TeSum: Human Generated Abstractive Summarization Corpus for Telugu

Existing Corpora

- **Web-Scraped**; assumption that good summaries are provided by the publishers.
- **Collection of Highlights**; often ungrammatical and incoherent
- **Article contains only an image/video;** Summary is just the caption
- **Highlights from the beginning**; initial sentences (prefixes) or lead-3 summaries

Duplicate, Irrelevant Summaries:

Text: మేషం :(అశ్విని, భరణి, కృత్తిక 1వపాదం) : మానసిక ఒత్తిడి …….పఠించడం మంచిది. [**#tokens = 578**]
Translated Text: Aries :(Ashwini, Bharani, Krittika 1vapadam) : Mental stress …….Good to recite.

Summary: "ఈ రోజు రాశిఫలాలు ఇలా ఉన్నాయి"
[**#tokens = 5**]
Translated Summary: "today's horoscope is:"

Total Occurrences: 527

Prefixes or Lead-N Summaries:

Text: 2018 ఫిబ్రవరి 24న దుబాయ్ లో శ్రీదేవి అనుమానాస్పద స్థితిలో మరణించారు. అతిలోక సుందరిగా శ్రీదేవి ఇండియా మొత్తం తిరుగులేని క్రేజ్ సొంతం చేసుకుంది….పలు చిత్రాల్లో నటిస్తోంది.
Translated Text: On February 24, 2018, Sridevi died under suspicious circumstances in Dubai. Sridevi is an unstoppable craze for the whole of India as a celestial beauty....She is acting in many films.

Summary: 2018 ఫిబ్రవరి 24న దుబాయ్ లో శ్రీదేవి అనుమానాస్పద స్థితిలో మరణించారు. అతిలోక సుందరిగా శ్రీదేవి ఇండియా మొత్తం తిరుగులేని క్రేజ్ సొంతం చేసుకుంది.
Translated Summary: On February 24, 2018, Sridevi died under suspicious circumstances in Dubai. Sridevi is an unstoppable craze for the whole of India as a heavenly beauty.

With TeSum, we …

- **Present a human generated, curated abstractive summarization corpus for Telugu**, with a total of 20329 article-summary pairs
  - Training set: 16295 pairs
  - Dev set: 2017 pairs
  - Test set: 2017 pairs
- **Propose a pipeline that crowd-sources summarization data** and then aggressively filtered the content via: automatic and partial expert evaluation.
  - # Summary Creators: 347
  - # Raters : 03
- **Demonstrate Improvement** over XL-SUM and MassiveSumm (Telugu versions)
- **Provide baselines** with various SOTA models
- **Propose automatic filtrations** for existing scraped datasets

	XL-Sum	MassiveSumm	TeSum
Total Size	**13025**	**119282**	**92941**
Empty + Duplicate + Duplicate Summary	143	18733	652
Prefixes	3	30741	1330
Minimum Length	384	10399	4205
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**

Human Evaluation

- Randomly selected 200 samples from each dataset
- Each sample is scored between 0-4, for each evaluation parameter.
  - **Relevance:** All or most of the relevant information contained in the article should be present in the summary.
  - **Readability and Coherence:** The summary should be coherent, readable and free of any grammatical errors.
  - **Creativity:** The summary should have novel sentential and phrasal structures.

	Average Scores		
	XL-SUM	**MassiveSumm**	**TeSum**
Relevance	1	2	**3**
Readability	3.5	2.9	**3.27**
Creativity	0.98	1.58	**3.28**

	# Samples with all parameters rated >=3		
	XL-SUM	**MassiveSumm**	**TeSum**
Relevance	4	43	**185**
Readability	176	144	**188**
Creativity	12	51	**170**

Author names: Ashok Urlana, Nirmal Surange*, Pavan Baswani, Priyanka Ravva, Manish Shrivastava email ID: ashok.urlana@research.iiit.ac.in*