LREC 2022, 13th Conference on Language Resources and Evaluation

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

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

What is TeSum?

Human Generated, Curated, Abstractive Summarization Dataset

- **Total article-summary pairs : 20329**
    - ○ Training set : 16295 pairs
    - ○ Development set : 2017 pairs
    - ○ Test set : 2017 pairs
- **# Summary Creators : 347**
- **# Raters : 03**

Existing Datasets for Telugu

- XL-Sum (Telugu)
  - **13,025** Article – Summary pairs
  - Collected automatically from multiple news sources

Average Article length:	**50.99**
Average Article word count:	**609.44**
Average Summary length:	**2.01**
Average Summary word count:	**24.58**
Average Compression:	~ 96%

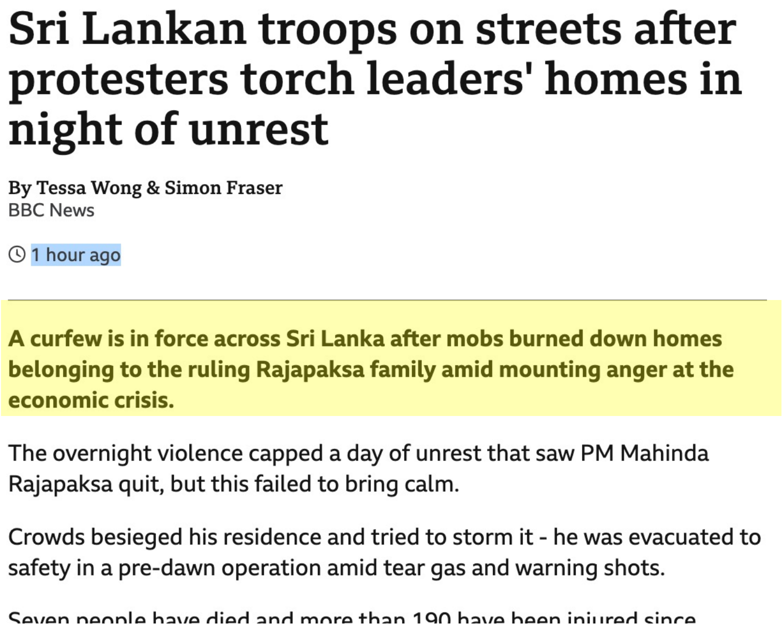
- MassiveSumm (Telugu)
  - **1,19,282** Article-Summary pairs
  - Collected automatically from multiple news sources

Average Article length:	**19.41**
Average Article word count:	**235.15**
Average Summary length:	**2.67**
Average Summary word count:	**29.86**
Average Compression:	~86%

- Collect “Highlights” as summary
- Advantages: Cost, Size …
- Disadvantages: Quality

HIGHLIGHT != SUMMARY

- Highlight
  - May be bullet points
  - Text in **BOLD** fonts at the beginning of the article
- True for many News outlets but not across Languages
- Common Problems:
  - Highlight is the often the same as initial Sentence(s)
  - Highlight is the start of the article
  - Article contains only an image/video; highlight is just the caption



Sri Lankan troops on streets after protesters torch leaders' homes in night of unrest

By Tessa Wong & Simon Fraser
BBC News

1 hour ago

A curfew is in force across Sri Lanka after mobs burned down homes belonging to the ruling Rajapaksa family amid mounting anger at the economic crisis.

The overnight violence capped a day of unrest that saw PM Mahinda Rajapaksa quit, but this failed to bring calm.

Crowds besieged his residence and tried to storm it - he was evacuated to safety in a pre-dawn operation amid tear gas and warning shots.

Seven people have died and more than 190 have been injured since

Questions on Highlights

- The motive behind “highlights”
  - Draw attention
  - Pique interest
  - Lead the user to the article
- Can be very short
  - Almost a headline
  - Lacks coverage
  - Often not even indicative
    - Irrelevant
- Editorial practice
  - Not enforced
  - Might not even intend to be a summary
  - Might be part of the article discourse
    - Cannot be separated

- Desiderata for an “abstractive” summary
  - Relevance
  - Coverage
  - Novelty/creativity
- Highlights often do not meet these criteria

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

Examples of Bad Samples XL-Sum (Telugu)

Duplicate Summaries:

ID: "international-54722433"
Text: "లాక్‌డౌన్ వల్ల చాలామంది ఇళ్లకే ... ... చేయండి.)"

Translation: many have been confined to their homes. the girl is not sitting idle. the makeup brush became popular. (facebook, instagram, twitter)

ID: "india-55219925"
Text: "ఐవరీకోస్ట్‌కు చెందిన కోఫీ ఎండ్రీ ... ... సబ్‌స్క్రైబ్ చేయండి.)"

Translation: the motorcycle was designed by a man named kofi endri palin of the iwakot. he said that the construction work of the building will be completed in three hours with the help of mini crane. (facebook, instagram, twitter)

Common Summary: "ఇవి కూడా చదవండి:"

Summary Translation: "also read:"

Total Occurrences: 17

Out of Context Summary:

ID: "international-41926617"
Text: "ప్రాణాలను గుప్పిట్లో పెట్టుకొని లక్షల మంది ... ... సబ్‌స్క్రైబ్ చేయండి.)"

Translation: lakhs of people have left the city. thousands have died. many have lost their families. this is the tragedy of the victim. the video was shot by bbc. other stories:(facebook, instagram and twitter)
Summary: "ఐఎస్ మిలిటెంట్లకు, సైనిక ... ... ధ్వంసమయ్యాయి."

Translated Summary: "the city of raqqa in syria was destroyed in the fighting between the isis and the military. all the houses were damaged."

ID: "india-55219925"
Text: "అయితే, ఆ జాబితా ... ... యూట్యూబ్‌లో సబ్‌స్క్రైబ్ చేయండి.)"
Translation: "However, the pilots say the list is inaccurate. The government has stated that the licenses are not fake and that there are loopholes in the testing process. While the Imran Khan government is taking corrective action, the Supreme Court has ordered action against those involved in the license scam. Article by BBC correspondent Shumaila Jaffrey. Also read: (Follow BBC Telugu on Facebook, Instagram, Twitter, Subscribe on YouTube.)"
Summary: "పాకిస్తాన్ విమానయాన ... ... విడుదల చేశారు."

Translated Summary: "A new crisis has begun in the Pakistani aviation sector. The country's aviation minister himself has released a list of 262 pilots in the country who have fake licenses."

Examples of Bad Samples MassiveSumm (Telugu)

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

Duplicate, irrelevant Summaries:

"**url":** "https://telugu.asianetnews.com/astrology/today-march13th-your-horoscope-poa61t"
Text: "Hyderabad, First Published Mar 13, 2019, 6:43 AM IST\n\nమేషం :(అశ్విని, భరణి, కృత్తిక 1వపాదం) : మానసిక ఒత్తిడి … … … … పరించడం మంచిది.\n\nLast Updated Mar 13, 2019, 6:43 AM IST"
Translated Text: "hyderabad, first published major 13, 2019, 6:43 am sham:(, ashwini, creative 1.) official thinking increases pressure. children have problems. administrative capacity will increase. … … … … you will get victory over enemies. credit is good. getting recognition in professional education. increased immunity. better read"

"url": "https://telugu.asianetnews.com/astrology/today-may-7th-2019-your-horoscope-pr41b9"
Text: "మేషం :(అశ్విని, భరణి, కృత్తిక 1వపాదం) : వాక్ చాతుర్యం తగ్గుతుంది … … … … లలితా సహస్రనామ పారాయణ ముఖ్యం.\n\nLast Updated May 7, 2019, 6:55 AM IST"
Translated Text: "aries:(ashwini, bhanu, krithika, and curd):wash your hair. value of words decreases. promises can cause problems. … … … … students are under pressure. you need to be careful while travelling. lalitha sahasranama parayanam is important. there is a pattern. interest in writing will increase. you will get cooperation with colleagues. conversations will be fruitful. communication brings satisfaction. publicity and resources will benefit."

"url": "https://telugu.asianetnews.com/astrology/todau-june5th-your-horoscope-pslpwd"
Text: "మేషం :(అశ్విని, భరణి, కృత్తిక 1వపాదం) : రచనలపై దృష్టి తగ్గుతుంది. కమ్యూనికేషన్స్ వల్ల అనుకూలత పెరుగుతుంది. పరామర్శలు చేస్తారు. … … … … పూజ చేసుకోవడం శుభ ఫలితాలనిస్తుంది.\n\nLast Updated Jun 5, 2019, 6:39 AM IST"
Translated Text: "aries:(ashwini, bhanu and krithika 1):the focus on writing will be reduced. connectivity increases. are reviewed. focus on publicity. friends will help. you need to be careful while travelling. friends will help. durga puja brings good results. … … … … maternal mortality is low. time is tough for students. there is pressure on ideas. worshipping durga brings good results."

