Automatic Data Retrieval for Cross Lingual Summarization

Nikhilesh Bhatnagar[1*] Ashok Urlana[2*] Vandan Mujadia[1]
Pruthwik Mishra[1] Dipti Misra Sharma[1]
IIIT Hyderabad[1] TCS Research, Hyderabad, India[2]
tingc9@gmail.com, ashok.urlana@tcs.com, vandan.mu@research.iiit.ac.in
pruthwik.mishra@research.iiit.ac.in, dipti.m@iiit.ac.in

Abstract

Cross-lingual summarization involves the summarization of text written in one language to a different one. There is a body of research addressing cross-lingual summarization from English to other European languages. In this work, we aim to perform cross-lingual summarization from English to Hindi. We propose pairing up the coverage of newsworthy events in textual and video format can prove to be helpful for data acquisition for cross lingual summarization. We analyze the data and propose methods to match articles to video descriptions that serve as document and summary pairs. We also outline filtering methods over reasonable thresholds to ensure the correctness of the summaries. Further, we make available 28,583 mono and cross-lingual article-summary pairs[*]. We also build and analyze multiple baselines on the collected data and report error analysis.

1 Introduction

In the field of Natural Language Processing (NLP), advancements in human language understanding and processing have often required extensive data. Cross-lingual summarization, a recent focus in NLP, has seen the emergence of well-known datasets like CrossSumm (Bhattacharjee et al., 2023) and PMIndiaSum (Urlana et al., 2023). In the context of Indian languages, recent datasets such as XL-Sum (Hasan et al., 2021), MassiveSumm (Varab and Schluter, 2021), and PMIndiaSum (Urlana et al., 2023) promote research in low-resource language setting.

Generating human-annotated datasets is a labor-intensive and costly endeavor (Urlana et al., 2022b). To address the scarcity of resources for low-resource languages, some efforts involve scraping news websites and use source document prefixes or headlines as summaries. However, this approach has two main shortcomings: 1) It often lacks diversity in information since all article-summary pairs originate from a single domain, and 2) Many source documents lack comprehensive coverage of specific events or incidents.

	En-En	**En-Hi**	**Hi-Hi**
Unfiltered pairs[*]	80868	210056	276599
Article Summary	78183	204975	274342
Deduplication	49116	133538	148421
Unigram Overlap[*]	25898	33054	68025
Cosine Similarity[*]	6340	7892	14351
Filtered pairs	**6340**	**7892**	**14351**

Table 1: Data filtering and statistics

In this work, we present the data collection and alignment approach aimed at enhancing both mono and cross-lingual summarization for Indian languages. Our core hypothesis is that news events can serve as a basis for aligning diverse sources of coverage in the context of summarization. Specifically, we concentrate on textual summarization and propose that by pairing up YouTube descriptions with corresponding news article coverage of different events from a multitude of sources, we can create a substantial summarization dataset comprising of 28,583 pairs. We have experimented with state-of-the-art multilingual summarization models and performed error analysis to assess the effectiveness of the pretrained models for both mono and cross-lingual summarization.

In the subsequent sections, we outline related work, our data collection methodology, the process of matching summarization pairs, and the subsequent filtering to derive the final dataset for our work in 1. Furthermore, we describe the training of various baseline models using the collected data and present our research findings.

[*] Equal Contribution

[*] https://github.com/tingc9/Cross-Sum-News-Aligned

2 Literature Survey

Significant progress has been made in the field of English language summarization, in stark contrast to the relatively limited efforts focused on Indian languages in the context of summarization and related Natural Language Generation (NLG) tasks (Urlana et al., 2022b,a, 2023), including headline generation. Nevertheless, recent times have witnessed a surge in active research in this domain, notably marked by the release of datasets like XL-Sum (Hasan et al., 2021), MassiveSumm (Varab and Schluter, 2021), and others. These multilingual datasets comprise pairs of articles and summaries extracted from publicly available news sources, encompassing a range of Indian languages, including Hindi, Gujarati, Bengali, and more.

