A fully automated and scalable Parallel Data Augmentation for Low Resource Languages using Image and Text Analytics

Prawaal Sharma
Infosys
Pune, Maharashtra, India
prawaal_sharma@infosys.com

Navneet Goyal
BITS Pilani
Pilani, Rajasthan, India
goel@pilani.bits-pilani.ac.in

Poonam Goyal
BITS Pilani
Pilani, Rajasthan, India
poonam@pilani.bits-pilani.ac.in

Vishnupriyan K R
Infosys
Chennai, Tamil Nadu, India
vishnupriyan.r02@infosys.com

ABSTRACT

Linguistic diversity across the world creates a disparity with the availability of good quality digital language resources thereby restricting the technological benefits to majority of human population. The lack or absence of data resources makes it difficult to perform NLP tasks for low-resource languages. This paper presents a novel scalable and fully automated methodology to extract bilingual parallel corpora from newspaper articles using image and text analytics. We validate our approach by building parallel data corpus for two different language combinations and demonstrate the value of this dataset through a downstream task of machine translation and improve over the current baseline by close to 3 BLEU points.

CCS CONCEPTS

• **Information systems** → **Multilingual and cross-lingual retrieval**; *Surfacing*; *Personalization*;

KEYWORDS

Low resource language, Parallel Data Augmentation, Information Mining

ACM Reference Format:
Prawaal Sharma, Navneet Goyal, Poonam Goyal, and Vishnupriyan K R. 2023. A fully automated and scalable Parallel Data Augmentation for Low Resource Languages using Image and Text Analytics. In *The 38th ACM/SIGAPP Symposium on Applied Computing (SAC '23), March 27-March 31, 2023, Tallinn, Estonia.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3555776.3577788

1 INTRODUCTION

Modern NLP research focuses largely on the languages on the internet, which consists of only 20 of the 7,000 languages of the world [11]. This leaves the majority of languages understudied, which are also referred to as low-resource languages (LRLs), and are spoken by a large section of world population. LRLs can be described as *resource scarce, under studied, less digitized, under privileged or less commonly taught*, among other denominations [29] [9]. There are more than 2.5 billion inhabitants using 2,000 LRLs, within India and Africa and any progress for these languages shall help in digital enablement of these semi-literate populations [30].

Most NLP tasks (e.g. neural networks) require large amounts of training data. Availability of data therefore becomes more relevant in context of LRLs where scarcity of digital data is the primary challenge for taking NLP to masses. In this paper, we describe a novel approach using image and text analytics to build a completely automated, scalable and language agnostic methodology for bilingual parallel dataset generation.

We use Konkani-Marathi as the primary example to establish our claims. Our choice for Konkani is based on the scarcity of digital resources along with its small population of native speakers [24]. Marathi, on the other hand is resource rich and hence the combination would be instrumental in wide NLP applications. It has been established that the significance of parallel corpora is best observed with *resource deficit-resource rich* language combination [10].

As illustrated in Figure 1, we observe that a lot of local newspapers in LRLs (published by large publishing houses) also have editions in other languages and they re-use the pictures across different language versions to optimise on resources. We apply this

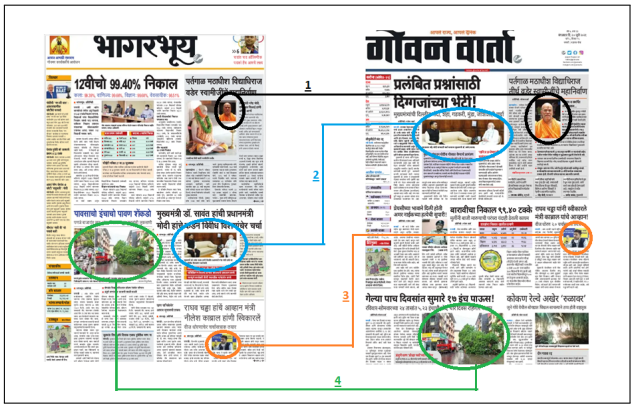

Figure 1: Article mapping using images as pivots.