Common Summary: "ఈ రోజు రాశిఫలాలు ఇలా ఉన్నాయి"

Translated Summary: "today's horoscope is :"

Total Occurrences: **527**

Examples of Bad Samples MassiveSumm (Telugu)

Prefixes or Lead-N Summaries:

"url": "https://telugu.asianetnews.com/entertainment-news/sridevi-s-second-death-anniversary-prayer-meet-in-chennai-q6oh3x"
Text: “Hyderabad, First Published Mar 4, 2020, 10:13 PM IST\n\n2018 ఫిబ్రవరి 24న దుబాయ్ లో శ్రీదేవి అనుమానాస్పద … … … లో పలు చిత్రాల్లో నటిస్తోంది.\n\nLast Updated Mar 4, 2020, 10:13 PM IST”

Translated Text: “**Sridevi died under suspicious circumstances on February 24, 2018 in Dubai**. As a heavenly beauty, Sridevi has created an unstoppable craze all over India. Sridevi's untimely death has caused a great deal of grief to the film industry as well as fans. On February 24, two years have passed since Sridevi's death. Meanwhile, according to Hindu traditions, the Bonnie Kapoor family held Sridevi's second funeral in Chennai. Bonikapur, Janvi Kapoor and other family members and friends participated in Sridevi's second Varthanthi celebrations. Janvi Kapoor shared the scenes on social media. Janvi Kapoor shines in a traditional look skirt. Shared photos with friends on Instagram. Janvi Kapoor made an emotional comment about her mother Sridevi on this occasion. Janvi commented that Amma wants you to stay here. Sridevi's dream is to become a Janvi star heroine. Janvi is currently starring in several films in Bollywood.”

Summary: “2018 ఫిబ్రవరి 24న దుబాయ్&nbsp;లో శ్రీదేవి అనుమానాస్పద స్థితిలో మరణించారు. అతిలోక సుందరిగా&nbsp;శ్రీదేవి ఇండియా మొత్తం తిరుగులేని క్రేజ్ సొంతం చేసుకుంది.”
Translated Summary: “**Sridevi died under suspicious circumstances on February 24, 2018 in Dubai**. Sridevi is the most beautiful woman in the world.”

"url": "https://telugu.asianetnews.com/entertainment/24-kisses-movie-trailer-ph5295"
Text: “మేషం :(అశ్విని, భరణి, కృత్తిక 1వపాదం) : రచనలపై దృష్టి తగ్గుతుంది. కమ్యూనికేషన్స్ వల్ల అనుకూలత పెరుగుతుంది. పరామర్శలు చేస్తారు. … … … … పూజ చేసుకోవడం శుభ ఫలితాలనిస్తుంది.\n\nLast Updated Jun 5, 2019, 6:39 AM IST”

Translated Text: “**'24 Kisses' is a movie starring Arun Adit and Hebba Patel as a couple. Directed by Ayodhya Kumar Krishnam Shetty, the movie's posters and teaser were so romantic that the youth focused on the movie. The film team has recently released the trailer of the movie**. … … … … concept is new. Last Updated Oct 25, 2018, 10:37 AM IST”

Summary: “అరుణ్ ఆదిత్, హెబ్బా పటేల్ జంటగా నటిస్తోన్న చిత్రం '24 కిస్సెస్'. అయోధ్యకుమార్ కృష్ణం శెట్టి డైరెక్ట్ చేస్తోన్న ఈ సినిమా పోస్టర్లు, టీజర్ చాలా రొమాంటిక్ గా ఉండడంతో సినిమా పై యూత్ దృష్టి పడింది. తాజాగా ఈ సినిమా&nbsp;ట్రైలర్ ని విడుదల చేసింది చిత్రబృందం.”
Translated Summary: “**'24 Kisses' is a movie starring Arun Adit and Hebba Patel as a couple. Directed by Ayodhya Kumar Krishnam Shetty, the movie's posters and teaser were so romantic that the youth focused on the movie. The film team has recently released the trailer of the movie.**”

Unsupervised Collection of Summarization Data

Based on the assumptions of :

- Uniform Editorial Practices
- Availability of Quality Summaries

  (Mostly Flawed Assumptions)

Data Quality :

- Questionable
- Valid pairs compression: 71%

	XL-Sum	MassiveSumm
Total Size	**13025**	**119282**
Empty + Duplicate + Duplicate Summary	143	**18733**
Prefixes	3	30741
Minimum Length	384	10399
Compression < 50%	10	1641
Compression > 80%	11920	46776
Abstractivity < 10%	0	6683
Abstractivity > 80%	227	303
Remaining Valid	**338**	**4006**

How??

