XL-Sum: Large-Scale Multilingual Abstractive Summarization for 44 Languages

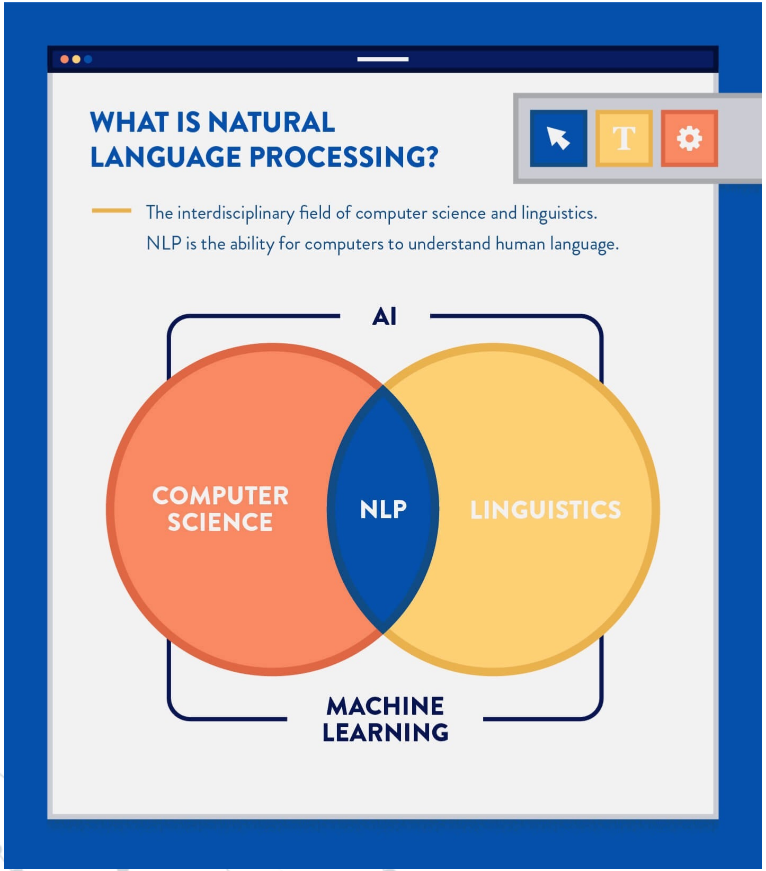
Presented by:
Kushan Hewapathirana – 229333P

Published in: ACL-IJCNLP 2021: Main Volume: August 2021

APPLICATION DOMAIN:

