Systematically Exploring Redundancy Reduction in Summarizing Long Documents

Wen Xiao and Giuseppe Carenini

University of British Columbia

What is Extractive Summarization?

- select sentences that can best represent the whole document
- can be regarded as a sequence labeling problem

(1) A 6.3-magnitude earthquake struck early sunday off Indonesia, according to the U.S. geological survey.
(2) The quake rattled a remote swath of sea between the Pacific and Indian oceans, north of Australia and east of Timor-leste, some 5.6 miles (9 kilometers) deep, according to the U.S. agency.
(3) It was centered approximately 212 miles (340 kilometers) west-northwest of Saumlaki in Indonesia 's Tanimbar Islands, 217 miles east-northeast of Dili, Timor-leste, and 226 miles of Ambon, Indonesia.
(4) Neither the Pacific Tsunami Warning Center nor the Japan Meteorological Agency issued Tsunami Warnings or advisories immediately after the tremor.

Properties of Good Summary

A good summary should be

Properties of Good Summary

A good summary should be

- informative

Properties of Good Summary

A good summary should be

- informative
- salient

Properties of Good Summary

A good summary should be

- informative
- salient
- **non-redundant**

Properties of Good Summary

A good summary should be

- informative
- salient
- **non-redundant**

Previous neural models focus more on the informativeness, and in this work, we aim to **reduce redundancy while keeping the informativeness** in the generated summary.

How to measure redundancy in the text?

- **Unique N-gram Ratio**: measures n-grams uniqueness. [PXS17a]

$$Uniq\_ngram\_ratio = \frac{|uniq\_n\_gram|}{|n\_gram|}$$

- **Normalized Inverse of Diversity (NID)**: captures redundancy, as the inverse of a diversity metric with length normalization. Diversity is defined as the entropy of unigrams in the document [FRBK17].

$$NID = 1 - \frac{entropy(D)}{log(|D|)}$$

Document is more redundant with low Unique N-gram Ratio and high NID.

Analyze Redundancy of Documents

- News: CNNDM, Xsum
- Scientific Paper: Pubmed, arXiv

Analyze Redundancy of Documents

- News: CNNDM, Xsum
- Scientific Paper: Pubmed, arXiv

Datasets	# Doc.	# w./doc.	# w./sent.	NID	Uni-%	Bi-%	Tri-%
Xsum	203k	429	22.8	0.188	54.00	90.22	97.28
CNNDM	270k	823	19.9	0.205	41.76	83.40	93.87
Pubmed	115k	3142	35.1	0.255	26.86	65.14	80.33
arXiv	201k	6081	29.2	0.267	22.51	61.81	82.93

Analyze Redundancy of Documents

- News: CNNDM, Xsum
- Scientific Paper: Pubmed, arXiv

Datasets	# Doc.	# w./doc.	# w./sent.	NID	Uni-%	Bi-%	Tri-%
Xsum	203k	429	22.8	0.188	54.00	90.22	97.28
CNNDM	270k	823	19.9	0.205	41.76	83.40	93.87
Pubmed	115k	3142	35.1	0.255	26.86	65.14	80.33
arXiv	201k	6081	29.2	0.267	22.51	61.81	82.93

Findings:

- Scientific paper tend to be much longer than the news articles

Analyze Redundancy of Documents

- News: CNNDM, Xsum
- Scientific Paper: Pubmed, arXiv

Datasets	# Doc.	# w./doc.	# w./sent.	NID	Uni-%	Bi-%	Tri-%
Xsum	203k	429	22.8	0.188	54.00	90.22	97.28
CNNDM	270k	823	19.9	0.205	41.76	83.40	93.87
Pubmed	115k	3142	35.1	0.255	26.86	65.14	80.33
arXiv	201k	6081	29.2	0.267	22.51	61.81	82.93

Findings:

- Scientific paper tend to be much longer than the news articles
- Redundancy is a more serious problem in scientific paper