Crowdsourcing

- 347 crowdsource workers
- 1 HIT = 50 articles (long duration)
- Collected 92,941 article-summary pairs

- Guidelines for Summary preparation
  - Relevance
  - Readability & Coherence
  - Creativity

- Why? ⇒ Reduces Cost

- Filter for:
  - Compression
  - Abstractivity
  - Length

- Why not specific compression etc.?
  - Do not want to restrict ; leads to unnatural sentence construction

How??

Curation

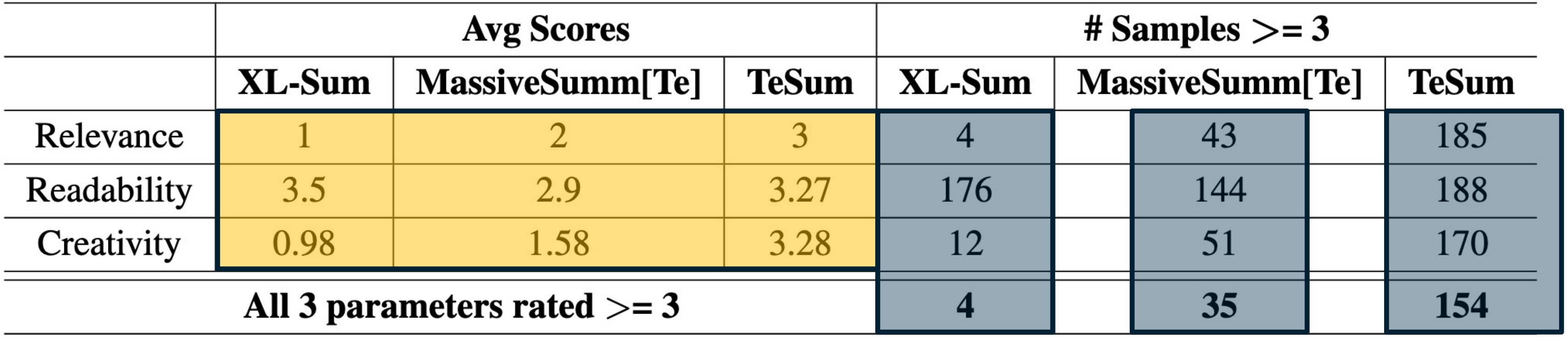
Empty + Duplicate + Duplicate Summary	652
Prefixes	1330
Minimum Length	4205
Not Compressed Enough	52802
Too Compressed	456
Not Abstractive Enough	5942
Too Abstractive	42

- Automatic Filtering
  - Removes noise
- Automatic Quality Control
  - Helps meet data requirements
    - Abstractivity
    - Compression
    - Length
  - Samples after filters:
    - 27,152 out of 92k
- Sampled-Manual Evaluation
  - Reduces evaluation cost

Sampled Manual Evaluation

- Evaluated ~25% of “filtered” samples from each HIT
  - Sampled randomly
  - Approximately 6900 samples

- Three expert raters
  - The number of evaluated samples in the final dataset:
    - 5089 article summary pairs



	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.

Baselines

- **Training on TeSum:**
    - Human generated summary as a reference
    - Feature balanced test set
        - Length
        - Compression
- **Training on Existing Datasets:**
    - Models trained on other two datasets do not perform well
- **Testing on Existing Datasets:**
    - Data characteristics are too divergent
    - Still, achieve good results

Baselines on TeSum

Model	R1	R2	RL
Pointer Generator	**39.37**	**22.72**	**32.15**
MLE+RL-wo	38.09	21.9	31.77
BertSumAbs	26.49	12.55	19.60
mT5	37.42	20.82	30.88

Thank You!

INTERNATIONAL INSTITUTE OF
INFORMATION TECHNOLOGY
HYDERABAD