇది నిజంగా మిస్టర్ జోర్డాన్ ఇక్కడ వచ్చాడు, క్రాఫ్టీ డెవిల్ ప్రజలు లిబిచ్ ట్రాల్ నుండి విసిరిపెట్టింది
minsan is so assuring
almost చాలా భయనీయంగా లేదు
being alone is always fun
ఒంటరిగా ఉండటం ఎప్పుడూ సరదాగా ఉండదు
because restaurant chain burritos are what i live for
నేను బర్రిటోస్ కోసం జీవితాన్ని వృధా చేస్తున్నాను

Figure 4: Low (coloured red) and high rated (coloured green) (based on Human evaluation) samples from the best Sarcasm Interpretation model outputs

for each score.

We see high BLEU and ROUGE scores from our fine-tuned models compared to SIGN models, implying the predictions were close to our expected interpretations. We see a lower PINC score, and we attribute this to our observation that the expected English interpretation sentences are close to the source sarcasm sentences, sometimes with only a few word changes: "I love working on Mondays" versus "I hate working on Mondays". We observed the best BLEU score with the BART-large model and the best ROUGE scores with the T5-large model; however, all the models performed comparatively equally.

In Table 2, we present BLEU and ROUGE metrics for both pipelines with various fine-tuned models. Using the mT5 model for translation did not yield great results; the mBART model for translation proved to be the best in our study. The pipeline A method using the T5-large model for interpretation and the mBart-large many-to-many model for translation gave us the best BLEU score of 35.80 overall. This two-stage approach of first interpreting sarcasm in English and then translating the honest interpretation allowed us to leverage the strengths of specialized models for each task, ultimately resulting in higher-quality sarcasm interpretation. However, the pipeline B technique did not fall too far behind, with a BLEU score of 31.69 with the same machine translation model. The ROUGE scores are also very close across the pipelines with the best being 15.68, 7.07, and 15.59 for ROUGE-1, ROUGE-2, and ROUGE-L respectively.

Even though the scores look very low compared to the English-to-English model scores, considering the challenges in automatically evaluating Telugu as we discussed in the Background & Related Work section and looking at how previous state-of-the-art BLEU scores for English to Telugu translation are only around on various translation tasks (Ramesh et al., 2023), we believe the scores reflect our fine-tuned models performed well on the task, which is also reflected in our Human evaluation results (Table 3)

Pipeline	Adequacy	Fluency
A	3.8	3.88
B	3.2	3.04

Table 3: Human evaluation results for Pipeline A and B

From Table 3, we can observe that the adequacy scores for both the Pipelines A and B are above 3 implying Good interpretation for most of the sentences. But, we can see that, along with a slight edge in adequacy, Pipeline A performs significantly better when it comes to Fluency. We can see it almost always produces Great fluent output. Based on the human evaluation ratings, we sampled out 2 of the lowest adequacy-rated sentences along with two highly rated sentences in Figure 4. One of our observations based on the human evaluation outputs is that the lower-length sentences often produce better results compared to the longer sentences. Another of our observations is how the pre-processing to remove the punctuation from the English sentences, makes the model assume the wrong context in the sentence structure thus producing improper sarcasm-interpreted translations.

Model	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	PINC
SIGN‡	66.96	70.34	42.81	69.98	47.11
T5-base†	84.34	87.89	80.90	87.37	15.97
T5-large†	85.29	89.28	82.83	88.95	13.83
BART-large†	86.32	86.40	80.73	86.21	11.06

Table 1: Comparison of SIGN model's best scores with our English to English monolingual sarcasm interpretation fine-tuning performance in Pipeline A. ‡best model from (Peled and Reichart, 2017) †fine-tuned model.

	Interpretation Model	Translation Model	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
Pipeline A	T5-base†	mBART-large-50-many-to-many-mmt†	35.39	15.17	6.04	14.85
	T5-large†		**35.80**	15.00	6.26	14.73
	BART-large†		35.69	15.34	6.69	15.04
	T5-base†	mBART-large-50-one-to-many-mmt†	33.92	15.44	6.63	15.34
	T5-large†		33.12	15.46	6.62	15.30
	BART-large†		33.47	**15.68**	*6.81*	**15.59**
	T5-base†	mT5-base†	17.64	11.00	3.73	10.91
	T5-large†		17.37	12.18	4.47	12.05
	BART-large†		17.01	11.84	4.02	11.61
Pipeline B		mBART-large-50-many-to-many-mmt†	*31.69*	14.48	**7.07**	14.42
		mBART-large-50-one-to-many-mmt†	30.39	*14.72*	6.97	*14.60*
		mT5-base†	13.54	8.88	2.95	8.86

Table 2: Comparison of English to Telugu sarcasm interpretation model evaluations of both pipelines. **Bold** scores (**0.00**) indicate overall best score, *italic* scores (*0.00*) indicate the best score within the pipeline, †fine-tuned model.

5 Conclusion and Future Work

In this paper, we explored two approaches with the goal of achieving accurate sarcasm interpretation and translation from English to Telugu. The first approach involves an indirect, two-stage process of first interpreting English sarcasm into honest English using seq2seq models, followed by translating to honest Telugu using machine translation models. The second is a direct approach to translating English sarcasm into honest Telugu interpretations. For effectively fine-tuning the machine translation models, we curated a manually annotated honest Telugu interpretation dataset and evaluated the performance using automatic metrics like BLEU and ROUGE along with human judgments. The results demonstrated that our two-stage pipeline performed better by leveraging specialized models for each subtask. While we see the direct translation model struggled a bit to accurately interpret and translate sarcasm, we attribute this behavior to the limited NLP resources of the Telugu language, which wouldn't be the case for a high-resource language, which was also demonstrated by the superior performance in the case of interpreting English sarcasm in a high-resource language (English). This research contributes towards bridging a crucial gap in existing translation systems, which often struggle with understanding and conveying nuanced expressions like sarcasm accurately across languages.

Our research is currently focused on interpreting sarcasm from English to Telugu and presents only the literal meaning in Telugu. Further development to enable direct translation of the sarcastic intent into Telugu or other languages would be a good future scope for this project. This could be achieved by expanding the training dataset to include examples of sarcastic translations and incorporating contextual cues to guide the appropriate sarcastic phrasing in the target language. Moreover, understanding sarcasm involves applying prior information and context recognition even in words that don't explicitly express sentiment. Therefore, additional research in this field is another prospect for the future.

References

Christos Baziotis, Prashant Mathur, and Eva Hasler. 2022. Automatic evaluation and analysis of idioms in neural machine translation.

David Chen and William B Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 190–200.

Poorav Desai, Tanmoy Chakraborty, and Md Shad Akhtar. 2022. Nice perfume. how long did you marinate in it? multimodal sarcasm explanation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10563–10571.

Jay Gala, Pranjal A Chitale, Raghavan AK, Sumanth Doddapaneni, Varun Gumma, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, et al. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *arXiv preprint arXiv:2305.16307*.

Lotem Peled and Roi Reichart. 2017. Sarcasm sign: Interpreting sarcasm with sentiment based monolingual machine translation. *arXiv preprint arXiv:1704.06836*.

T Venkateswara Prasad and G Mayil Muthukumaran. 2013. Telugu to english translation using direct machine translation approach. *International Journal of Science and Engineering Investigations*, 2(12):25–32.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

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. 2023. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning.