FIRE 2022
Kolkata, India

Indian Language Summarization using Pretrained Sequence-to-Sequence Models

Authors: Ashok Urlana, Sahil Manoj Bhatt, Nirmal Surange and Manish Shrivastava

Language Technologies Research Center, KCIS, IIIT Hyderabad, India

Presenter: Sahil Manoj Bhatt

Problem Statement

- Summarization – a widely explored NLP problem.
- State of research in summarization for Indic languages; lack of datasets.
- ILSUM: Shared task for Indian Language Summarization.
- Focus languages: Hindi (340 million+ speakers), Gujarati (56 million+ speakers) and 'Indian' English.

Source: https://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers

Challenges

- Little work in Indic language summarization.
- Primary reason – lack of quality data
- Recent progress in terms of dataset availability – XL-Sum, MassiveSumm.
- IndicNLG Suite – sentence summarization and headline generation tasks
- TeSum – Telugu summarization dataset
- Dataset provided for ILSUM has code-mixing and script-mixing.

Dataset

	English		Hindi		Gujarati	
#Pairs	12564		7957		8457	
	Text	Summary	Text	Summary	Text	Summary
#Avg Words	595	36.24	553	40.17	414.43	32.26
(Min, Max) Words	(1, 5717)	(1, 113)	(17, 5034)	(6, 113)	(25, 2839)	(1, 408)
#Avg Sentences	10.29	1.26	18.1	1.7	21.28	1.57
(Min, Max) Sentences	(1, 169)	(1, 17)	(1, 157)	(1, 9)	(1, 187)	(1, 46)

Article: Trump rules out changing date of November presidential election. US President Donald Trump has ruled out making any changes in the date of the November 3 presidential election because of the coronavirus pandemic. “I never even thought of changing the date of the election. Why would I do that? November 3, it's a good number,” Trump told reporters at his White House news conference.His likely Democratic opponent Joe Biden last week said Trump was considering changing the date. ""Mark my words, I think he is gonna try to kick back the election somehow, come up with some rationale why it can't be held,"" Biden said during an online fundraiser.“No, I look forward to that election and that was just made a propaganda not by him but by some of the many people that are working writing little segments. I see all of the time statements made you say something statement made per Joe Biden, Sleepy Joe,” Trump said.“He didn't make those statements. Somebody did but they said he made it. No, let him know I am not thinking about it at all. Not at all,” he said.

Summary: US President Donald Trump has ruled out making any changes in the date of the November 3 presidential election because of the coronavirus pandemic.

Article: देशभर में कोरोना के मामले लगातार बढ़ रहे हैं। दिल्ली में 3-4 दिनों से कोरोना के मामलों में कमी देखी गई है। हरियाणा में भी कुछ ऐसा ही नज़र आ रहा है। लेकिन हरियाणा के स्वास्थ्य मंत्री अनिल विज से जब कोरोना के बेकाबू होने के बारे में पूछा गया तो उन्होंने इसके लिए दिल्ली को जिम्मेदार ठहराया। अनिल विज ने कहा, 'हरियाणा के 3 जिलों में कोरोना के मामलों में तेजी से उछाल देखा गया है। दिल्ली के नज़दीक होने के कारण यहां कोरोना के मामले बढ़ रहे हैं, लेकिन घबराने की कोई जरूरत नहीं है।'इसके जवाब में दिल्ली के स्वास्थ्य मंत्री सत्येंद्र जैन ने कहा, 'ये राजनीतिक बातचीत है। मैं भी कह सकता हूं कि कितने हरियाणा के लोग दिल्ली में आकर पॉजिटिव पाए जा रहे हैं। दिल्ली में रोज़ाना करीब 1 हजार केस तो बाहर के ही होते हैं।' दिल्ली के कोरोना केस पर बोलते हुए जैन ने कहा, 'दिल्ली में आज 14-15 हजार नए केस आएंगे। ये कल के मुकाबले काफी कम हैं। दिल्ली में करीब 2.85 करोड़ वैक्सीनेशन की डोज़ लगाई जा चुकी है।'सत्येंद्र जैन ने आगे कहा, 'एलिजिबल लोगों में 100 प्रतिशत को टीका लगाया जा चुका है, 80 प्रतिशत को दूसरी खुराक लगाई गई है और 1.28 लाख लोगों को प्रिकोशनरी डोज़ लग चुकी है।' दिल्ली में रविवार को कोरोना वायरस संक्रमण के 18,286 मामले सामने आए और 28 रोगियों की मौत हुई। इसके अलावा संक्रमण दर गिरावट के साथ 27.87 फीसदी रही, जो एक दिन पहले 30.64 फीसदी थी। स्वास्थ्य विभाग के आंकड़ों में यह जानकारी दी गई है। दिल्ली में शनिवार को संक्रमण के 20,718 मामले सामने आए थे और 30 रोगियों की मौत हुई थी।देश में कोरोना की रफ्तार धीमी होती नज़र आ रही है। बीते 24 घंटे में 2 लाख 58 हजार 089 नए मामलों की पुष्टि हुई है। इस दौरान 385 लोगों की मौत हुई है। वहीं, 1 लाख 51 हजार 740 लोग ठीक हुए हैं। अभी देश में 16 लाख 56 हजार 341 एक्टिव केस हैं। पॉजिटिविटी रेट 119.65 प्रतिशत रिकॉर्ड की गई है। ओमिक्रॉन के 8 हजार 209 मामलों की पुष्टि हुई है। कल यानी 16 जनवरी के मुकाबले आज 13 हजार 113 कम मामले सामने आए हैं।

Summary: सत्येंद्र जैन ने कहा, 'ये राजनीतिक बातचीत है। मैं भी कह सकता हूं कि कितने हरियाणा के लोग दिल्ली में आकर पॉजिटिव पाए जा रहे हैं। दिल्ली में रोज़ाना करीब 1 हजार केस तो बाहर के ही होते हैं।'

Article: વિશ્વના બીજા નંબરના સૌથી મોટા સિરામીક ઉદ્યોગનુ હબ બનેલા મોરબી શહેરમાં ગેરકાયદેસર બાંધકામો છડેચોક બંધાઈ રહ્યા છે. ત્યારે ભુતકાળમાં ઈમ્પેકટ ફી મામલે થયેલી ગેરરીતિઓ ઉપરથી પરદો ઉચકવા નવનિયુક્ત ચીફ ઓફિસરે સાત દિવસમાં ઈમ્પેકટ ફી વસુલાતનું સઘળું રેકોર્ડ રજૂ કરવા જવાબદાર કર્મચારીઓને તાકીદ કરી હતી, છતાં જવાબદાર કર્મચારીઓ દ્વારા રેકર્ડ રજૂ ન કરતા આજે ચીફ ઓફિસર દ્વારા ત્રણ હંગામી, ફિક્સ વેતન કર્મચારીઓની સેવાનો અંત લાવી તાત્કાલિક અસરથી છુટા કરવા આદેશ આપ્યા હતા. જેને લઈ નગરપાલિકા કચેરીમાં ખળભળાટ મચી ગયો છે.આજ રોજ મોરબી પાલિકાના નવનિયુક્ત ચીફ ઓફિસર સંદીપસિંહ ઝાલા દ્વારા નગરપાલિકા વિસ્તારમાં ઈમ્પેકટ-ફી અધીનીયમ હેઠળ જે ગેરરીતીઓ હાથ ધરાયેલી હતી તેની તપાસ કરી દિન-7 (સાત)માં તેનો અહેવાલ અત્રેની સમક્ષ રજુ કરવાનો હુકમ કરવામાં આવ્યો હતો. પરંતુ આ મામલે જવાબદાર એવા ત્રણ કર્મચારીઓ દ્વારા અક્ષમ્ય બેદરકારી દાખવવામાં આવી હતી. નગરપાલિકાની શાખ અને સ્વભંડોળને થઇ રહેલા નુકશાન પહોચાડ્યાનું જણાવી હંગામી, ફિક્સ વેતન કર્મચારી જયદીપ સોરઠીયા, ધીરુભાઈ સુરેલીયા અને વિવેક દવેની સેવાઓને કરારને તાત્કાલિક અસરથી અંત લાવવાનો હુકમ કર્યો હતો.વધુમાં ચીફ ઓફિસર દ્વારા નગરપાલિકાના કાયમી કર્મચારી વિનુભાઈ બારહટને આ કાર્યાલય આદેશ હેઠળ બે દિવસમાં ઈમ્પેકટ-ફી અધિનિયમ હેઠળનું સમગ્ર રેકર્ડ રજુ કરવાનો આદેશ આપ્યો હતો. તેમજ કામમાં વિલંબ બદલ દિન-3 (ત્રણ)માં લેખિતમાં વિલંબ બદલનો ખુલાસો રજુ કરવા આદેશ કરતા નગરપાલિકાના કાયમી કર્મચારીઓમાં ભારે ગણગણાટ શરૂ થયો છે.

Summary: વસુલાતનું સઘળું રેકોર્ડ રજૂ કરવા કર્મચારીઓને તાકીદ કરી હોવા છતાં રેકર્ડ રજૂ કર્યો ન હતોકાયમી કર્મચારીને બે દિવસમાં તમામ રેકર્ડ રજૂ કરવા આદેશ કરાયા

Approach

- We experimented (abstractive approach) with various pretrained models (PLMs) on the ILSUM dataset.
- PLMs – pretrained using massive amounts of unlabeled text data; stimulates universal representations and improved generation quality.
- Also shown to perform extremely well on most summarization datasets.
- We thus finetune PLMs on the ILSUM dataset and perform various experiments to determine the best model for each language.

Experiments

- In many cases, models finetuned on foldwise-data gave better validation scores than models finetuned on the whole training data.
- Tuning parameters and hyperparameters (epochs, learning rate, dropout)
- Using adapters on pretrained models.
- Training on augmented data. Examples:
  - Hindi and Gujarati ILSUM datasets together
  - Hin./Guj. ILSUM dataset augmented with IndicNLG Hin./Guj. data

Experiments

Experimental setup and parameters settings

Parameters	BART	T5	ProphetNet	PEGASUS	BRIO	MBart	MT5	IndicBART
Max source length	512	512	512	512	512	512	512	512
Max target length	75	75	75	75	75	75	100	75
Batch Size	2	1	1	2	2	4	2	2
Epochs	5	5	5	5	5	5	10	10
Vocab Size	50265	32128	30522	96103	50264	250054	250112	64015
Beam Size	4	4	5	4	4	4	4	4
Learning Rate	5e-5	5e-5	5e-5	5e-4	5e-5	5e-5	5e-5	5e-5

Results

ILSUM Experiments on Validation Data. *Finetuned on the combination of Hindi and Gujarati Data

			Validation Scores		
Lang	**Model**	**Full Data / k-fold**	**R-1**	**R-2**	**R-4**
English	PEGASUS	Full Data	**56.85**	**45.92**	**43.36**
	$T5_{large}$	Full Data	56.05	45.03	42.36
	$BART_{large}$	k-fold	54.83	43.58	40.71
	PEGASUS xsum	Full Data	54.66	43.48	40.64
	BRIO	Full Data	53.57	41.86	38.81
	$BART_{large}$ xsum	k-fold	53.35	41.74	38.75
	$T5_{base}$ + Adapter	k-fold	51.91	40.07	37.1
	ProphetNet	k-fold	49.51	36.98	33.83
Hindi	IndicBART	k-fold	**60.73**	**51.26**	**47.57**
	$MT5_{base}$	k-fold	60.04	50.72	46.82
	$MT5_{base}$*	Full Data	58.65	49.09	45.08
	IndicBART-SentSumm	k-fold	58.09	47.99	43.72
	$MBart_{large}$50 + Adapters	Full Data	56.26	45.56	41.21
	$MBart_{large}$50	Full Data	55.76	44.96	40.59
Gujarati	$MBart_{large}$50	Full Data	**26.20**	**16.44**	**12.16**
	$MT5_{base}$	Full Data	25.11	15.81	11.68
	$MT5_{base}$*	Full Data	24.16	14.68	10.79
	IndicBART	k-fold	23.38	13.34	9.35
	$MBart_{large}$50 + Adapter	Full Data	21.63	13.04	9.56

Results

ILSUM scores on Test Data

			Test Scores		
Lang	**Model**	**Full Data / k-fold**	**R-1**	**R-2**	**R-4**
English	PEGASUS	Full Data	**55.83**	**44.58**	**41.8**
	$T5_{large}$	Full Data	54.73	43.08	40.12
Hindi	$MT5_{base}$	k-fold	**60.72**	**51.02**	**47.11**
	IndicBART	k-fold	58.38	48.31	44.25
Gujarati	$MBart_{large}50$	Full Data	**26.11**	16.51	12.41
	$MBart_{large}50$	Full Data (dropout=0.2)	26.07	**16.60**	**12.58**

Our team (MT-NLP, IIIT-H), ranked 1st across all three language sub-tasks (Eng., Hin. and Guj.)

Results

Article: " કિર્તેશ પટેલ, સુરત : શહેરની સાંકડી શેરીઓમાં કે જ્યાં ફાયર બ્રિગેડના (Fire Brigade) વાહનો આસાનીથી ન જઈ શકે અથવા ફાયર કર્મીઓ, અધિકારીઓ માટે જોખમ હોય તેવા વિસ્તારોમાં આગ ઓલવવાનું કામ રોબોટ (Fire Robot) કરશે. આ રોબોટિક વાહનમાં અત્યાધુનિક થર્મલ ઈમેજીંગ કેમેરા (Fire Robot Camera) વગેરેની સુવિધા હશે, આગ અકસ્માતના કોલમાં ફસાયેલી વ્યકિતઓને શોધી શકાશે અને ફસાયેલી વ્યકિતને બહાર કાઢીને જાનહાનિમાંથી બચાવી શકાશે. 1 કરોડ 42 લાખના ખર્ચે ખરીદાયેલ આ રોબોટ (Fire Robot in Surat) શુક્રવારે લોન્ચ કરવામાં આવ્યું હતું. ચોક બજાર કિલ્લાના પટાંગણમાં ફાયર ફાઈટીગ રોબોટ, 3 ફોમ કમ વોટર ટેન્ડર અને 2 ફાયર એન્જીન, 2 વોટર બોવરનને ખુલ્લું મૂકવામાં આવ્યું હતું. રોબોટ મશીન સાથે 7 ફાયર ફાયટરની ગાડીઓ પણ વાસવામાં આવ્યા છે સીસીટીવી કેમેરા સાથેનું છે આ રોબોટ મશીન. આ રોબોટ અમદાવાદની કંપની પાસેથી 1.4 કરોડનું મનપાએ ખરીદયું છે.150 મીટરની રેન્જમાં પાણી નો ફલો જઈ શકે છે. હવે કોઈ પણ આગની દુર્ઘટનમાં રોબોટથી આગ કાબુ મેળવી શકાશે. રિમોટ થી ઓપરેટ થતું રોબોટિક ફાયર મશીન અનેકરીતે સબળ છે. આજે .રોબોટ મશીનનો ડેમો કરાયો હતો. સુરતમાં અગ્નિકાંડની ઘટના 22 લોકો હોમાયા હતા નોંધનીય છે કે તક્ષશિલા દુર્ઘટના પછી સુરત ફાયર વિભાગે અગ્નિશમન સાધનો અને વાહનોમાં સતત વધારો કર્યો છે. આગ બુઝાવવાની કામગીરીમાં મદદરુપ થઇ શકે તેવા અત્યાધુનિક સાધનો વિદેશોથી મંગાવવામાં આવી રહ્યા છે. આ પહેલા ટર્ન ટેબલ લેડર, હાઈટેક કેમેરા સાથેના ફાયર ફાઇટિગ સાધનો અને હવે રોબોટની મદદથી સુરત ફાયર વિભાગ પોતાની તાકાત વધારવા જઈ રહ્યું છે."

Generated summary: "Fire Brigade in Surat - આ રોબોટિક વાહનમાં અત્યાધુનિક થર્મલ ઈમેજીંગ કેમેરા (Fire Robot Camera) વગેરેની સુવિધા હશે, આગ અકસ્માતના કોલમાં ફસાયેલી વ્યકિતઓને શોધી શકાશે"

Data Quality Assessment

- ILSUM dataset – good step forward in the field of Indic language summarization; certain limitations.
- What makes a good summary? Subjective.
- Are ROUGE scores the best way to go?
- Filters to keep in mind while creating summaries. For example:
  - Empty article text and/or summary
  - Duplicate summaries
  - Presence of prefixes and/or ellipses (...)
  - Length of summary vs length of input article text (Summary compression)
- We apply filters and carry out experiments as well.