Furthermore, the IndicNLG Suite (Kumar et al., 2022) has contributed by providing datasets designed for various Indic language NLG tasks, such as sentence summarization and headline generation. Despite these promising developments, there is still a need for continued effort in this area to attain performance levels in Indian language summarization that is on par with their English language counterpart.

3 Data Creation

There are three steps for data preparation - Data Collection, Pair Matching and lastly, Pair Filtering.

3.1 Data Collection

For the potential summaries, we crawl the language specific youtube channels of 4 sources - NDTV, IndianExpress, OneIndia and TheQuint. In this work, we focus on English and Hindi languages as a proof of concept. We used yt-dlp to crawl the descriptions of the above mentioned channels and populate our description store. For the potential documents, we crawl the respective language websites for each of the outlined sources using news-please (Hamborg et al., 2017). In total, we crawled 150K youtube descriptions and 350K news articles. We preprocess and compute the distUSE (Yang et al., 2019) vectors for the text and the extracted title of the respective articles and descriptions.

3.2 Pair Matching

The aggregated data needs to be aligned, resulting in pairs of articles and descriptions that include multiple pairs of document-summaries for a single event. A quick analysis of the data reveals that for the purposes of abstractive summarization, the fraction of common content words in the summary and the document must be high and the two texts must be semantically similar. We have three factors driving the pair matching 1

1. Date of incidence: The article and description must be published within 2.5 days of each other.

2. Unigram Overlap: The fraction of non stop word unigrams in the summary that are present in the document. For the cross-lingual case we perform a parallel dictionary lookup or transliterate, if the lookup fails.

3. Semantic Similarity: The cosine similarity of the DistUSE (Yang et al., 2019) text embeddings for the two texts.

Some samples are presented in Figure 2.

3.3 Pair Filtering

We overgenerate the pairs based on the above factors and employ the following mechanisms to filter out bad summaries.

1. Preprocessing: Deduplication and length-based outlier removal.

2. Compression Ratio: Summaries within one s.d. of the compression ratio are kept.

3. Unigram Overlap: Summaries with more than 0.4 for monolingual and 0.3 overlap for crosslingual case are kept.

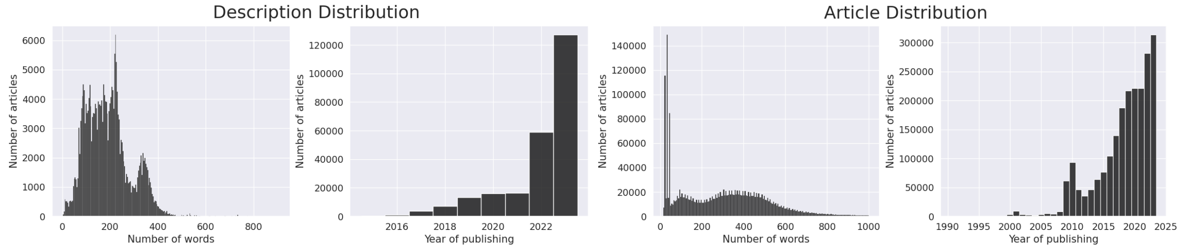
4. Similarity Threshold: Only pairs with both title and text similarity above 0.7 are kept.

After the whole process, we are left with 8K article summary pairs for cross lingual summarization and 14K pairs for monolingual summarization. We release the combined data of 28,583 article-summary pairs for en-en, en-hi and hi-hi language pairs.