Analyze Redundancy of Documents

- News: CNNDM, Xsum
- Scientific Paper: Pubmed, arXiv

Datasets	# Doc.	# w./doc.	# w./sent.	NID	Uni-%	Bi-%	Tri-%
Xsum	203k	429	22.8	0.188	54.00	90.22	97.28
CNNDM	270k	823	19.9	0.205	41.76	83.40	93.87
Pubmed	115k	3142	35.1	0.255	26.86	65.14	80.33
arXiv	201k	6081	29.2	0.267	22.51	61.81	82.93

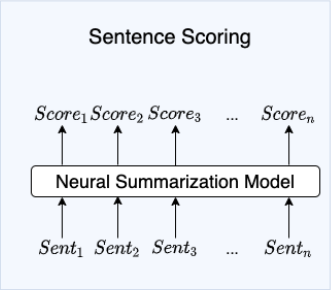
Findings:

- Scientific paper tend to be much longer than the news articles
- Redundancy is a more serious problem in scientific paper
- The sentences in the scientific paper datasets tend to be longer than in the news datasets

Thus in this paper, we focus only on the scientific paper domain.

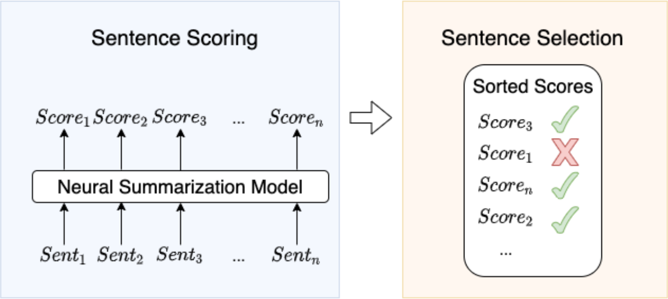
A Common Framework of Neural Summarizers

- **Sentence Scoring**: measure the importance of each sentence in the document.



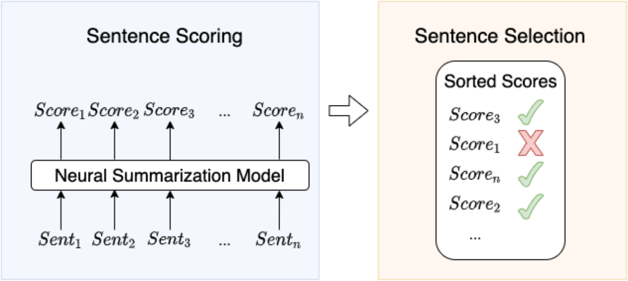
A Common Framework of Neural Summarizers

- **Sentence Scoring**: measure the importance of each sentence in the document.
- **Sentence Selection**: select sentences based on the importance score (and/or other measurements).



Categories of Redundancy Reduction Methods

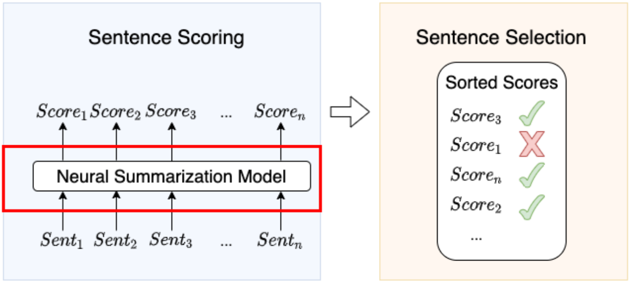
Based on When and How the redundancy is considered, we organize the redundancy reduction methods into three categories:



Categories of Redundancy Reduction Methods

Based on When and How the redundancy is considered, we organize the redundancy reduction methods into three categories:

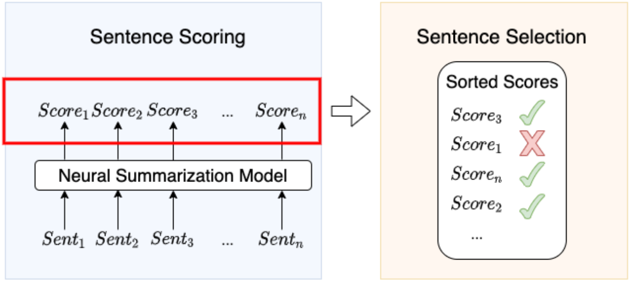
A When Design The Architecture, Implicitly



Categories of Redundancy Reduction Methods

Based on When and How the redundancy is considered, we organize the redundancy reduction methods into three categories:

- **A** When Design The Architecture, Implicitly
- **B** When Compute Scores For Sentences, Explicitly



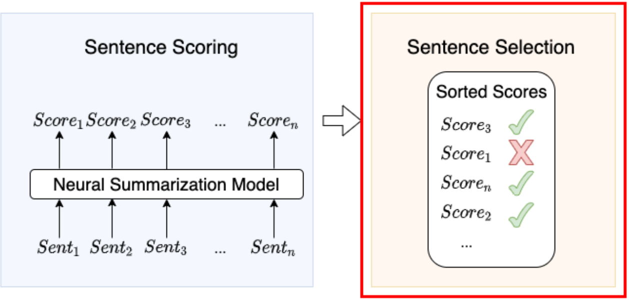
Categories of Redundancy Reduction Methods

Based on When and How the redundancy is considered, we organize the redundancy reduction methods into three categories:

A When Design The Architecture, Implicitly

B When Compute Scores For Sentences, Explicitly

C When Select Setences Based On Scores, Explicitly



The Baseline Models - Naive MMR

- traditional extractive summarization method
- ranks the candidate sentences with a balance between **informativeness** and **redundancy** with a balance factor λ

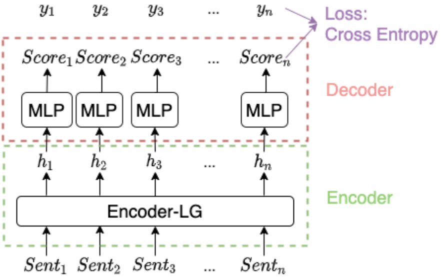
$$\begin{aligned} MMRScore = \arg \max_{s_i \in D \setminus \hat{S}} [\lambda Sim_1(s_i, Q) \quad & \#Informativeness \\ - (1-\lambda) \max_{s_j \in \hat{S}} Sim_2(s_i, s_j)] \quad & \#Redundancy \end{aligned}$$

The Baseline Models - ExtSum-LG

To compare different redundancy reduction methods fairly, we adapt all the methods into the baseline model - ExtSum-LG[XC19], as it

- is the SOTA summarizer on both scientific paper datasets
- is a non auto-regressive model
- doesn't consider redundancy aspect.

Sentence Scoring:



Sentence Selection: Greedily pick top k sentences

Overview of Current Methods

Categ.	*Methods*	*Sent. Scor.*			*Sent. Sel.*
		Encoder	*Decoder*	*Loss Func.*	
BSL	Naive MMR	Cosine Similarity			MMR Select
BSL	ExtSum-LG	Enc. LG	MLP	Cross Entropy (CE)	Greedy

Category A

SR Decoder:

- Auto-regressive SummaRuNNer Decoder [NZZ17], taking consideration of previous predictions.

NeuSum Decoder:

- Auto-regressive NeuSum Decoder[ZYW+18]
- Learn the relative gain of each sentence
- Loss function: KL Divergence

Overview of Current Methods

Categ.	*Methods*	*Sent. Scor.*			*Sent. Sel.*
		Encoder	*Decoder*	*Loss Func.*	
BSL	Naive MMR	Cosine Similarity			MMR Select
BSL	ExtSum-LG	Enc. LG	MLP	Cross Entropy (CE)	Greedy
A	+ SR Decoder	Enc. LG	SR Dec.	CE	Greedy
A	+ NeuSum Decoder	Enc. LG	NeuSum Dec.	KL Divergence	Greedy

Category B - RdLoss

- Add a redundancy loss term L_{rd} to the original loss function
- Explicitly learn to reduce the score of redundant sentences.

$$
\begin{aligned}
L &= \beta L_{ce} + (1-\beta) L_{rd} \\
L_{rd} &= \sum_{i=1}^{n} \sum_{j=1}^{n} P(y_i) P(y_j) Sim(s_i, s_j)
\end{aligned}
$$