Natural Language Processing – Multilingual Summarization

Introduction to Natural Language Processing

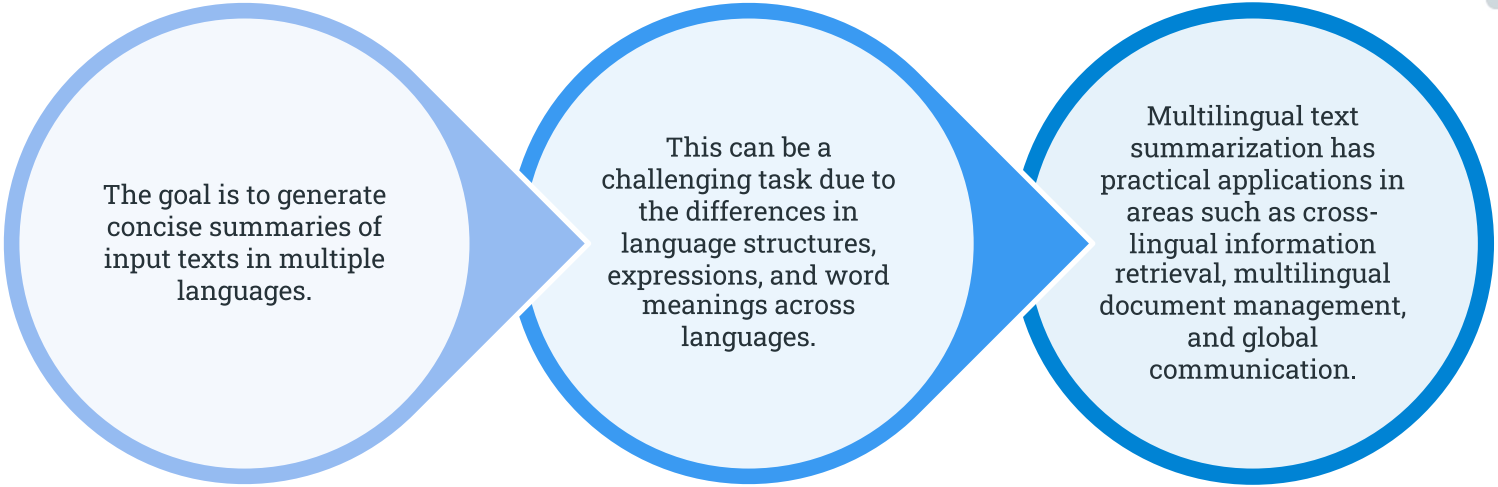


WHAT IS NATURAL LANGUAGE PROCESSING?

The interdisciplinary field of computer science and linguistics. NLP is the ability for computers to understand human language.

- Speech recognition
- Part of speech tagging
- Word sense disambiguation
- Named entity recognition
- Co-reference resolution
- Sentiment analysis
- Natural language generation

Introduction to Multiligual Summarization



The goal is to generate concise summaries of input texts in multiple languages.

This can be a challenging task due to the differences in language structures, expressions, and word meanings across languages.

Multilingual text summarization has practical applications in areas such as cross-lingual information retrieval, multilingual document management, and global communication.

PROBLEM STATEMENT:

Limitations in analyzing cross-document interactions and capturing global knowledge element interactions

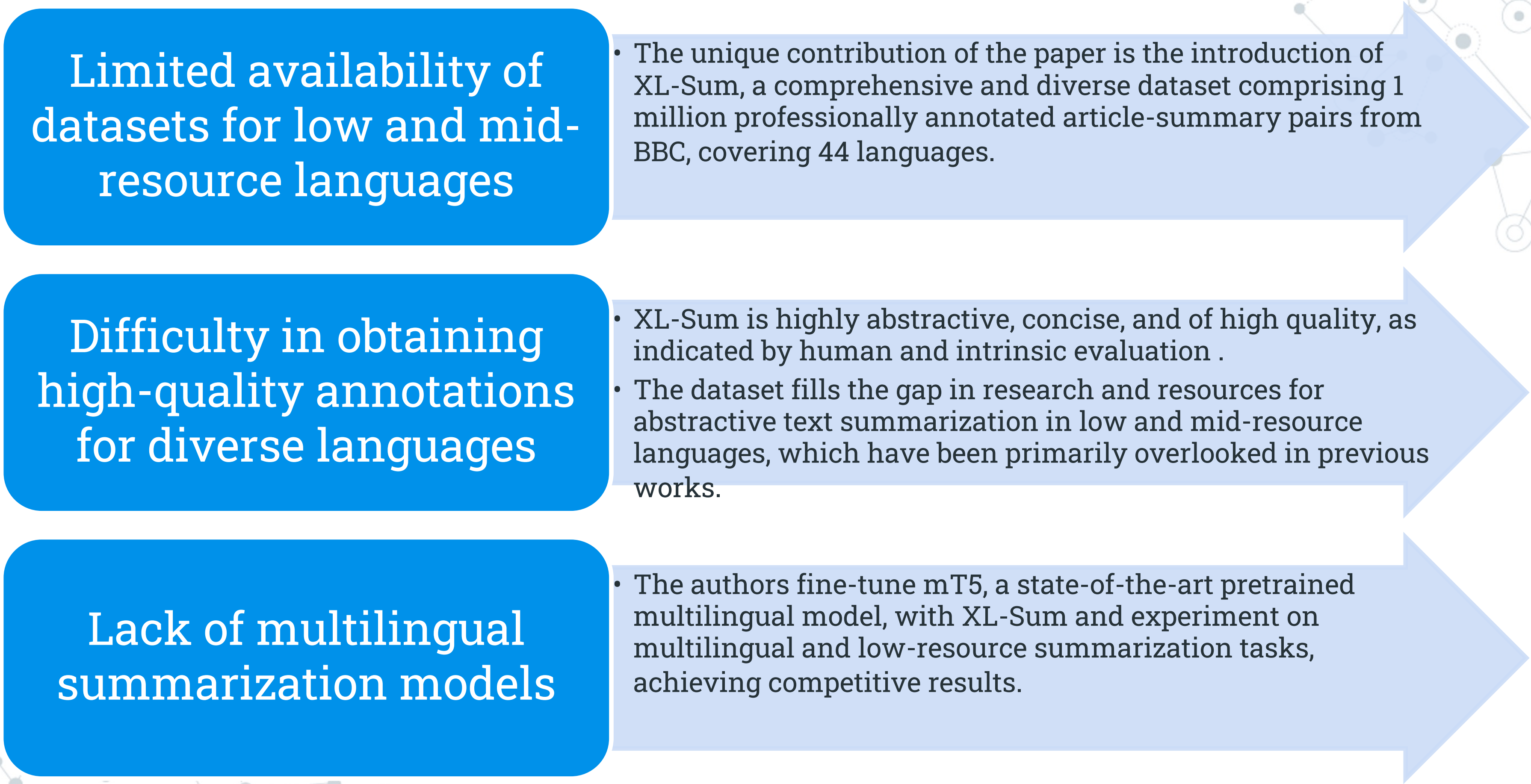
Challenges in Multiligual Summarization

Limited availability of datasets for low and mid-resource languages hinders progress in multilingual summarization.

Difficulty in obtaining high-quality annotations for diverse languages poses a challenge in developing effective summarization models.

Lack of benchmark datasets and models for low and mid-resource languages limits the evaluation and comparison of multilingual summarization systems.

Bridging the Research Gap: Unique Contribution



Limited availability of datasets for low and mid-resource languages

- The unique contribution of the paper is the introduction of XL-Sum, a comprehensive and diverse dataset comprising 1 million professionally annotated article-summary pairs from BBC, covering 44 languages.

Difficulty in obtaining high-quality annotations for diverse languages

- XL-Sum is highly abstractive, concise, and of high quality, as indicated by human and intrinsic evaluation .
- The dataset fills the gap in research and resources for abstractive text summarization in low and mid-resource languages, which have been primarily overlooked in previous works.

Lack of multilingual summarization models

- The authors fine-tune mT5, a state-of-the-art pretrained multilingual model, with XL-Sum and experiment on multilingual and low-resource summarization tasks, achieving competitive results.

Methodology

Dataset Development

•The dataset was created using a data curation tool that automatically crawled and extracted article-summary pairs from BBC, ensuring the dataset can be expanded over time.

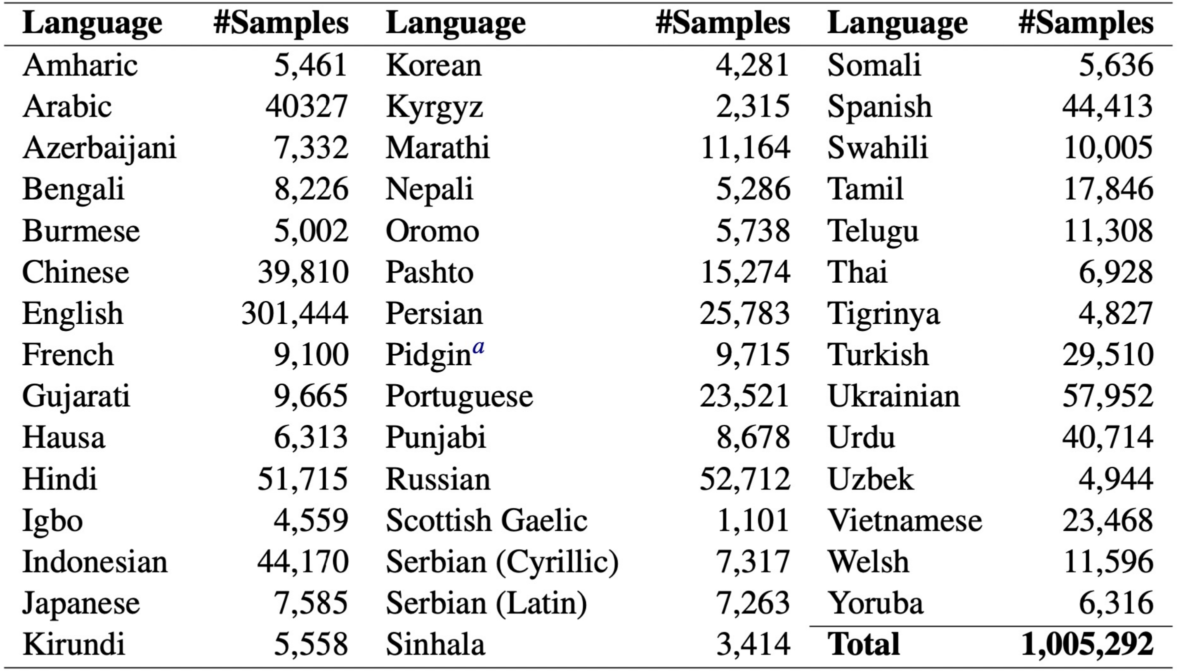
•The dataset was extracted using carefully designed heuristics, taking advantage of the consistent editorial style of the BBC articles .

•The annotators labelled each article-summary pair based on three properties: Property A, which assesses if the summary conveys the main idea; Property B, which checks for consistency between the article and the summary; and Property C, which identifies additional information present in the summary .

•Inter-annotator agreement was measured using Cohen's kappa coefficient, and most scores showed high agreement between the evaluator

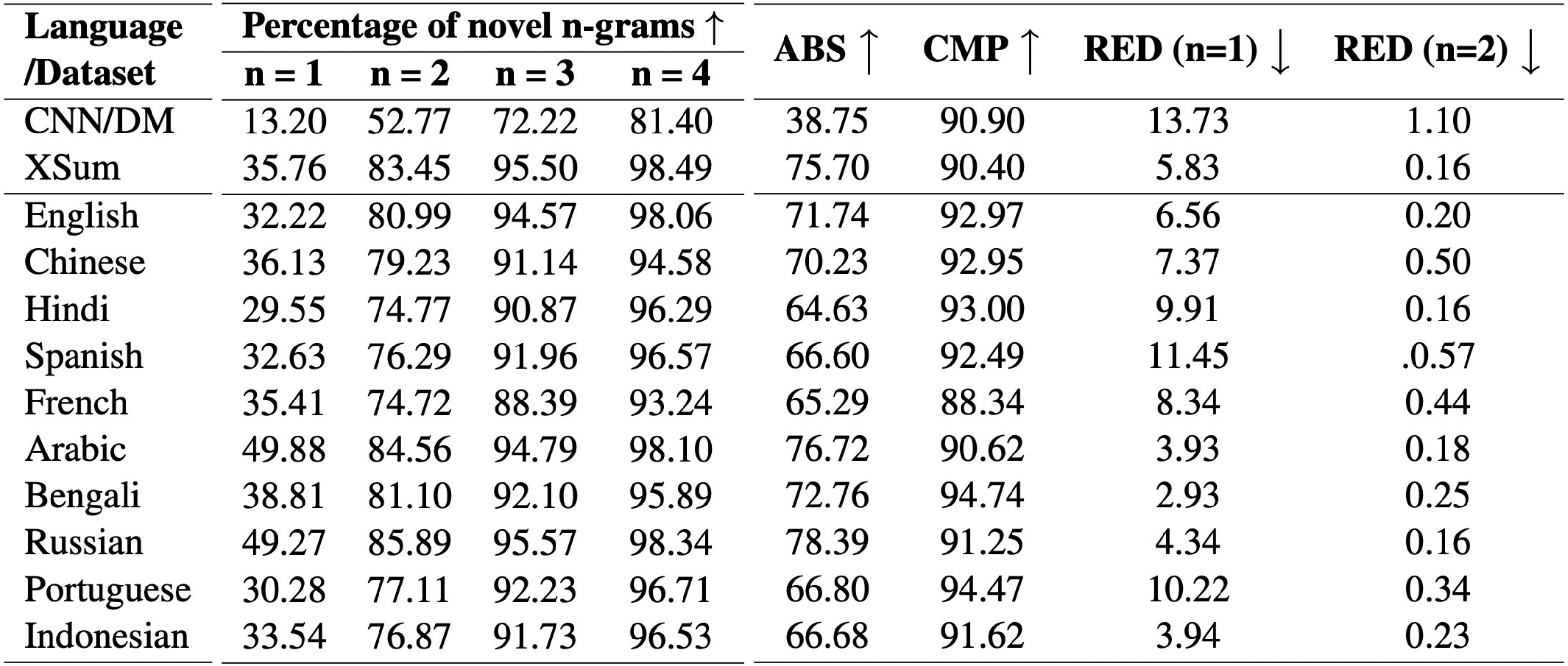
•The dataset is highly abstractive, concise, and of high quality, as indicated by human and intrinsic evaluation

Dataset Statistics



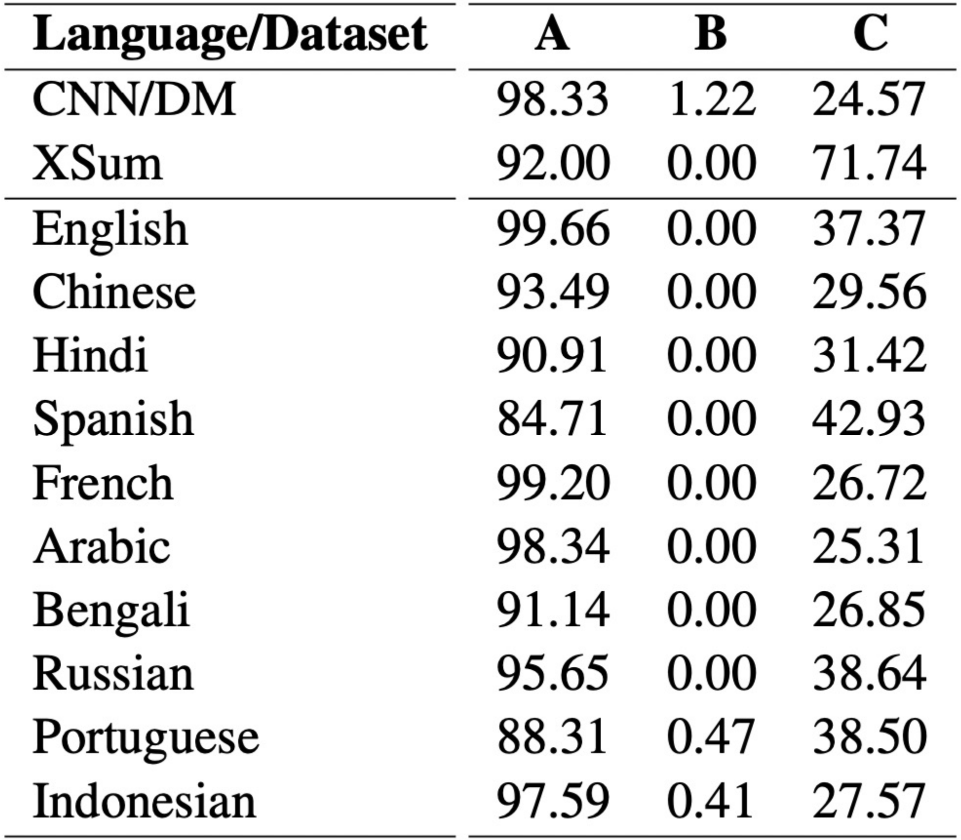
Language	#Samples	Language	#Samples	Language	#Samples
Amharic	5,461	Korean	4,281	Somali	5,636
Arabic	40327	Kyrgyz	2,315	Spanish	44,413
Azerbaijani	7,332	Marathi	11,164	Swahili	10,005
Bengali	8,226	Nepali	5,286	Tamil	17,846
Burmese	5,002	Oromo	5,738	Telugu	11,308
Chinese	39,810	Pashto	15,274	Thai	6,928
English	301,444	Persian	25,783	Tigrinya	4,827
French	9,100	Pidgin[a]	9,715	Turkish	29,510
Gujarati	9,665	Portuguese	23,521	Ukrainian	57,952
Hausa	6,313	Punjabi	8,678	Urdu	40,714
Hindi	51,715	Russian	52,712	Uzbek	4,944
Igbo	4,559	Scottish Gaelic	1,101	Vietnamese	23,468
Indonesian	44,170	Serbian (Cyrillic)	7,317	Welsh	11,596
Japanese	7,585	Serbian (Latin)	7,263	Yoruba	6,316
Kirundi	5,558	Sinhala	3,414	**Total**	**1,005,292**

