Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 5712–5722
Marseille, 20-25 June 2022

TeSum: Human-Generated Abstractive Summarization Corpus for Telugu

Ashok Urlana*, **Nirmal Surange***, **Pavan Baswani, Priyanka Ravva, Manish Shrivastava**
Language Technologies Research Center, KCIS, IIIT Hyderabad,
India
{ashok.urlana,nirmal.surange,pavan.baswani,priyanka.ravva}@research.iiit.ac.in,m.shrivastava@iiit.ac.in

Abstract

Expert human annotation for summarization is definitely an expensive task, and can not be done on huge scales. But with this work, we show that even with a crowd sourced summary generation approach, quality can be controlled by aggressive expert informed filtering and sampling-based human evaluation. We propose a pipeline that crowd-sources summarization data and then aggressively filters the content via: automatic and partial expert evaluation. Using this pipeline we create a high-quality **Te**lugu Abstractive **Sum**marization dataset (**TeSum**) which we validate with sampling-based human evaluation. We also provide baseline numbers for various models commonly used for summarization. A number of recently released datasets for summarization, scraped the web-content relying on the assumption that summary is made available with the article by the publishers. While this assumption holds for multiple resources (or news-sites) in English, it should not be generalised across languages without thorough analysis and verification. Our analysis clearly shows that this assumption does not hold true for most Indian language news resources. We show that our proposed filtration pipeline can even be applied to these large-scale scraped datasets to extract better quality article-summary pairs.

Keywords: Summarization, Abstractive Summarization, Telugu Dataset, Low Resource Languages

1. Introduction

Summarization is a task that has held the interest of the NLP community since the beginning of the computational processing of languages. In NLP literature going back to the early years of NLP, we find a large number of different goals being described as the task of summarization. The DUC (2003-2007) guidelines define summarization as a task of generating a very short text which gives a general idea about the source article. Earlier, Hovy and Lin (1998) asked the question: What is a summary precisely? They proceed to answer the question as :

A summary is a text that is produced out of one or more (possibly multimedia) texts, that contains (some of) the same information of the original text(s), and that is no longer than half of the original text(s).

Hovy and Lin (1998) and KS Jones (1998) set out to define the task of summarization as something more than a small piece of the text providing an indication of the content of the source article. In fact, Hovy and Lin (1998) follow and extend (Jones, 1998) to provide fine categorization of summary types based on the broader aspects of input, purpose, and output.

In recent years, much work has been done to advance the state of the art of summarization for multiple languages across the world. But, most of these works adhere more closely to the DUC summarization challenge rather than the nuanced definitions presented by Hovy and Lin (1998). We find that collection of summarization data is reduced to mass scraping of various news sources across the world, in order to source the articles from the real world. At the same time, the expensive task of summary creation is reduced to clever partitioning of the content already available on online news media . In Hermann (2015), we first saw, the usage of news articles along with their highlights available on certain reputed news outlets for the purpose of cloze-kind question answering. This data was later repurposed for the summarization task by using the highlights as a proxy to the summary itself, trusting an implicit assumption which is based on strict editorial policies implemented by some news publications.

The underlying assumption is that these highlights or bullet points preceding an article are editorially required to convey a broad idea about the content of the article. While this assumption is an inspired one, and makes sense for a large number of news sources with strict editorial quality control, unfortunately, the assumption cannot be blindly extended to a vast majority of news sources. And it can be easily found that this assumption fails even for a large number of English news summarization datasets. Even so, many recent works have followed this strategy for creating massive summarization data sets fit for the training of deep neural networks. There are other problems as well. While such an approach might give us some "summary", one cannot guarantee if 1) the summary is abstractive, unless explicitly measured for it, and 2) the summary is coherent.

We look at a particular Indian language, Telugu, for which such datasets are available as XL-Sum (2021) and MassiveSumm (2021) which have been collected from various sources. Both rely on above mentioned

* Authors contributed equally

assumptions. Even with a casual observation, we find that these assumptions, and therefore these datasets do not stand up to the test. Therefore, we find that there is a urgent and immediate need for a dataset and a dataset creation methodology which stays true to the essence of the task of summarization as defined in (Hovy and Lin, 1998). Such a dataset must be created with human involvement and thorough evaluation. While we acknowledge that a completely human summary creation task might be unacceptably expensive, we propose a methodology which ensures high quality dataset without depending on very high degree of human involvement. In such tasks, creation and evaluation both are expensive processes, with evaluation often costing more than the creation itself. With this in mind we propose a Human-generated summary creation pipeline.

We propose a combination of automated and human evaluations to ensure a high quality dataset for Telugu (can be extended to other languages). We present, TeSum, a Human-Generated, curated Abstractive Summarization data set for Telugu[1] (Table 1).

We compare the resultant dataset with the existing datasets for the language and show that in the light of some well-motivated criteria, both XL-Sum[2](Telugu) and MassiveSumm [3](Telugu) do not live up to the expectations.

2. Related Work

In recent years, many automatic abstractive summarization datasets have been proposed. Initial inspiration for many of these came from the DUC tasks (2004) of generating a summary, for one or more given articles, in response to the given topic of the article. Nallapati et al. (2016) took a step ahead and proposed CNN/DM dataset. They followed the work of Hermann (2015) on creating large-scale data for reading comprehension tasks. This dataset was soon followed by Gigaword (2015), Newsroom (2018) and XSUM (2018). These monolingual datasets led the way to multilingual datasets in the form of XL-Sum, MassiveSumm, etc. While this approach of scraping news and highlights does lead to large numbers which are suitable for training large deep learning models, it is safe to assume that if the quality of the data is not up to the mark, the model outputs would also suffer.

For Telugu,we evaluated XL-Sum and MassiveSumm datasets, and came to the conclusion that the dataset qualities could not be considered as human summarization, therefore we set up for a task of creating human generated and curated summarization dataset for Telugu. We find that the curation policies can even be extended to scraped data such as XL-Sum and MassiveSumm.

3. Crowd-Sourced Corpus Creation

We propose a crowd sourced summary creation phase followed by a curation phase by trained experts. We work with 347 "creators" and 3 expert "raters" for this task. The "creators" are provided with specific guidelines to ensure the quality of generated content. The content is then aggressively filtered to retain high quality article-summary pairs. Human experts then evaluate a subset of the collected article-summary pairs to remove substandard tuples.

3.1. Source

We scrape Telugu news sites for source articles, under fair usage policy and divide them into sets of 50 articles each. The copyright of the original articles remains with the original authors/publishers of the articles. We release TeSum dataset as a list of URLs and summary pairs. These articles are then processed to remove HTML tags, non-Telugu content and common irrelevant phrases (article dates, city names etc.). The dataset is released at a later date.

3.2. Manual Summarization

The human summarization task is posed as a crowdsourcing activity. Each HIT (Human Intensive Task) for a creator consists of 50 news articles set created earlier. The creator is expected to create summaries for all the articles in the HIT following the guidelines given below. Each HIT submitted by the creator undergoes thorough automatic and human evaluation steps in order to ensure quality based on a criteria which maintains the essence of the task of summarization. The creator needs to ensure that:

1. Relevance: All or most of the relevant information contained in the article should be present in the summary.

2. Readability and Coherence: The summary should be coherent, readable and free of any grammatical errors.

3. Creativity: The summary should have novel sentential and phrasal structures.

The human summary creators were given the guidelines presented in Section 3.2.1 based on the above 3 properties.

3.2.1. Guidelines for Abstractive Summary Creation

Summary creators were instructed to carefully follow these guidelines and write one abstractive summary per article.

1. **Relevance and Coverage:** All the pertinent information conveyed in the source article should be

[1] https://github.com/manshri/TeSum/

[2] XL-Sum data is taken from the publicly available repository at https://github.com/csebuetnlp/xl-sum

[3] We thank the authors of MassiveSumm for graciously providing us with the entire MassiveSumm datasets, for our experiments.

	Train		Validation		Test	
# Pairs	16295		2017		2017	
Avg Compression%	58.26		58.08		58.28	
	Text	Summary	Text	Summary	Text	Summary
# Unique Words	183641	113723	49038	28873	49620	28777
Avg Unique Words	88.93	42.78	89.19	43.14	90.74	43.81
(Min, Max) Words	(30, 536)	(12, 213)	(32, 685)	(10, 248)	(36, 592)	(12, 261)
Avg Words	120.8	50.02	122.56	50.8	124.82	51.69
Avg Sentences	9.23	3.22	9.50	3.19	9.51	3.17

Table 1: TeSum Statistics

captured in summary while discarding any irrelevant information. Redundant information or information unrelated to the major topic of the article may be considered irrelevant.

- **Missing important information:** A summary has to cover all the important aspects of the original article.
- **Including irrelevant information:** A summary should not include any irrelevant information. No personal opinion(s) or non-factual details should be included.
- **Redundant information:** Summary should not contain any repetitive phrases/sentences.

2. **Readability:** If the summary is understandable by a native speaker without looking at the source article, it is considered "Readable". Bad grammar, pronouns that cannot be resolved within the summary, and unnatural sentential/phrasal structures would make the summary difficult to understand. Also, creators are instructed that the summary should stand as an independent article, and the reader should not need the original article to understand it fully.

   - **Disjoint sentences:** While paraphrasing, sentences should be joined in such a way that the composite sentence must be meaningful.
   - **Anaphora issue:** In summary, pronouns should be used only after the antecedent has appeared at least once.
   - **Disordering of sentences:** The summary should be coherent to convey the proper context of the original article.
   - **Not readable:** The summary should be free from any syntactic and semantic errors.

3. **Creativity:** Since this is an abstractive summarization task, we require the summaries to have novelty in terms of sentential structures such as lexical choices (vocabulary used, is other than the given article), phrasal constructions, and sentence formations.

   - **Missing novel sentence structure:** The summary should contain novel sentence structures (using some novel words) compared to the original article.
   - **Lengthy summary:** The summary should be a new shorter text that conveys the most crucial information of the original article.
   - **Sentence level summary:** The summary should not be created by just altering words/phrases in individual sentences.

4. Corpus Curation Process

As expected with any crowd-sourced text annotation task, the summaries generated by the creators had a wide variety of errors. As shown in the guidelines, we have not instructed the creators to create their summaries within a pre-specified character or word limit. This is done to ensure that the creators do not feel restricted while writing the summaries in order to fit within a pre-specified limit. An artificial limit to the length of the summary at the creation phase might introduce unnatural structures or phrasing in the sentences of summaries. Instead, we decide to filter these generated summaries on the basis of multiple automated criteria after the summaries have already been created. The crowdsourcing activity resulted in a collection of 92941 article-summary pairs. These articles were then filtered based on the following two stages.