Overview of Current Methods

Categ.	*Methods*	*Sent. Scor.*			*Sent. Sel.*
		Encoder	*Decoder*	*Loss Func.*	
BSL	Naive MMR	Cosine Similarity			MMR Select
BSL	ExtSum-LG	Enc. LG	MLP	Cross Entropy (CE)	Greedy
A	+ SR Decoder	Enc. LG	SR Dec.	CE	Greedy
A	+ NeuSum Decoder	Enc. LG	NeuSum Dec.	KL Divergence	Greedy
B	**+ RdLoss**	Enc. LG	MLP	CE + Red. Loss1	Greedy

Category C - Trigram Blocking

- A simplified version of MMR method [PXS17b]
- Widedly used in recent summarization models (e.g. BERTSUM [LL19])
- In the sentence selection phase, the current candidate is added to the summary only if it **does not have trigram overlap** with the previous selected sentences
- Otherwise, the current candidate sentence is ignored and the next one is checked

Category C - MMR-Select

- Inspired by the traditional MMR method
- Balance the informativeness and redundancy in a more soft and flexible way

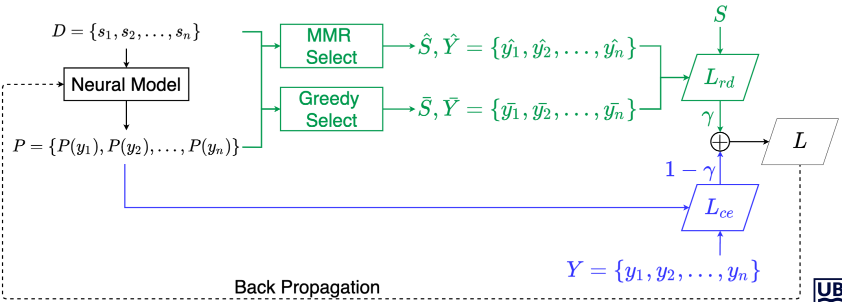
$$\text{MMR-Select} = \arg \max_{s_i \in D \setminus \hat{S}} [\text{MMR-score}_i]$$

$$\text{MMR-score}_i = \lambda P(y_i) - (1 - \lambda) \max_{s_j \in \hat{S}} Sim(s_i, s_j)$$

λ is a balance factor.

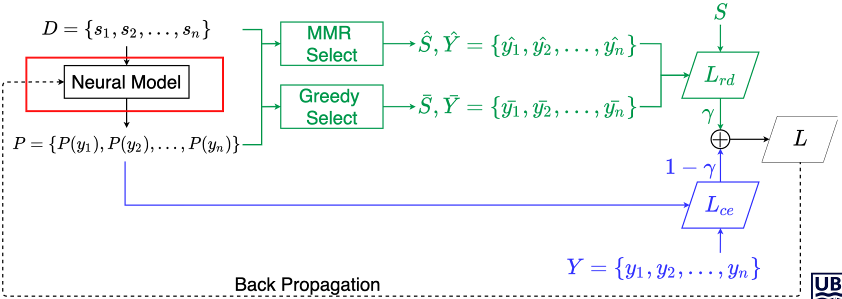
Category C - MMR-Select+

- Finetune the neural model based on MMR-Select
- To promote synergy between Sentence Scoring and Sentence Selection phases
- The Sentence Scoring combines three components:



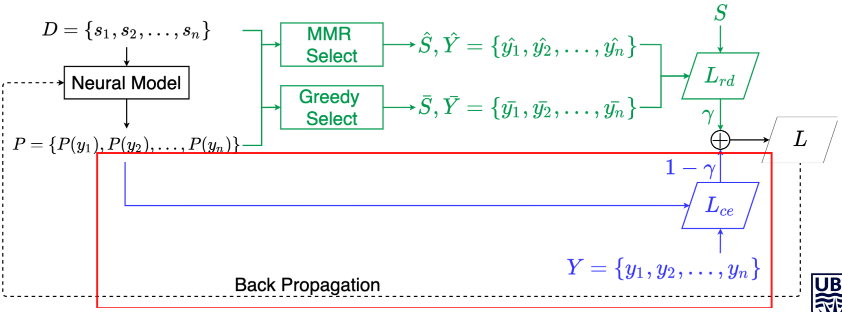
Category C - MMR-Select+

- Finetune the neural model based on MMR-Select
- To promote synergy between Sentence Scoring and Sentence Selection phases
- The Sentence Scoring combines three components:
  - The neural model