Dataset Statistics…



Language /Dataset	Percentage of novel n-grams ↑				ABS ↑	CMP ↑	RED (n=1) ↓	RED (n=2) ↓
	n = 1	n = 2	n = 3	n = 4				
CNN/DM	13.20	52.77	72.22	81.40	38.75	90.90	13.73	1.10
XSum	35.76	83.45	95.50	98.49	75.70	90.40	5.83	0.16
English	32.22	80.99	94.57	98.06	71.74	92.97	6.56	0.20
Chinese	36.13	79.23	91.14	94.58	70.23	92.95	7.37	0.50
Hindi	29.55	74.77	90.87	96.29	64.63	93.00	9.91	0.16
Spanish	32.63	76.29	91.96	96.57	66.60	92.49	11.45	.0.57
French	35.41	74.72	88.39	93.24	65.29	88.34	8.34	0.44
Arabic	49.88	84.56	94.79	98.10	76.72	90.62	3.93	0.18
Bengali	38.81	81.10	92.10	95.89	72.76	94.74	2.93	0.25
Russian	49.27	85.89	95.57	98.34	78.39	91.25	4.34	0.16
Portuguese	30.28	77.11	92.23	96.71	66.80	94.47	10.22	0.34
Indonesian	33.54	76.87	91.73	96.53	66.68	91.62	3.94	0.23

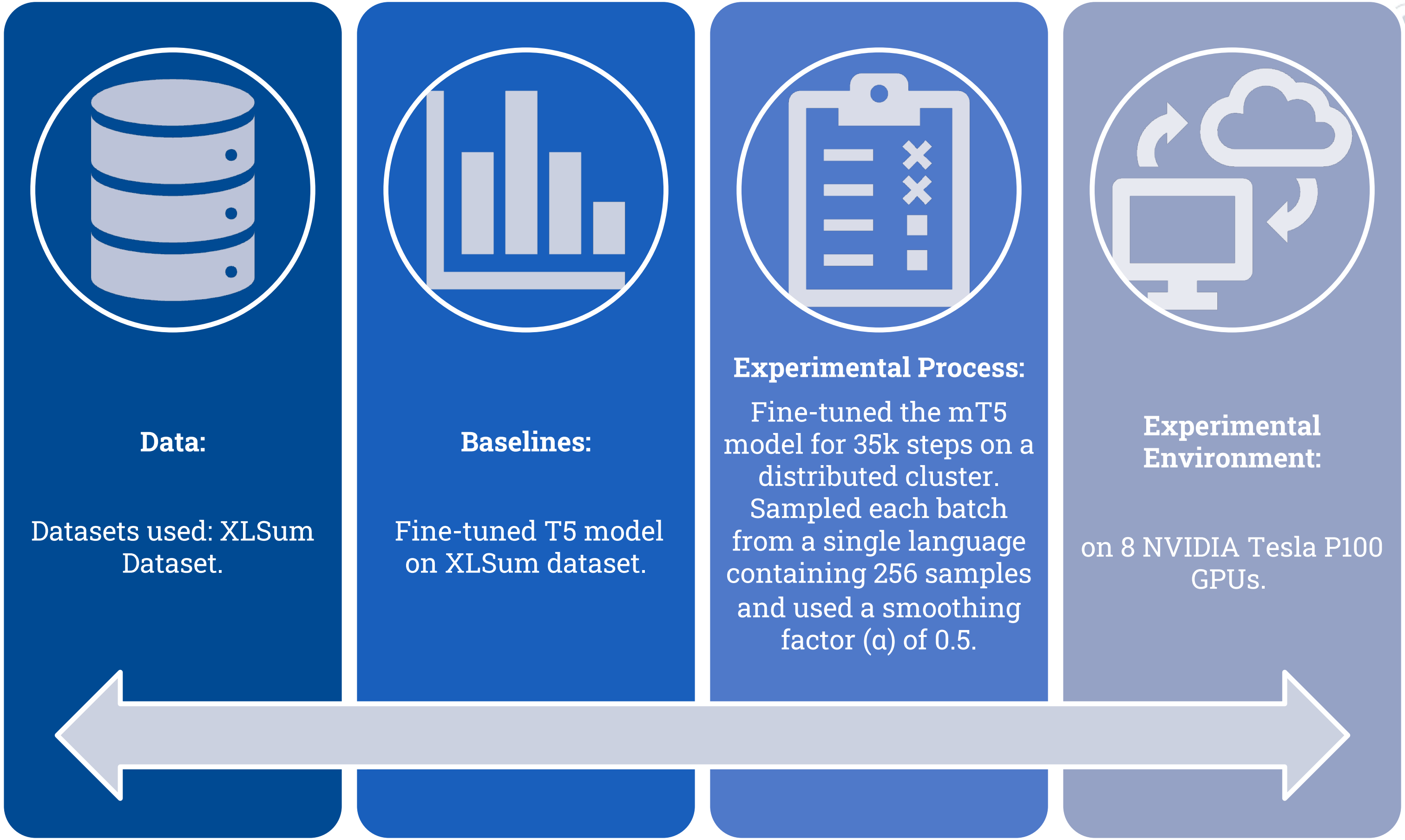
Dataset Statistics...