[*] Only pairs with >= 0.5 similarity are selected

[*] 0.4 for mono-lingual summarization and 0.3 for cross-lingual summarization

[*] Pairs above 0.7 threshold were selected



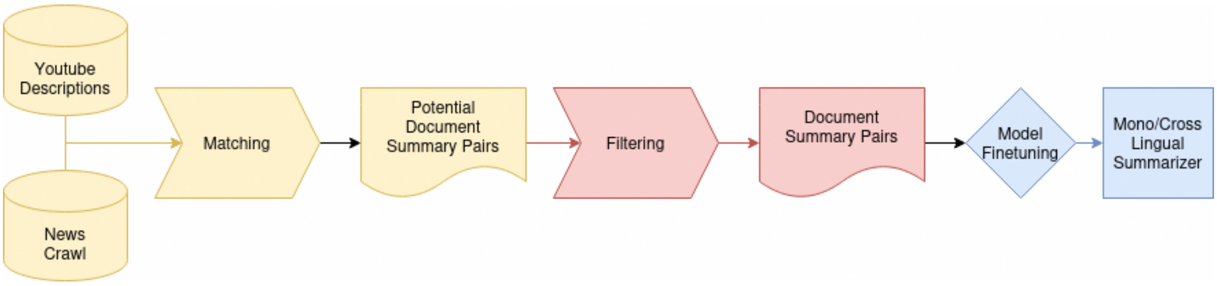


Figure 1: Overall architecture

4 Methodology

Our objective is to establish a benchmark for evaluating traditional summarization methods. Below, we present the following paradigms and specify the language settings utilized for testing with each approach.

4.1 Training free-baselines:

Extractive Method: In this approach, we incorporate two training free-baselines:

- Selection of the lead sentence.
- Scoring each sentence in the document against a reference and choosing the best one in an oracle-like manner.

	LEAD			EXT-ORACLE		
	R-1	R-2	R-L	R-1	R-2	R-L
en-en	43.2	26.3	34.9	50.1	33.1	41.7
hi-hi	10.4	4.1	9.6	24	11.3	21.9

Table 2: ROUGE scores of training free-baselines

Lang-pair	Train	Valid	Test
en-en	5072	634	634
en-hi	6314	789	789
hi-hi	11482	1435	1435

Table 3: Data splits counts

4.2 Pre-trained models

- **IndicBART** (Dabre et al., 2022) is a sequence-to-sequence model (244M parameters) trained on 11 Indic languages alongside English. Its training approach follows a masked span reconstruction objective, similar to MBart.
- **mBART:** BART (Lewis et al., 2020) is a denoising auto-encoder, the multilingual version mBART (Chipman et al., 2022) (610M parameters) designed for the pretraining of sequence-to-sequence models.
- **mT5:** The T5 (Raffel et al., 2020) employs an encoder-decoder Transformer architecture fine-tuned on the C4 corpus. We use the multilingual mT5 (Xue et al., 2021) model for finetuning on our data.

5 Experiments and results

We fine-tune on three pretrained models namely IndicBART, mBART and mT5-base. The dataset splits are detailed in the Table 3. We have used 80% for the training purpose and 10% each for the validation and testing. We report the average ROUGE(Lin, 2004) f1 scores in 4. We can see from Table 2, that for en-en samples, more than half the samples the summary is the same as the first sentence of the document. As mentioned in the Table 4, for all the three language pairs mBART-large50 model outperforms the remaining two models (IndicBART, mT5). Compared to IndicBART, mBART outperforms.