SAC '23, March 27-March 31, 2023, Tallinn, Estonia

ACM ISBN 978-1-4503-9517-5/23/03.
https://doi.org/10.1145/3555776.3577788

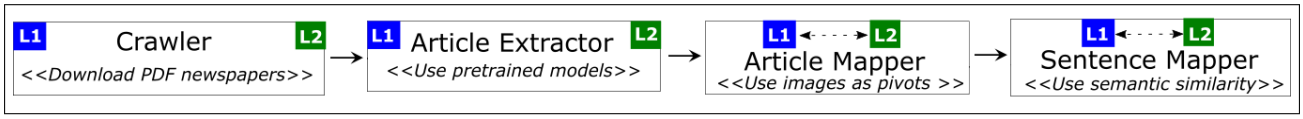

Figure 2: Proposed data augmentation pipeline

observation and use images as pivots for article mapping. Once articles are mapped, article text is extracted and sentences are mapped to form parallel corpus, which is then empirically evaluated.

In a nutshell, the contributions from our work are three fold:

- We use newspaper article images as pivots to map articles, which has not been explored earlier in similar context.
- We use language agnostic embeddings for sentence mapping (on LRL combinations) and empirically substantiate this.
- Our final Konkani-Marathi corpus is largest available dataset created without human annotations.

2 RELATED WORK

Our work leverages the principles from image and text analytics to build sizeable parallel data corpus augmented from newspaper articles.

2.1 Image Analytics

The hypothesis for our work is that, news articles with same (or quasi same) images would have same information in text. We break this into primarily two parts (a) segmenting newspaper page into various article regions including pictures and (b) Matching images for article mapping.

Marking boundaries for articles, specially in a newspaper image (non machine readable) is a very difficult task. Most newspapers have multi column format with no explicit boundaries marked for individual articles. Existing article segmentation approaches include heuristic-based [15], graph embedding based [14] and deep learning based [31].

Image matching is based on feature detection of images. Feature detection is an abstraction of the image information and making a local decision at every image point to see if there is an image feature of the given type existing in that point. This should ideally be robust to image transformations such as rotation, scale, illumination, noise and affine transformations. The most popular image matching algorithms for this include scale invariant feature transform (SIFT) [19], speed up robust feature (SURF) [3] and robust independent elementary features (BRIEF) [7]. In the recent times, variants of neural networks including Convolutional Neural Network (CNN) has been explored and found more efficient than traditional image processing based techniques [16].

2.2 Text Analytics

For our work, the extracted and mapped articles need further processing on (a) text extraction and (b) sentence level mapping.

An Optical Character Recognition (OCR) system is a framework by which embedded textual information is repossessed through application of character extraction. Girdher et al. have done an extensive survey on Devanagari OCR and multiple approaches to accomplish this task [12].

Sentence alignment across languages is an critical step for building bilingual (or multilingual) corpus. The existing approaches for sentence alignment are based on length based heuristics [5][8], lexical correspondences [32] and the recent deep learning based approaches which makes use of language agnostic sentence embedding (like LaBSE, which uses BERT like architecture), that can be compared using cosine-similarity [13, 27].

2.3 NLP for Konkani

Existing NLP research for Konkani is limited to more elementary NLP tasks including POS tagging, sentiment analysis, NER etc. [17] [22], [23]. Indian Languages Corpora Initiative (ILCI) is a project of Technology Development in Indian Languages (TDIL)[1], an agency of the Indian government. They have been engaged to create corpus for low resource Indian languages to facilitate research and avoid its extinction. They have created, two sets (across two different domains) of Hindi-Konkani corpus containing 25,000 sentences accomplished by human annotators [25] [18]. The same dataset is used to accomplish the task of Neural Machine Translation (NMT) and a BLEU [21] score of 23.5 has been achieved [28]. This is considered as current benchmark for Konkani related machine translation.

3 METHODOLOGY

As illustrated in Figure 2, the proposed data augmentation pipeline contains four components (a) Crawler, (b) Article Extractor, (c) Article Mapper and (d) Sentence Mapper. While *crawler* and *article extractor* work on the two languages (L1 and L2 as represented) independently, the *mappers* (Article and Sentence) use the languages in conjunction and perform mappings.

3.1 Crawler - Raw data collection

Crawler, helps download copies of newspapers[2] from online sources. Downloaded files are not machine readable and the content is "locked" in a snapshot-like image. Crawler, helps in splitting the files into individual pages and label appropriately (with dates, page numbers and language code) to ensure easy referencing in the downstream processes.

3.2 Article Extractor - Segmentation

Article Extractor, helps with two functions, (a) Marking boundaries for individual articles (b) Extraction of images and text (using OCR) within the marked article. Embedded articles are considered parts of parent articles in our work due to its strong association with parent article. We use layout analysis dataset by Pattern Recognition and Image Analysis Research Lab (PRImA) for article boundary

[1] http://tdil.meity.gov.in/

[2] Bhaangarbhuin for Konkani and Goan Varta for Marathi

Figure 3: Article sub parts and nomenclature

detection [2] [31]. To extract regions of interest (ROI) from segmented articles OpenCV [4] is used to find contours, apply simple size based heuristics and remove noise using Run Length Smoothing Algorithm (RSLA). For text extraction we consider a combination of EasyOCR[3], PaddleOCR[4] and Tessaract[5] and use majority voting for final decision.

As illustrated in Figure 3 we partition the article into into four regions of interest (ROI) (a) Headlines (H) (which includes sub headlines) (b) Images (I) (c) Picture Captions (P) and (d) Contents (C). In case of embedded articles (e.g. H_2^0 and C_2 in Figure 3), indexing is done to maintain the hierarchy with parent article.

$$a \equiv \begin{cases} H_1^0, H_1^1, H_1^2 ... & \text{Main article headlines} \\ H_2^0, H_2^1, H_2^2 ... & \text{Embedded article headlines} \\ I_1, I_2, I_3 ... & \text{Images} \\ P_1, P_2, P_3 ... & \text{Picture captions} \\ C_1, C_2, C_3 ... & \text{Content} \end{cases} \quad (1)$$

Finally, each extracted article is stored as a set of image files and text file. Each ROI within the article is labelled with markers within the text file as illustrated in Equation 1.

3.3 Article mapper - Match articles

Article Mapper, compares the articles images (I_i,I_j) across the two languages ($L1, L2$) for similarity (for same date (dt)) and builds the set of mapped article tuples (a_i^{L1}, a_i^{L2}) when image similarity score is beyond the set threshold. Embedded articles are mapped by comparing headline similarity.

$$\{(a_1^{L1}, a_1^{L2}), (a_2^{L1}, a_2^{L2}) ...\} \equiv \theta(\, I_i^{L1}, I_j^{L2}) \quad (2)$$

$$\theta \textit{ is the image matching algorithm function} I_i^{L1} \in \forall \textit{ (images for date (dt) \& language (L1))} I_i^{L2} \in \forall \textit{ (images for date (dt) \& language (L2))}$$

[3]https://github.com/JaidedAI/EasyOCR.git
[4]https://github.com/PaddlePaddle/PaddleOCR
[5]https://github.com/tesseract-ocr

We use SIFT as the image matching algorithm (θ) for our work, since the valid image combinations are exact copies of each other and differ only in scale, illumination and shifts.

3.4 Sentence mapper - Match sentences

The extraction and alignment of data till article level is straight forward task. However, the sentences between the mapped articles may not be positioned sequentially. Hence it becomes the most critical step in our experiment and impacts the overall accuracy of our experiment.