4.1. Automatic Filtering

Even with basic sanity checks at the crowdsourcing stage, we encounter a large number of errors in these submitted HITs. These article summary pairs need to be filtered out in order to maintain the high quality of the dataset.

1. *Remove Empty:* We remove any pairs where either the summary/article or both are empty.

2. *Remove Duplicates:* Duplicate pairs and duplicate summaries were removed. We do not want duplicate article-summary pairs. Two distinct articles should not have the same summary. We find it is unlikely that two distinct articles would share the

	XL-Sum	MassiveSumm	TeSum
Dataset Size	**13025**	**119282**	**92941**
Empty	2	5579	2
Duplicate Pairs	0	10456	515
Duplicate Summary	141	2698	135
Prefixes	3	30741	1330
Article $<$ 4 Sentences	10	4953	4195
Article $<$ 40 Tokens **Summary $<$ 10 Tokens**	374	5446	10
Compression $<$ 50%	10	1641	52802
Compression $>$ 80%	11920	46776	456
Abstractivity $<$ 10	0	6683	5942
Abstractivity $>$ 80	227	303	42
Human-Eval(TeSum)	-	-	7183
Final Valid	**338**	**4006**	**20329**
Valid %	**2.6%**	**3.36%**	**21.9%**
Avg Abstractivity scores	68.23	36.44	31.08
Avg Compression(%)	71.71%	73.34%	58.24%

Table 2: Pre-processing and Filtration Details. Here, the bottom 2 rows show the average abstractivity scores and average compression% for the final valid pairs of the 3 datasets.

same summary, therefore we also remove the pairs which share a common summary.

3. *Remove Prefixes:* We remove all prefix cases, that is any pair where the summary is just the first few sentences of the article. We should note that though MassiveSumm has claimed to follow steps $1-3$, we find in Table 2 that applying these steps to MassiveSumm, a large volume of their samples still fell in to these categories. Duplicate summaries case holds true for XL-Sum also.

4. *Remove Article Length $<$ 4 Sentences:* We removed 4452 pairs with less than 4 sentence articles.

5. *Remove Article Length $<$ 40 Tokens and / or Summary Length $<$ 10 Tokens:* Very small article lengths are not indicative of the general distribution of news article data.

4.2. Automatic Quality Control

- *Compression ranges:* If an article is compressed (Bommasani and Cardie, 2020) too much then we loose significant amount of information from the article, which contradicts the first property of summarization that all/most of the relevant information of article must be present in the summary. Though, a summary should also result in a significant amount of reduction in the size of the article but not at the cost of relevance. Therefore, we set compression% limits to be between 50-80. While the upper limit is higher than many previous datasets, (which, usually, set this to 30%) we find, particularly in news domain which is information dense, there can be large number of examples where slightly more content is required in the summary. Figure 1 shows the article counts of TeSum for compression% ranges.

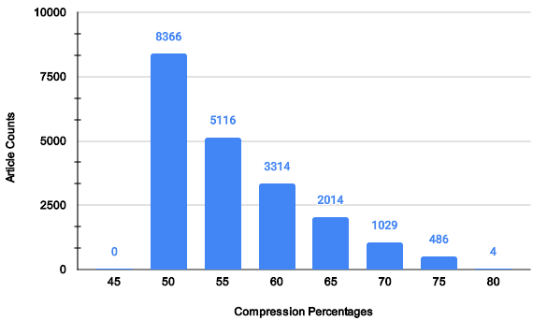

Figure 1: Article Count Vs. Compression

- *Abstractivity ranges:* We want novelty in the summary, the content should be different from the source on both sentential as well as phrasal levels. We often find that even with the best editorial practices, the content in the highlights is often a conjunction of multiple disjoint phrases or absolute copy of phrases from the article. Which apart from being non-coherent, beats the third property of creativity. Therefore, we take the measure of abstractivity from (Bommasani and Cardie, 2020) and apply 10-80 range of filtration. Even though we want the summary to be abstractive, we still need to copy some n-grams from the article which corresponds to factual information (names etc.) as presented in the article. Therefore, we restrict Abstractivity at 80 which is still a fairly lenient limit.

  On the lower side, for shorter articles which are information dense, it is possible that the copied unigrams or bigrams will constitute a large chunk of the summary in order to retain facts. This is especially true in the news domain where the summary creator has to copy smaller n-grams but would change the phrasal structure for novelty (paraphrasing). If abstractivity as proposed by Bommasani and Cardie (2020) goes very close to zero then we get very high degree of copying in higher n-grams also. And on the other side, if the abstractivity is very high then we loose important information in terms of verbs, nouns, etc also which need to be there.

At this stage, after filtering by compression and abstractivity, we are left with 27512 article-summary pairs. Table 2 shows the number of article-summary pairs getting affected by each filter.

	Relevance	Readability	Creativity
Score 0	0 - 10% relevant information	Not understandable	Copied verbatim from the 'original' article
Score 1	10 - 40% relevant information	Largely ungrammatical	Most of the sentences copied from the 'original article'
Score 2	40 - 60% relevant information	Approx 50% ungrammatical	Half the summary is copied from the original article
Score 3	60 - 90% relevant information	Minor grammatical error	Most of the summary is novel, but some is copied verbatim
Score 4	Everything is relevant and all the relevant information is covered	Free from any grammatical, spelling and punctuation errors.	The entire summary, except the factual information (names, dates etc.), is novel

Table 3: Abstractive Summarization Evaluation Criteria

5. Human Evaluation

To maintain quality, one has to ensure that the human summarization guidelines are well understood by the creators and creators are, by an large, sticking to the guidelines. Though, it is impossible in any such task to have all the submissions manually evaluated, if a reasonable percentage of all the submissions are evaluated and found to be of high quality, it can be safely assumed that the rest of the submissions are also of high quality. Over the course of large number of evaluations, the expected percentage of lower quality samples in the total data can be estimated.

For human evaluation, the raters were asked to rate a minimum of 25% of the pairs from each HIT, for the 3 parameters *Relevance*, *Readability* and *Creativity* as per the Table 3. Each rater is supposed to rate a sample by giving scores, ranging between 0 to 4, for each parameter.

5.1. Special Cases

- If all the sentences are copied verbatim from the original article, scores are [0 0 0] for Relevance, Readability, and Creativity.
- In case of syntactic errors (spelling, spacing, punctuation), if that particular word/phrase deviates the overall meaning/context significantly, then scores will be deducted in Readability as well as Relevance.
- In case of tense issues, simultaneously, the scores can be reduced in Creativity and Relevance.
- The addition of irrelevant information or outside the context of the article leads to obtaining less scores in Creativity and Relevance.
- For anaphora-related issues, both Readability and Creativity scores will be reduced.
- Improper usage of novel words/phrases causes a reduction in Creativity score. If that particular word/phrase deviates from the original article's meaning, there will also be a reduction in the Relevance score.

5.2. Inter Rater Reliability:

The inter-rater-reliability was established by following the guidelines (as mentioned in section 5). We ramdomly extracted 500 samples from the total collected articles. These 500 samples were then rated by 3 expert raters to compute the ICC3 scores. The agreement scores were then computed using the Intra-class Correlation Coefficient (ICC) (Shrout and Fleiss, 1979) following the guidelines given by Koo and Li (2016). We report **ICC3** scores, which correspond to fixed raters and individual (single) reliability. We specifically chose this model (ICC3), because each sampled article-summary pair, from the HITs, is then evaluated by one rater, and not all 3.

For our three parameters: Relevance, Readability and Creativity, our raters achieved **0.89**, **0.94** and **0.90** reliability scores respectively. These scores indicate good to excellent reliability.

5.3. Human Evaluation Process

Each HIT was evaluated by one rater, by randomly selecting a minimum of 25% from the HIT and distributing among the 3 raters, such that each pair of this 25% was evaluated by a single rater. If on an average the combination of these 25% pairs do not rate 3 or above for each individual parameter, then the entire HIT is rejected based on the assumption that there is a higher percentage of low quality submissions in this HIT. This process resulted in a total reduction of 7183 pairs. Giving us the final 20329 pairs.

Since we are evaluating only a percentage of the samples submitted in each HIT, we need to be aware of the possibility of some errors in the final dataset. To estimate this, we take the 5089 evaluated samples (25% of the final 20329) and find individual samples which have lower scores. These were found to be 3.6%. As, 25% is a fair enough sample size, we can safely extend the same error percentages to the entire dataset. Therefore resulting in a dataset which, while being smaller than other existing datasets, is of high quality. But if we subject the existing datasets to the same high standards that we expect from our dataset, we find that our dataset size is not low at all, in comparison with their resultant dataset sizes.

	Avg Scores			# Samples >= 3		
	XL-Sum	**MassiveSumm[Te]**	**TeSum**	**XL-Sum**	**MassiveSumm[Te]**	**TeSum**
Relevance	1	2	3	4	43	185
Readability	3.5	2.9	3.27	176	144	188
Creativity	0.98	1.58	3.28	12	51	170
All 3 parameters rated >= 3				**4**	**35**	**154**

Table 4: Human Evaluation of XL-Sum[Te], MassiveSumm[Te] and TeSum on 200 samples each.

6. Evaluating Existing Datasets

For comparison, when we applied the filters mentioned in Section 4, to the existing datasets MassiveSumm and XL-Sum, we found some surprising results. As mentioned in the Table 2, MassiveSumm ended up with only 3.36% of their original dataset size. Similarly, XL-Sum also reduced to only 2.6% of their original size. Even if we relax the constraints a little bit, it does not help their end results much. We will be releasing all the filtering scripts along with the lists of IDs/URLs for basic problem cases from both the datasets.

Human Evaluation of MassiveSumm and XL-Sum: Due to surprising final numbers of the XL-Sum and MassiveSumm datasets after the filtration, we decided to validate this finding by manually evaluating randomly selected 200 article-summary pairs from each of the 3 datasets using the same raters. All samples were completely annonymized/randomized in order to avoid dataset bias. We found that the summaries are of low quality for XL-Sum and MassiveSumm on almost all parameters, Table 4 shows the average numbers obtained by each dataset for each parameter individually. Also, the counts of article-summary pairs from each dataset which gained 3 or above ratings are shown. The bottom line shows final count of valid pairs (out of the 200) for each dataset which were rated 3 or above, for all the 3 parameters.

Other Languages: As the numbers on the existing datasets were too surprising, we wondered if it was for this particular language. Therefore, we extended our analysis to some other languages (Hindi, Gujarati and Marathi from both XL-Sum and MassiveSumm) that we could evaluate (could read). We found similar issues in all of these datasets. We show the detailed filtration counts in Table 5. We also show examples of some of the common problem cases (from the respective datasets) in the Appendix- A.

7. Baseline Models

We present some common baselines used for summarization by other authors, to demonstrate the impact of the datasets on summarization using various models.

7.1. Models

In order to show the proof of quality of TeSum dataset on the summarization task itself, and to provide various baselines, we trained and tested several existing summarization models with TeSum data.

Pointer-Generator (PG): This model is implemented using sequence-to-sequence Recurrent Neural Networks (RNN) (Sutskever et al., 2014) with attention mechanism (Bahdanau et al., 2014). Further, we also implemented the pointer-generator(See et al., 2017) with coverage mechanism model. Pointer-generator mechanism helps in deciding whether to copy words from the source text or to generate from the vocabulary. Hence, it effectively handles the Out Of Vocabulary (OOV) words problem. The coverage mechanism, prevents the model from attending to the same phrases multiple times, which helps in handling the redundancy issue in summary generation.

MLE+RL, with intra-attention: This model is implemented using the intra-attention mechanism proposed by Paulus (2017), that attends over the input document and continuously generates decoder output separately to reduce the problem of repetitive and incoherent phrases in the summaries. They further introduced a new training method that combined with supervised and Reinforcement Learning (RL) prevents from exposure bias problems and can produce readable summaries.

Text summarization with Pretrained Encoders (BertSumAbs): This model is based on the novel document-level encoder by Liu and Lapata (2019) which uses Bidirectional Encoder Representations from Transformers (BERT). For abstractive summarization, this method adopts the encoder-decoder architecture with a new fine-tuning approach where the encoder is a pre-trained BERT and the decoder is a randomly initialized Transformer. For this model, we have used the embeddings by Marreddy et al. (2021) (trained on 8M+ Telugu sentences).

mT5: We fine-tuned the Multi-lingual Text To Text Transfer Transformer (mT5) model by Xue et al. (2020) on TeSum dataset. This model is a multi-lingual variant of the T5 (Raffel et al., 2019) model trained on common crawl English dataset. We have used mT5-small for our experiments.

7.2. Experimental Setup

To create train, dev and test splits of TeSum dataset, we divide the total 20329 pairs into carefully selected subparts of about 80%, 10% and 10% respectively. The selection of pairs is done in a way that preserves the balance in terms of length of articles, compression(%)

	HINDI		MARATHI		GUJARATI	
	XL-Sum	**MassiveSumm**	**XL-Sum**	**MassiveSumm**	**XL-Sum**	**MassiveSumm**
Dataset Size	**88472**	**563477**	**13627**	**127838**	**11397**	**43830**
Empty	5	20936	1	1488	0	3797
Duplicate Pairs	4	48461	0	614	1	525
Duplicate Summary	698	5626	465	4507	59	878
Prefixes	19	4225	3	4015	6	99
Article $<$ 4 Sentences	164	27845	6	6811	104	5307
Article $<$ 40 Tokens Summary $<$ 10 Tokens	377	125372	154	60489	97	14303
Compression $<$ 50%	13	1990	6	145	5	163
Compression $>$ 80%	85028	286696	11985	47659	10611	18481
Abstractivity $<$ 10	4	10668	1	843	0	55
Abstractivity $>$ 80	29	643	411	128	91	11
Final Valid	**2131**	**31015**	**595**	**1139**	**423**	**211**
Valid %	**2.4%**	**5.5%**	**4.37%**	**0.89%**	**3.71%**	**0.48%**

Table 5: Filtration counts of XL-Sum and MassiveSumm for the other 3 languages; Hindi, Marathi and Gujarati.

and abstractivity levels across the three splits. Table 1 details the statistics for the 3 splits.

For the experiments and baseline training, we have used Word2Vec (Mikolov et al., 2013) (Telugu Wikipedia pre-trained) embeddings. Apart from mT5, which was fine-tunned using 2 GPUs and 20 CPUs, the rest had system config of 1 GPU and 10 CPUs. Further details on hyper-parameter settings and configuration is listed in Table 6. Here, 'PG+' represents PG and PG+Coverage models, and 'MLE+' represents MLE, MLE+RL and RL models.

Necessary Concessions: As, after our filtration steps, the originally large-scale existing-datasets ended up with a very low percentage of their total article-summary pairs. Which extrinsically does not make for a fair comparison. Therefore, before going ahead with the model training and experiments, for evaluating the effect of these curations of the datasets for the task of summarization, we are forced to make some concessions for XL-Sum and MassiveSumm.

As a concession for MassiveSumm, we decided to concede compression from 80% to 90% and we find that it added a fairly high number of articles to the valid set for MassiveSumm (giving us a total of 17248 pairs, which we then divide into about 80%-10%-10% to get the train, dev and test splits). Relaxing the compression further would increase the numbers, but we also note that the authors themselves have presented their results on a randomly selected 12633 pairs (not made available by the author), therefore we take a comparative number, which according to us should be of a better quality due to the aggressive quality control.