Document	Summary
Janata Dal (United) chief Nitish Kumar resigned from the post of chief minister on Tuesday, 9 August, ending his alliance with the Bharatiya Janata Party (BJP) in the state. Kumar, in a meeting held with party leaders in Patna, said that the BJP is conspiring to break the JD(U), sources said. Following Kumar's submission of his resignation to Governor Phagu Chauhan, the duo of Kumar and Rashtriya Janata Dal (RJD) leader Tejashwi Yadav staked a claim to the government. Amid the political turmoil, here are the top developments from Bihar:	Nitish Kumar & RJD leader Tejashwi Yadav left for the residence of Rabri Devi in Patna. Kumar resigned from his position as Chief Minister on August 9.
The massive fire that engulfed Assam's Baghjan oil well two weeks after a blowout in the oil field in Tinsukia district has spread to nearby villages. At least 30 houses have been damaged so far, and locals have reported loss of flora and fauna as well. Two people have also been killed in the massive fire The Baghjan Oil Well has been blowing out since 27 May. What started as an oil spill and gas leak due to the failure of pressure control systems, has now turned into a massive fire. The fire rapidly spread to nearby villages and set ablaze several houses and other properties. Close to 30-35 houses near the Baghjan oil field have been destroyed in the fire, which is visible 4-5 km away from the site.	The Baghjan Oil Well located near the Dibru Saikhowa National Park in Assam has been blowing out since 27 May. What started as an oil spill and gas leak due to the failure of pressure control systems, has now turned into a massive fire.
After being denied bail multiple times, the Bombay High Court granted bail to Aryan Khan, Arbaaz Merchantt and Munmun Dhamecha in the Mumbai cruise drugs case on Thursday, 28 October. The court said it will release a detailed order on Friday, 29 October. The three of them were arrested by the Narcotics Control Bureau (NCB) following a raid onboard a luxury cruise on 2 October. They were sent to judicial custody on 8 October. Here are the arguments made by NCB and Aryan Khan in court on Thursday.	मुंबई ड्रग्स मामले में आर्यन खान (Aryan Khan) को आखिरकार बड़ी राहत मिली है. बॉम्बे हाईकोर्ट ने आर्यन को जमानत दे दी है. पिछले तीन दिनों से जमानत याचिका पर सुनवाई चल रही थी. आर्यन खान के अलावा अरबाज मर्जेंट और मुनमुन धमेचा को भी हाईकोर्ट ने जमानत दी है. इन तीनों को एक साथ एनसीबी ने गिरफ्तार किया था.
With the beginning of the new year, team India is all set to play their first game against Sri Lanka. The Sri Lankan cricket team will be visiting India for the shortest format accompanied by a three-series match. The first game will be played in the iconic Wankhede stadium in Mumbai. This will be the first time that Hardik Pandya will be leading the team in the absence of Virat Kohli, KL Rahul, and Rohit Sharma. Hardik Pandya is expected to begin his full-time T20 captaincy tenure on a bold note in India vs Sri Lanka T20I 2023. The three-match series against Sri Lanka will begin on Tuesday, 3 January 2023. The followers of Indian cricket have gotten a vision of Hardik's captaincy from his victory in the rain-hit T20 series in New Zealand. Now, let's know the timings, venue, and live streaming details like when and where to watch India vs Sri Lanka T20I match?	