Data Quality Assessment

Filtration counts of ILSUM data

Filters	Hindi	Gujarati	English
Dataset Size	7957	8457	12565
Empty	0	0	1
Duplicate Pairs	23	0	0
Duplicate Summary	15	113	117
Prefixes	2518	135	486
Compression <50%	11	37	182
Final Valid	5390	8172	11779
Valid %	**67.74%**	**96.63%**	**93.74%**

Data Quality Assessment

Validation set ROUGE scores on ILSUM corpus. This table reports the mean ROUGE scores and its standard deviation over 10 runs

Lang	Model	Data composition	R-1	R-2	R-L
English	PEGASUS	Original Data	52.51 ± 1.1	40.91 ± 1.36	47.81 ± 1.16
		Original + Filtered Data	51.65 ± 1.14	40.07 ± 1.25	46 ± 3.67
		Filtered Data	51.88 ± 1.25	40.37 ± 1.39	47.32 ± 1.31
		Filtered + Original Data	53.28 ± 1.18	41.82 ± 1.3	48.67 ± 1.2
	T5-large	Original Data	**53.45 ± 0.95**	**42.16 ± 1.13**	**48.97 ± 1.05**
		Original + Filtered Data	53.22 ± 1.23	42.04 ± 1.41	48.85 ± 1.31
		Filtered Data	51.9 ± 1.37	40.49 ± 1.53	47.38 ± 1.46
		Filtered + Original Data	53.33 ± 0.83	42.1 ± 0.96	48.92 ± 0.86
	BART-large	Original Data	50.25 ± 1.52	38.15 ± 1.85	45.46 ± 1.63
		Original + Filtered Data	51.42 ± 0.88	39.85 ± 1.11	46.93 ± 1
		Filtered Data	51.21 ± 1.3	39.83 ± 1.57	46.79 ± 1.38
		Filtered + Original Data	52.45 ± 1.05	40.98 ± 1.29	48 ± 1.17
Hindi	IndicBART	Original Data	26.36 ± 1.02	12.66 ± 0.73	26.28 ± 0.98
		Original + Filtered Data	21.58 ± 0.66	9.84 ± 0.76	21.45 ± 0.6
		Filtered Data	21.27 ± 0.88	9.75 ± 0.56	21.12 ± 0.86
		Filtered + Original Data	25.67 ± 1.04	12.16 ± 0.82	25.57 ± 1
	MT5-base	Original Data	**27.04 ± 1.22**	**13.21 ± 0.61**	**26.96 ± 1.22**
		Original + Filtered Data	20.33 ± 0.91	9.26 ± 0.8	20.2 ± 0.92
		Filtered Data	20.61 ± 1.55	9.47 ± 0.67	20.51 ± 1.53
		Filtered + Original Data	26.73 ± 1.11	12.83 ± 0.61	26.64 ± 1.1
Gujarati	MBart Large 50	Original Data	20.36 ± 0.67	11.65 ± 1.13	20.01 ± 0.72
		Original + Filtered Data	16.04 ± 1.12	9.23 ± 0.76	15.83 ± 1.15
		Filtered Data	12.82 ± 2.28	6.6 ± 1.54	12.38 ± 2.36
		Filtered + Original Data	19.55 ± 0.74	11.42 ± 0.43	19.2 ± 0.72
	MT5-base	Original Data	**21.55 ± 0.77**	**11.81 ± 0.78**	**21.19 ± 0.83**
		Original + Filtered Data	18.63 ± 0.93	9.23 ± 0.5	18.19 ± 0.92
		Filtered Data	9.66 ± 0.97	4.84 ± 0.56	9.53 ± 0.92
		Filtered + Original Data	20.29 ± 0.62	10.7 ± 0.52	19.84 ± 0.56

Conclusion

- For the ILSUM task, PEGASUS, MT5 and MBart give us the best results for English, Hindi and Gujarati respectively. We conclude that the transformer-based pretrained seq2seq models are capable of generating high-quality summaries for the ILSUM shared task.
- Better models finetuned exclusively on Indian languages - benefit research in the area of Indian Language Summarization
- Creation of larger, high-quality datasets for such languages will surely lead to progress in this field.

Thank you