For XL-Sum, the only option was to remove all constraints, as the original size itself was quite small. Therefore, we considered the original splits of XL-Sum (Telugu) for our experiments.

Parameters	**PG+**	**MLE+**	**BertSumAbs**	**mT5**
Max source length	400	400	512	512
Max target length	100	100	200	256
Min target length	35	35	50	30
Batch Size	8	8	140	2
Epochs/Iterations	100k iter	100k iter	50k iter	10 epochs
Vocab Size	50k	50k	28996	250112
Beam Size	4	4	5	4
Learning Rate	0.15	0.001 (MLE) 0.0001 Others	lr_bert = 0.002 lr_dec = 0.2	5e-4

Table 6: Experimental setup and parameter settings

8. Results and Analysis

For better comparison, experiments were conducted by training on each dataset's training split and then testing on all 3 datasets' test set. Table 7 shows the ROUGE scores[4] for some of the selected best performing model configurations. Here, 'wo' with MLE+RL and RL models stands for 'without intra attention', and 'Pointer Generator' represents the PG+Coverage model.

Looking at this table our first observation is that models trained on TeSum end up performing well across the board, but do not end up beating models trained and tested on the same dataset for almost all models. We surmise that this is because the fundamental nature of these summarization datasets is different. While MassiveSumm and XL-Sum summaries are primarily small number of disjoint sentences, TeSum summaries are coherent discourses in themselves. This means that a model trained to avoid copying and trained to generate coherent discourse would fail on MassiveSumm and XL-Sum.

While we accept the contributions made by XL-Sum and MassiveSumm, which bring value to this field for any given language, we claim that this scraping and the initial pre-processing is just the first step. The data need to be held to higher standards. Even if it is achieved

[4]Multilingual Rouge from XL-Sum `https://tinyurl.com/MLTERouge`

Trained on TeSum									
	Tested on TeSum			Tested on MassiveSumm			Tested on XL-Sum		
Model	**T-R1**	**T-R2**	**T-RL**	**M-R1**	**M-R2**	**M-RL**	**XL-R1**	**XL-R2**	**XL-RL**
Pointer Generator	**39.37**	**22.72**	**32.15**	25	**13.8**	20.74	**9.73**	**2.29**	**7.32**
MLE + RL- wo	38.09	21.9	31.77	**25.05**	13.41	**20.78**	8.56	2.03	6.54
RL- wo	31.19	17.6	24.86	20.16	10.58	16.75	8.28	1.96	6.45
BertSumAbs	26.49	12.55	19.6	18.61	8.24	14.69	6.21	1.34	4.96
mT5-small	37.42	20.82	30.88	24.37	12.5	20.2	8.8	2.06	6.7

Trained on MassiveSumm									
	Tested on TeSum			Tested on MassiveSumm			Tested on XL-Sum		
Model	**T-R1**	**T-R2**	**T-RL**	**M-R1**	**M-R2**	**M-RL**	**XL-R1**	**XL-R2**	**XL-RL**
Pointer Generator	26.31	14.45	22.16	28.38	16.33	24.95	**9.85**	**2.18**	**8.34**
MLE + RL- wo	26.3	14.62	22.36	**30.46**	**18.69**	**27.17**	9.82	2.03	8.23
RL- wo	15.59	7.93	13.75	13.82	7.76	12.82	4.27	0.89	3.88
BertSumAbs	**29.73**	13.59	22.02	23.76	11.47	19.28	7.69	1.54	6.11
mT5-small	27.67	**15.08**	**23.03**	29.43	17.41	26.25	9.67	1.91	8.14

Trained on XL-Sum									
	Tested on TeSum			Tested on MassiveSumm			Tested on XL-Sum		
Model	**T-R1**	**T-R2**	**T-RL**	**M-R1**	**M-R2**	**M-RL**	**XL-R1**	**XL-R2**	**XL-RL**
Pointer Generator	17.13	2.2	10.06	12.45	1.59	8.73	5.41	0.28	4.35
MLE + RL- wo	3.7	0.44	3.18	2.88	0.43	2.63	1.17	0.04	1.13
RL- wo	2.4	0.28	2.14	1.61	0.17	1.49	0.68	0.04	0.66
BertSumAbs	13.8	2.41	10.26	11.46	2.06	9.19	6.55	1.44	5.73
mT5-small	**18.42**	**9**	**15.88**	**19.44**	**9.12**	**17.46**	**12.24**	**3.6**	**11.18**

Table 7: ROUGE scores achieved by various baseline models.

by scraping, filtering and then evaluating a percentage of randomly selected samples of the resultant, it would ensure a much more valuable dataset than just scraping.

9. Conclusion

Dataset creation for any task is an expensive and complex activity. With the increased demand for data for deep-learning models, it is often infeasible to create datasets which reach the desired sample counts. It then does make sense to make do with data collected "from the wild". It is our opinion that such collected data, while useful, should also be subjected to quality control. At the same time, we should adopt pipelines which can establish a balance between quality control and cost. This is especially critical for Low Resource Languages which need to make do with low sample numbers.