Language/Dataset	A	B	C
CNN/DM	98.33	1.22	24.57
XSum	92.00	0.00	71.74
English	99.66	0.00	37.37
Chinese	93.49	0.00	29.56
Hindi	90.91	0.00	31.42
Spanish	84.71	0.00	42.93
French	99.20	0.00	26.72
Arabic	98.34	0.00	25.31
Bengali	91.14	0.00	26.85
Russian	95.65	0.00	38.64
Portuguese	88.31	0.47	38.50
Indonesian	97.59	0.41	27.57

EXPERIMENTS AND RESULTS

Experimental Setup



Data:

Datasets used: XLSum Dataset.

Baselines:

Fine-tuned T5 model on XLSum dataset.

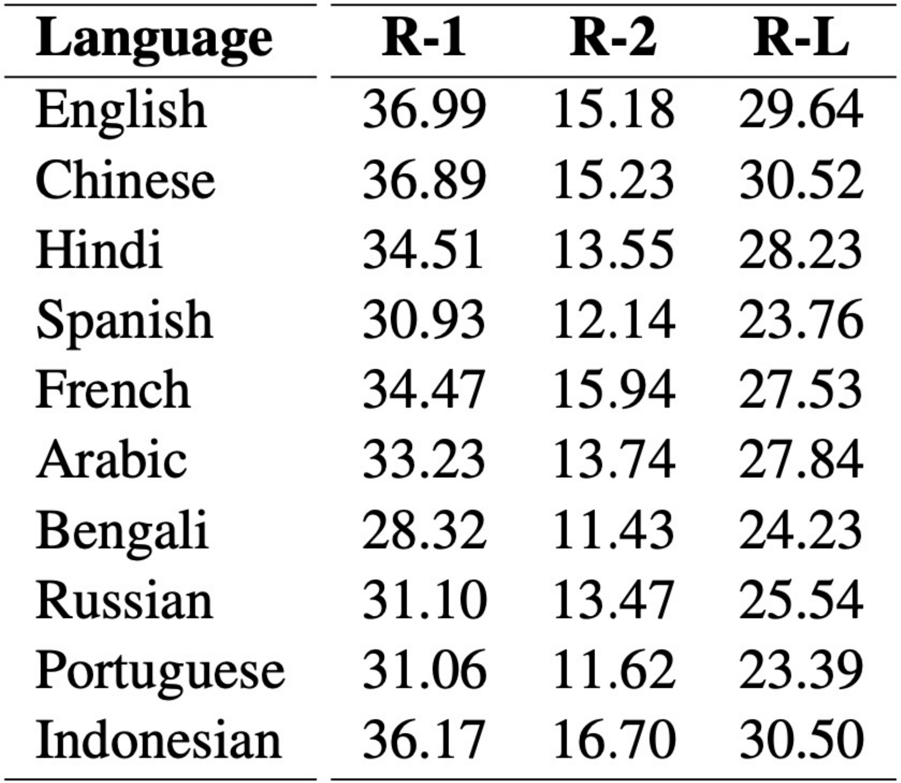
Experimental Process:

Fine-tuned the mT5 model for 35k steps on a distributed cluster. Sampled each batch from a single language containing 256 samples and used a smoothing factor (α) of 0.5.

Experimental Environment:

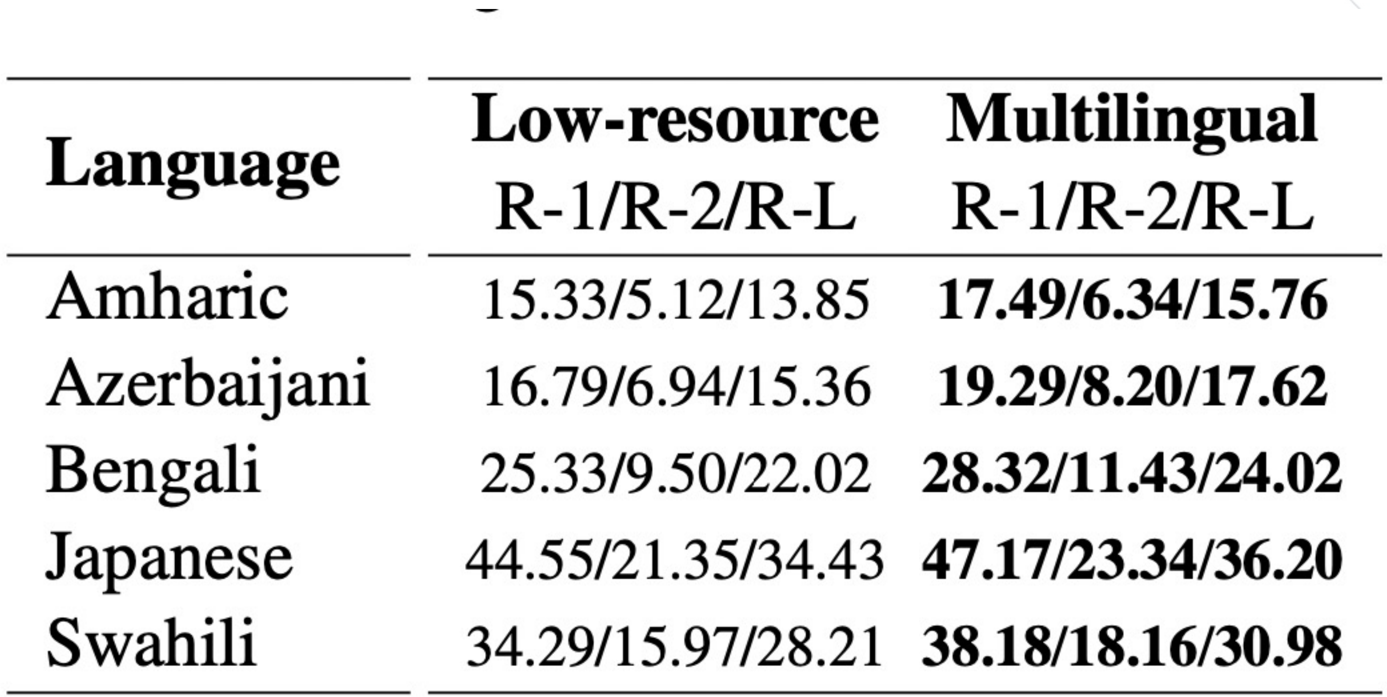
on 8 NVIDIA Tesla P100 GPUs.

Test Results



Language	R-1	R-2	R-L
English	36.99	15.18	29.64
Chinese	36.89	15.23	30.52
Hindi	34.51	13.55	28.23
Spanish	30.93	12.14	23.76
French	34.47	15.94	27.53
Arabic	33.23	13.74	27.84
Bengali	28.32	11.43	24.23
Russian	31.10	13.47	25.54
Portuguese	31.06	11.62	23.39
Indonesian	36.17	16.70	30.50

Test Results...



Language	**Low-resource**	**Multilingual**
	R-1/R-2/R-L	R-1/R-2/R-L
Amharic	15.33/5.12/13.85	**17.49/6.34/15.76**
Azerbaijani	16.79/6.94/15.36	**19.29/8.20/17.62**
Bengali	25.33/9.50/22.02	**28.32/11.43/24.02**
Japanese	44.55/21.35/34.43	**47.17/23.34/36.20**
Swahili	34.29/15.97/28.21	**38.18/18.16/30.98**

CONCLUSIONS

- The paper presents XL-Sum, a comprehensive and diverse dataset comprising 1 million professionally annotated article-summary pairs from BBC, covering 44 languages ranging from low to high-resource.
- The dataset is highly abstractive, concise, and of high quality, as indicated by human and intrinsic evaluation.
- The authors fine-tuned the mT5 model with XL-Sum and achieved competitive results on multilingual and low-resource summarization tasks.
- Multilingual training with the mT5 model demonstrated positive transfer between sister languages with morphological similarity, leading to better summarization performance

- Encouraging future research on multilingual abstractive summarization by releasing the XL-Sum dataset, curation tool, and summarization model checkpoints.

- The authors suggested investigating the use of the XL-Sum dataset for other summarization tasks, such as cross-lingual summarization as future work.

THANK YOU…