भारत बनाम श्रीलंका (India vs Sri Lanka) के बीच तीन मैचों की सीरीज का पहला टी20 मुकाबला मुंबई (Mumbai) के वानखेड़े (Wankhede) में खेला जाएगा. ये सीरीज हार्दिक पांड्या (Hardik Pandya) के लिए बतौर कप्तान काफी जरूरी होने वाली है. वहीं, टीम इंडिया (Team India) भी यही चाहेगी की साल की शुरूआत एक यादगार जीत के साथ ही की जाए.
महिला हॉकी खिलाड़ियों के लिए डिप्टी सीएम दुष्यंत चौटाला ने भेजी 50 किट, कहा- नाम करेंगी रौशन पुलिस के सामने से डायल-112 लेकर फरार हुआ चोर, 10 km दौड़ाया, फिर चाबी फेंक कर हो गया फुर्र छात्रसंघ चुनाव के लिए Digvijay Chautala ने शुरू किए यूनिवर्सिटी और कॉलेजों के दौरे, कही ये खास बात हरियाणा के युवाओं के लिए उपलब्ध करवाए 75 प्रतिशत रोजगार, सीएम खट्टर ने कहा हरियाणा में सीएम पद पर दावे को लेकर अभी से भिड़ गए कांग्रेसी, चुनाव से साल भर पहले क्यों शुरू हुआ विवाद? जानिए हरियाणा से सटे राजस्थान के जिलों में नेताओं का जमावड़ा, उम्मीदवारों को जीताने के लिए कर रहे धुआंधार प्रचार TV-9 Cicero exit poll results 2019: हरियाणा में भाजपा को भारी बहुमत, कांग्रेस-इनेलो का पत्ता साफ Haryana oi-Akarsh Shukla नई दिल्ली। हरियाणा विधानसभा चुनाव के लिए मतदान आज संपन्न हो गया है। राज्य की जनता बढ़-चढ़कर वोटिंग में हिस्सा लिया और दिन के अंत तक 53.78 फीसदी मतदान हुआ। हरियाणा की 90 विधानसभा सीटों के लिए एग्जिट पोल्स के अनुमान सामने आ चुके हैं। ज्यादातर एग्जिट पोल्स में हरियाणा में बीजेपी की सरकार बनती दिख रही है। साफ शब्दों में कहें तो एग्जिट पोल्स के मुताबिक, हरियाणा में फिर से मनोहर लाल खट्टर की सरकार बनते हुए दिखाई दे रही है। एग्जिट पोल्स के मुताबिक राज्य में भाजपा को 69 सीटें तो कांग्रेस को 11 सीटें मिल रही हैं वहीं, अन्य पार्टियां 10 सीट पर ही सिमट गई हैं। हरियाणा विधानसभा चुनाव 2019 में मुख्यमंत्री मनोहरलाल खट्टर, कांग्रेस नेता भूपेंद्र सिंह हुड्डा, बबीता फोगाट, योगेश्वर दत्त, सोनाली फोगाट समेत कई दिग्गजों की प्रतिष्ठा दांव पर है। उल्लेखनीय है कि एग्जिट पोल सिर्फ एक अनुमान हैं, ये गलत भी हो सकते हैं। सही परिणाम 24 अक्टूबर को होने वाली मतगणना के दिन ही पता चलेगा। बता दें, 2014 में हुए हरियाणा विधानसभा चुनाव में भाजपा ने पहली बार राज्य में बहुमत हासिल किया था। भाजपा ने दस साल से सत्ता पर काबिज कांग्रेस को सत्ता से बाहर किया था। 2014 में हरियाणा की 90 सीटों में 47 पर भाजपा को जीत मिली थी, जबकि इंडियन नेशनल लोकदल को 19 और कांग्रेस को 15 सीटों पर जीत मिली थी। अन्य को 9 सीटें मिली थी। Oneindia की ब्रेकिंग न्यूज़ पाने के लिए . पाएं न्यूज़ अपडेट्स पूरे दिन. Allow Notifications	हरियाणा और महाराष्ट्र विधानसभा चुनावों के लिए वोटिंग खत्म हो चुकी है. नतीजों से पहले अलग-अलग न्यूज चैनल्स और एजेंसियों के एग्जिट पोल सामने आने शुरू हो चुके हैं. एबीपी न्यूज के एग्जिट पोल के मुताबिक, महाराष्ट्र में बीजेपी की सरकार एक बार फिर से बनती नजर आ रही है. पोल में बीजेपी गठबंधन को 204 सीट, कांग्रेस गठबंधन को 69 सीटों का अनुमान जताया गया है. अन्य को 15 सीटें मिलती दिख रही हैं.