Sentence Mapper, helps to map sentences within the mapped articles (a_i^{L1}, a_i^{L2}) for all sentences combinations applying semantic sentence similarity algorithms and build the set of mapped sentence tuples (s_j^{L1}, s_j^{L2}).

$$\{(s_1^{L1}, s_1^{L2}), (s_2^{L1}, s_2^{L2}) ...\} \equiv \delta(\, S_k^{L1}, S_k^{L2}) \quad (3)$$

$$\delta \textit{ is the semantic sentence similarity algorithm} S_k^{L1} \in \forall \textit{ (sentences for article } a_i^{L1}) S_k^{L2} \in \forall \textit{ (sentences for article } a_i^{L2})$$

We consider three types of metrics for sentence similarity (δ), in our experiment (a) Language agnostic sentence embedding based cosine similarity (LAS), (b) Simple Length base heuristics (SLAS) and (c) Lexical overlap based metrics (LO).

Language Agnostic Sentence embeddings converts sentence text into vectors to capture semantic information. These models are based on BERT-like architecture and is trained on 119 languages of different origins [13]. It claims to work universally including for the languages not part of its training corpus. We convert sentences into these vectors independently first and then find the cosine similarity between them, and refer as Language agnostic sentence embedding similarity (LAS) in our work.

Sentence Length Alignment heuristics is another metric to find the semantic similarity based on sentence length, and its position within the article. Sentence similarity is calculated as a function of number of words in a sentence, after filtering punctuation marks and applying adjustment factors for average sentence length across languages along with adjustment for the size of articles mapped [6].

Lexical Overlap is based on precision, recall and F-Score on common words across the candidate sentence set to measure the overlap statistically using a pivot language. We use English as a pivot language and perform lexical translation to English for both the languages. We use google translate for Marathi to English lexical conversion and English-Konkani dictionary by Maffei and Xavier [20] for Konkani to English lexical conversion.

4 EXPERIMENTAL SETUP AND RESULTS

We run our experiment in two parts, (a) Intrinsic evaluation using multiple language combinations, (b) Extrinsic evaluation using downstream task of machine translation (MT) . We use Konkani-Marathi as the primary language combination to study multiple sentence mapping methods, and later apply the same process on Punjabi-Hindi language pair to establish that our methodology is language agnostic and works universally.

Mapping Strategy	Sentence Lengths (1-10) words	(11-19) words	20+ words	Article Lengths (1-5) sentences	(6-15) sentences	(16+) sentences
LAS	3.8	3.7	3.8	3.8	3.8	3.7
SLAS	3.4	3.4	3.2	3.1	3.5	3.3
LO	2.9	3.0	2.6	2.8	2.9	2.9

Table 1: STS for various sentence and article size

Variables	Value Konkani-Marathi	Punjabi-Hindi
Mapped articles	1,320	150
Mapped sentences	14,448	2,200
Human evaluation	600	100
STS Score	3.70	3.73

Table 2: Summary view on varioud heads

4.1 Konkani-Marathi parallel corpus evaluation

We apply our methodology (as described in Section 3) and build a parallel corpus for Konkani-Marathi using all sentence mapping methods independently. We analyse the quality of each corpus by applying human annotation on a subset of sentence pairs from our experiment. For defining the annotation scores, we use Semantic Textual Similarity (STS), characterised by six ordinal levels ranging from complete semantic equivalence (5) to complete semantic dissimilarity (0) [1]. This is the most widely used evaluation metric for parallel data augmentation tasks, and used by multiple people working on this field [26].

We have sampled a total of 900 sentences, in two phases. In the first phase we have sampled 200 pairs from each sentence alignment strategy (making a total of 600 sentence pairs) and analysed the results. In the second phase we have sampled, another 300 pairs for the most appropriate sentence mapping strategy as evident from first phase analysis. The sampling was stratified, shuffled randomly such that no ordering is preserved. We illustrate the evaluation in two parts as below.

Phase 1: As illustrated in Table 1 the human evaluation for different sentence and article sizes clearly indicate that LAS gives the best results and is most appropriate for our task. We also observe that LAS and SLAS have moderate corelation with each other.

Phase 2: Based on the results from first phase, we use LAS as our final alignment method. As illustrated in Table 2 our Konkani-Marathi parallel corpus contains 14,448 sentence-pairs (aggregating a total of 28,896 sentences extracted). We evaluate STS on a corpus of 500 sentences (200 from first phase LAS evaluation and another 300 in second phase) and observe that the average STS in our corpus is 3.7 and more than 92% of our mapped sentences, have STS score of greater than 3.