To this effect, we constructed a high quality Human-curated Abstractive summarization dataset for Telugu. We also compared the dataset properties with existing Telugu summarization datasets and claim that these existing datasets can also benefit from the quality control measures that we have proposed.

Though, purely on the basis of size, our work also started with a huge collection of 92k+ article-summary pairs like the existing datasets, but by making use of human expertise at both annotation and quality assessment stages, we show that after applying the same quality measures our dataset performs significantly better then the automated ones. And as a result we outperform the other datasets in terms of final size as well.

10. Acknowledgements

We thank Lokesh Madasu and Gopichand Kanumolu for their assistance with the manual evaluations. We are thankful to K. Ravikanth (Faculty, RGUKT-Basar), J. Chakravarthi (Faculty, RGUKT-Nuzvid) and Satish Kumar Ch (Faculty, RGUKT-Srikakulam) for their constant and unyielding support and effort.

11. Bibliographical References

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bommasani, R. and Cardie, C. (2020). Intrinsic evaluation of summarization datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8075–8096.

Grusky, M., Naaman, M., and Artzi, Y. (2018). Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. *arXiv preprint arXiv:1804.11283*.

Hasan, T., Bhattacharjee, A., Islam, M. S., Mubasshir, K., Li, Y.-F., Kang, Y.-B., Rahman, M. S., and Shahriyar, R. (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, August. Association for Computational Linguistics.

Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., and Blunsom, P. (2015). Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28:1693–1701.

Hovy, E. and Lin, C.-Y. (1998). Automated text summarization and the summarist system. In *TIPSTER TEXT PROGRAM PHASE III: Proceedings of a Workshop held at Baltimore, Maryland, October 13-15, 1998*, pages 197–214.

Jones, K. S. (1998). Automatic summarising: factors and directions. *ArXiv*, cmp-lg/9805011.

Koo, T. K. and Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2):155–163.

Liu, Y. and Lapata, M. (2019). Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345*.

Marreddy, M., Oota, S. R., Vakada, L. S., Chinni, V. C., and Mamidi, R. (2021). Clickbait detection in telugu: Overcoming nlp challenges in resource-poor languages using benchmarked techniques. In *2021 International Joint Conference on Neural Networks (IJCNN)*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Nallapati, R., Zhou, B., Gulcehre, C., Xiang, B., et al. (2016). Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*.

Narayan, S., Cohen, S. B., and Lapata, M. (2018). Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *arXiv preprint arXiv:1808.08745*.

Over, P. and Yen, J. (2004). An introduction to duc-2004. *National Institute of Standards and Technology*.

Paulus, R., Xiong, C., and Socher, R. (2017). A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Rush, A. M., Chopra, S., and Weston, J. (2015). A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*.

See, A., Liu, P. J., and Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada, July. Association for Computational Linguistics.

Shrout, P. E. and Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Varab, D. and Schluter, N. (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.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2020). mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Appendix

A. Examples of Sub-standard Summaries

We are giving some examples of un-acceptable article-summary pairs from pre-existing datasets. Instead of Telugu script, we have used phonetic transcription of Telugu using ISO15919, which is similar to IAST, for better readability.

Corpus	XL-Sum
IDs	international-54722433 international-55923039
URLs	https://www.bbc.com/telugu/international-54722433 https://www.bbc.com/telugu/international-55923039
Text	Article-1: lākḍaun valla cālāmaṁdi iḷḷakē parimitaṁ ayyāru. ī yuvati khāḷīgā kūrcōkuṁḍā mēkap braṣ paṭṭukuni pramukhullā tayārai prācuryaṁ poṁdāru. (bībīsī telugunu phēsbuk , insṭāgrām , ṭviṭarlō phālō avvaṁḍi. yūṭyūblō sabskraib cēyaṁḍi.) Article-2: 2013 lō vēls rājadhāni kārḍiph nuṁci siriyā veḷḷi aisislō cērāru koṁdaru ṭīnējarlu. vāḷḷu aisislō cēraḍāniki kāraṇālēṁṭō telusukōvālani bībīsī pratinidhi olīviyā vārini iṁṭarvyū cēśāru. iṁtakī dēśālu dāṭi veḷḷi aisislō cērina vāḷḷaṁtā akkaḍa telusukunna vāstavālēṁṭi ? (bībīsī telugunu phēsbuk , insṭāgrām, ṭviṭarlō phālō avvaṁḍi. yūṭyūblō sabskraib cēyaṁḍi.)
Summary	ivi kūḍā cadavaṁḍi:
Remark	17 different articles have the same summary