Figure 2: Summary examples

	IndicBART			**mBART**			**mT5-base**		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
en-en	46.7	28.1	36.8	**49**	**31**	**40.1**	42.4	24.9	34.1
en-hi	20	7.8	14.8	**25.7**	**10.7**	**17.6**	18	5.1	13.1
hi-hi	28.2	15	21.5	**40.7**	**26**	**32.2**	30.6	15	22.3

Table 4: Multi-lingual baselines ROUGE scores

Error Types	**en-en**			**hi-hi**			**en-hi**		
	IndicBART	mBART	mT5-base	IndicBART	mBART	mT5-base	IndicBART	mBART	mT5-base
Comprehensibility	0	0	0	3	1	0	12	0	1
Grammar	29	2	0	21	14	19	2	1	0
Factuality	6	6	6	10	11	12	20	36	28
Omission	15	18	29	20	11	8	20	23	12
Redundancy	10	13	6	7	5	17	17	3	34
No error	8	22	16	4	16	6	0	1	0

Table 5: Error analysis on different models and various language combinations

Parameters	**mBART**	**mT5**	**IndicBART**
Max source length	512	512	512
Max target length	128	128	128
Batch Size	2	1	4
Epochs	5	5	20
Vocab Size	50265	32128	64015
Beam Size	4	4	4
Learning Rate	5e-5	5e-5	5e-5

Table 6: Experimental setup and parameters settings

5.1 Error analysis

We performed the error analysis by following the guidelines mentioned in the PMIndiaSum (Urlana et al., 2023). In case of English and Hindi monolingual summarization, IndicBART makes lot of grammatical errors by omitting the relevant information. For cross-lingual (En-Hi) summarization, we have observed that all three models make faithfulness errors by omitting the relevant information. Overall, we find that mBART performs better in case of both mono and cross-lingual summarization methods.

6 Conclusions

In this study, we have developed a mono and cross-lingual summarization dataset for English-Hindi language pairs by leveraging a variety of news events and their corresponding YouTube descriptions. Our experimentation with multilingual models revealed that mBART consistently outperforms the IndicBART and mT5 models. Our error analysis indicates significant opportunities for the development of more effective multilingual models for low-resource languages.

7 Limitations

The dataset is scraped from various news sources, there is a possibility of code-mixed samples in the document-summary pairs. Moreover, because of slight differences in the content of each source, not all summaries are equal.

8 Future Work

It is likely that for other Indian Languages, the collected data from this method would be a fraction so it would be worth looking into training a single model for cross lingual summarization where the parameter sharing could benefit low resource languages.

References

Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, Yuan-Fang Li, Yong-Bin Kang, and Rifat Shahriyar. 2023. CrossSum: Beyond English-centric cross-lingual summarization for 1,500+ language pairs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2541–2564, Toronto, Canada. Association for Computational Linguistics.

Hugh A Chipman, Edward I George, Robert E McCulloch, and Thomas S Shively. 2022. mbart: multidimensional monotone bart. *Bayesian Analysis*, 17(2):515–544.

Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh Khapra, and Pratyush Kumar. 2022. IndicBART: A pre-trained model for indic natural language generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1849–1863, Dublin, Ireland. Association for Computational Linguistics.

Felix Hamborg, Norman Meuschke, Corinna Breitinger, and Bela Gipp. 2017. news-please: A generic news crawler and extractor. *Proceedings of the 15th International Symposium of Information Science*.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Aman Kumar, Himani Shrotriya, Prachi Sahu, Amogh Mishra, Raj Dabre, Ratish Puduppully, Anoop Kunchukuttan, Mitesh M. Khapra, and Pratyush Kumar. 2022. IndicNLG benchmark: Multilingual datasets for diverse NLG tasks in Indic languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5363–5394, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Ashok Urlana, Sahil Manoj Bhatt, Nirmal Surange, and Manish Shrivastava. 2022a. Indian language summarization using pretrained sequence-to-sequence models.

Ashok Urlana, Pinzhen Chen, Zheng Zhao, Shay B Cohen, Manish Shrivastava, and Barry Haddow. 2023. Pmindiasum: Multilingual and cross-lingual headline summarization for languages in india. *arXiv preprint arXiv:2305.08828*.

Ashok Urlana, Nirmal Surange, Pavan Baswani, Priyanka Ravva, and Manish Shrivastava. 2022b. TeSum: Human-generated abstractive summarization corpus for Telugu. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5712–5722, Marseille, France. European Language Resources Association.

Daniel Varab and Natalie Schluter. 2021. MassiveSumm: a very large-scale, very multilingual, news summarisation dataset. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10150–10161, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernández Ábrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2019. Multilingual universal sentence encoder for semantic retrieval. *CoRR*, abs/1907.04307.