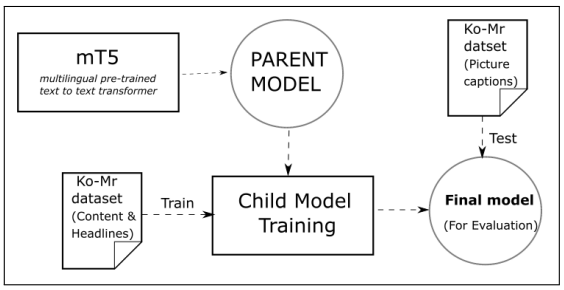

Figure 4: Proposed model for our evaluation

4.2 Punjabi-Hindi evaluation

We execute the same experiment on Punjabi-Hindi language combination, following the exact same steps on a smaller scale. We use Ajit[6] as our source of raw news corpus for Punjabi and Hindi news. We use a smaller set of news articles across one month's time span. Our intent of doing this as an extended experiment is to analyse and validate the significance of our work on other language pairs. As illustrated in Table 2 the results on this language pair is at par with our primary experiment.

Our results (considering both language combinations) indicate that our methodology is agnostic of language combinations and reasonably good to build a scalable model for parallel corpus generation[7].

4.3 Case study: MT for Konkani-Marathi

To access the quality of the entire Konkani-Marathi aligned corpus, we perform extrinsic evaluation using the downstream machine translation (MT) task which shall leverage the entire aligned corpus as an input.

As illustrated in Figure 4, we use mT5 (multilingual pre-trained text to text transformer) as our parent model [33], and further fine tune with our Konkani-Marathi parallel corpus (headlines and article content). We have used picture captions as ground truth to test our translation model and achieved the BLEU score of 26.4 which is an improvement over the current baseline for Konkani, by around 3 BLEU points (compared to existing baseline of 23.5 as mentioned in Section 2.3). This further validates the overall quality of the dataset.

5 CONCLUSION AND FUTURE WORK

The paper presents a simplistic novel methodology for parallel data augmentation using newspaper articles as raw input, empirically validated for multiple language combinations. We use a simple DA architecture applying images as pivots for article mapping and multilingual sentence embedding for sentence mapping and achieve reasonable results. We conclude that in order to boost the quality of the dataset additional constraints may need to be provided. The scale of the dataset can be enhanced by considering images clicked by different people depicting same event. We aim to explore these in our future work.

REFERENCES

[1] Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez Agirre, Rada Mihalcea, German Rigau Claramunt, and Janyce Wiebe. 2016. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation.

[6] https://www.ajitjalandhar.com/

[7] https://github.com/prawaal/Konkani-Marathi-Data-Corpus/

In *SemEval-2016. 10th International Workshop on Semantic Evaluation; 2016 Jun 16-17; San Diego, CA. Stroudsburg (PA): ACL; 2016. p. 497-511.* ACL (Association for Computational Linguistics).

[2] Apostolos Antonacopoulos, David Bridson, Christos Papadopoulos, and Stefan Pletschacher. 2009. A realistic dataset for performance evaluation of document layout analysis. In *2009 10th International Conference on Document Analysis and Recognition*. IEEE, 296–300.

[3] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. 2006. SURF: Speeded up robust features. *Computer Vision-ECCV 2006* 3951, 404–417. https://doi.org/10.1007/11744023_32

[4] Gary Bradski and Adrian Kaehler. 2000. OpenCV. *Dr. Dobb's journal of software tools* 3 (2000), 2.

[5] 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*. 169–176.

[6] 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*. 169–176.

[7] Michael Calonder, Vincent Lepetit, Christoph Strecha, and Pascal Fua. 2010. Brief: Binary robust independent elementary features. In *European conference on computer vision*. Springer, 778–792.