Table 8: XL-Sum: Duplicate Summary example

Corpus	MassiveSumm
URLs	https://telugu.asianetnews.com/astrology/today-may-1st-2019-your-horoscope-pqt03m https://telugu.asianetnews.com/astrology/today-2nd-july-2019-tuesday-your-horoscope-ptzqhq
Text	Article-1: mēṣaṁ : (aśvini , bharaṇi , kṛttika 1 vapādaṁ) peddalaṁṭē gauravaṁ perugutuṁdi . ādhyātmika ciṁtana perugutuṁdi śāstra parijñānaṁ pai dṛṣṭi ērpaḍutuṁdi . viśāla bhāvālu uṁṭāyi . vidya nērcukōvaḍaṁ valla vaccē gauravaṁ perugutuṁdi . . . .rājakīyālapai dṛṣṭi sāristāru. gauravaṁ peṁcukunē prayatnaṁ. vṛtti udyōgādullō ottiḍulu uṁṭāyi . śrī mātrē namaḥ japaṁ maṁcidi. Article-2: mēṣaṁ : (aśvini , bharaṇi , kṛttika 1 vapādaṁ) racanalapai dṛṣṭi taggutuṁdi . kamyūnikēṣans valla anukūlata perugutuṁdi . parāmarśalu cēstāru . pracārālapai dṛṣṭi ērpaḍutuṁdi . baṁdhuvula sahakāraṁ labhistuṁdi . prayāṇāla valla jāgratta avasaraṁ. . . . . .vidyārthulaku kaṭhinamena samayaṁ . ālōcanallō ottiḍi ērpaḍutuṁdi . durgādēvi pūja cēsukōvaḍaṁ śubha phalitālanistuṁdi
Summary	ī rōju rāśiphalālu ilā unnāyi
Remark	Many articles (with different URLs) have the same summary

Table 9: MassiveSumm: Duplicate summary example

Corpus	MassiveSumm
URL	https://telugu.asianetnews.com/entertainment-news/sridevi-s-second-death-anniversary-prayer-meet-in-chennai-q6oh3x
Text	2018 phibravari 24 na dubāy lō śrīdēvi anumānāspada sthitilō maraṇiṁcāru. atilōka suṁdarigā śrīdēvi iṁḍiyā mottaṁ tirugulēni krēj soṁtaṁ cēsukuṁdi. śrīdēvi akāla maraṇaṁ ceṁdaḍaṁtō citra pariśrama tōpāṭu abhimānulu kūḍā tīvra viṣādāniki gurayyāru gata phibravari 2 4 na ku śrīdēvi maraṇiṁci reṁḍēḷḷu pūrtayiṁdi. . . . . .amma nuvvu ikkaḍē uṁḍālani kōrukuṁṭunnā ani jānvī kāmeṁṭ peṭṭiṁdi . jānvī sṭār hīrōyin gā rāṇiṁcālanēdi śrīdēvi kala . prastutaṁ jānvī bālīvuḍ lō palu citrāllō naṭistōṁdi .
Summary	2018 phibravari 24 na dubāy lō śrīdēvi anumānāspada sthitilō maraṇiṁcāru. atilōka suṁdarigā śrīdēvi iṁḍiyā mottaṁ tirugulēni krēj soṁtaṁ cēsukuṁdi
Remark	The highlighted content is the prefix information

Table 10: MassiveSumm: Prefix example

Corpus	XL-Sum
Language	Telugu
ID	international-41926617
URL	https://www.bbc.com/telugu/international-41926617
Text	prāṇālanu guppiṭlō peṭṭukoni lakṣala maṁdi prajalu śaraṇārthulugā nagarānni vadaliveḷlāru. vēla maṁdi maraṇiṁcāru. eṁtō maṁdi kuṭuṁba sabhyulanu kōlpōyi tīvra vēdanaku guravutunnāru. alā sarvaṁ kōlpōyina ō bādhituḍi vyatha idi. ī vīḍiyōnu bībīsī arabik rūpoṁdiṁciṁdi. mā itara kathanālu : (bībīsī telugunu phēs buk, in sṭāgrām, ṭviṭar lō phālō avvaṁḍi. yūṭyūb lō sab skraib cēyaṁḍi.)
Summary	aies miliṭeṁṭlaku, sainika balagālaku madhya jarigina pōrulō siriyālōni rakhā nagaraṁ dhvaṁsamaiṁdi.iḷḷannī dhvaṁsamayyāyi.

Table 11: XL-Sum(Telugu): Out of the context example

Corpus	XL-Sum
Language	Marathi
ID	media-54453803
URL	https://www.bbc.com/marathi/media-54453803
Text	up-rāṣṭrādhyakṣālā rāṣṭrādhyakṣācā raniṁg mēṭ mhaṇajēc sāthīdār mhaṭalaṁ jātaṁ . up - rāṣṭrādhyakṣācyā umēdavārācyā pratiṣṭhēvar āṇi kāryakṣamatēkaḍē pāhūnahī matadān karaṇārā varg amērikēt āhē . aśāvēḷī ḍēmōkrēṭik umēdavār kamalā hêris yāṁcyāviṣayī tumhālā adhik jāṇūn ghyāyacaṁ āhē . pāhā hā vhīḍiō . . hēhī pāhilaṁt kā ? (bībīsī marāṭhīcē sarv apaḍēṭs miḷavaṇyāsāṭhī tumhī āmhālā phēsabuk ,insṭāgrām , yūṭyūb , ṭviṭar var phôlō karū śakatā . ’ bībīsī viśv ’ rōj saṁdhyākāḷī 7 vājatā JioTV êp āṇi yūṭyūbavar nakkī pāhā .)
Summary	amērikētē prēsiḍēṁśiēl āṇi vhāīs prēsiḍēṁśiēl ḍibēṭalā mahattv asataṁ. yā dōnhī ḍibēṭamuḷē sarakāracī diśā āṇi dhyēyadhōraṇaṁ lōkāṁnā samajatāt .

Table 12: XL-Sum(Marathi): Out of the context example