[8] Kenneth Church. 1993. Char_align: A program for aligning parallel texts at the character level. In *31st Annual Meeting of the Association for Computational Linguistics*. 1–8.

[9] Christopher Cieri, Mike Maxwell, Stephanie Strassel, and Jennifer Tracey. 2016. Selection criteria for low resource language programs. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. 4543–4549.

[10] Irene Doval and M Teresa Sánchez Nieto. 2019. Parallel corpora in focus. *Parallel corpora for contrastive and translation studies: New resources and applications* 90 (2019), 1.

[11] Long Duong. 2017. *Natural language processing for resource-poor languages*. Ph.D. Dissertation.

[12] Heena Girdher, Harmohan Sharma, and Akant Gupta. 2022. Comprehensive Survey on Devanagari OCR. *Available at SSRN 4033489* (2022).

[13] GoogleAI. 2020. Language-Agnostic BERT Sentence Embedding.

[14] Philip Hausner and Michael Gertz. 2021. News Article Extraction Using Graph Embeddings. (2021).

[15] David Hebert, Thomas Palfray, Stephane Nicolas, Pierrick Tranouez, and Thierry Paquet. 2014. Automatic article extraction in old newspapers digitized collections. In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*. 3–8.

[16] Jun-Ki Hong. 2022. Analysis of Image Similarity Using CNN and ANNOY. *International Journal of Software Innovation (IJSI)* 10, 2 (2022), 1–11.

[17] Diksha N Prabhu Khorjuvenkar, Megha Ainapurkar, and Sufola Chagas. 2018. Parts of speech tagging for Konkani language. In *2018 Second International Conference on Computing Methodologies and Communication (ICCMC)*. IEEE, 605–607.

[18] KM Chaman Kumar, Shailendra Aswale, Pratiksha Shetgaonkar, Vijaykumar Pawar, Deepmala Kale, and Sweta Kamat. 2020. A Survey of Machine Translation Approaches for Konkani to English. In *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*. IEEE, 1–6.

[19] David G Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision* 60, 2 (2004), 91–110.

[20] Angelus Francis Xavier Maffei. 1883. *An English-Konkani Dictionary*. Basel Mission Press.

[21] 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*. 311–318.

[22] Annie Rajan and Ambuja Salgaonkar. 2020. Sentiment Analysis for Konkani Language: Konkani Poetry, a Case Study. In *ICT Systems and Sustainability*. Springer, 321–329.

[23] Annie Rajan and Ambuja Salgaonkar. 2022. Named Entity Recognizer for Konkani Text. In *ICT with Intelligent Applications*. Springer, 687–702.

[24] Annie Rajan and Ambuja Salgaonkar. 2022. Survey of NLP Resources in Low-Resource Languages Nepali, Sindhi and Konkani. In *Information and Communication Technology for Competitive Strategies (ICTCS 2020)*. Springer, 121–132.

[25] Annie Rajan, Ambuja Salgaonkar, and Ramprasad Joshi. 2020. A survey of Konkani NLP resources. *Computer Science Review* 38 (2020), 100299.

[26] Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Divyanshu Kakwani, Navneet Kumar, et al. 2021. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages. *arXiv preprint arXiv:2104.05596* (2021).

[27] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* (2019).

[28] Karthik Revanuru, Kaushik Turlapaty, and Shrisha Rao. 2017. Neural machine translation of indian languages. In *Proceedings of the 10th annual ACM India compute conference*. 11–20.

[29] Anil Kumar Singh. 2008. Natural Language Processing for Less Privileged Languages: Where do we come from? Where are we going?. In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*.

[30] Yulia Tsvetkov. 2017. Opportunities and challenges in working with low-resource languages. *Slides Part-1* (2017).

[31] Manchester University of Salford. 2004. Pattern Recognition and Image Analysis.

[32] Dániel Varga, László Németh, Péter Halácsy, András Kornai, Viktor Trón, and Viktor Nagy. 2005. Parallel corpora for medium density languages In Proceedings of the RANLP 2005. (2005).

[33] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mT5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934* (2020).