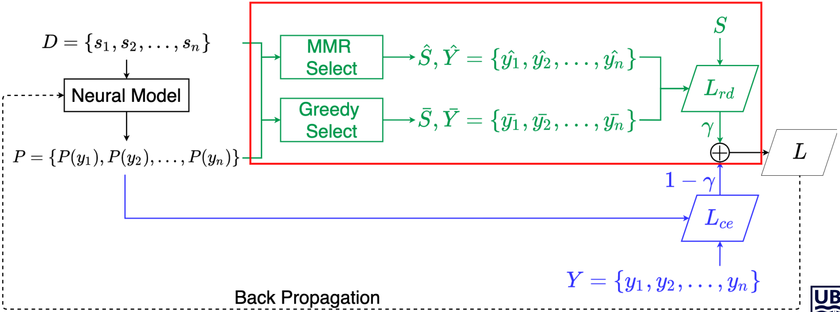
Category C - MMR-Select+

- Finetune the neural model based on MMR-Select
- To promote synergy between Sentence Scoring and Sentence Selection phases
- The Sentence Scoring combines three components:
  - The neural model
  - The original cross-entropy loss L_{ce}



Category C - MMR-Select+

- Finetune the neural model based on MMR-Select
- To promote synergy between Sentence Scoring and Sentence Selection phases
- The Sentence Scoring combines three components:
  - The neural model
  - The original cross-entropy loss L_{ce}
  - An RL mechanism whose loss is L_{rd}



Category C - MMR-Select+

$$L_{rd} = -(r(\hat{S}) - r(\bar{S})) \sum_{i=1}^{n} \log P(\hat{y}_i)$$

- L_{rd} is the **inverse expected reward** based on the ROUGE score of $\hat{S}$ (generated by MMR-Select) weighted by the probability of the $\hat{Y}$ labels in the log space.
- We adopt the **self-restriction strategy**[PXS17a] by adding a baseline summary $\bar{S}$, which is generated by Greedy algorithm on $P(y)$
- It only positively reward summaries which are better than the baseline.

Overview of All Methods

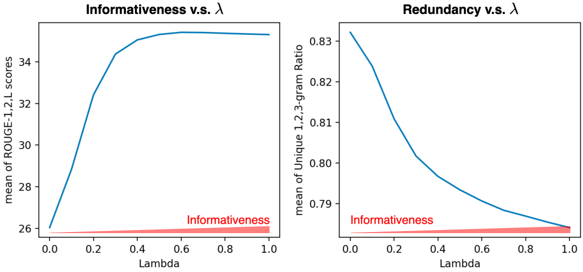
Categ.	*Methods*	*Sent. Scor.*			*Sent. Sel.*
		Encoder	*Decoder*	*Loss Func.*	
BSL	Naive MMR	Cosine Similarity			MMR Select
BSL	ExtSum-LG	Enc. LG	MLP	Cross Entropy (CE)	Greedy
A	+ SR Decoder	Enc. LG	SR Dec.	CE	Greedy
A	+ NeuSum Decoder	Enc. LG	NeuSum Dec.	KL Divergence	Greedy
B	**+ RdLoss**	Enc. LG	MLP	CE + Red. Loss1	Greedy
C	+ Trigram Blocking	Enc. LG	MLP	CE	Trigram Blocking
C	**+ MMR-Select**	Enc. LG	MLP	CE	MMR Select
C	**+ MMR-Select+**	Enc. LG	MLP	CE + Red. Loss2	MMR Select

Experiment

- Dataset: Pubmed, arXiv
- Metric for informativeness: ROUGE-1,2, L
- Metric for redundancy: Unique N-gram Ratio, NID

[0] All the hyper-parameter settings can be found in the paper.

Informativeness & Redundancy With MMR-Select

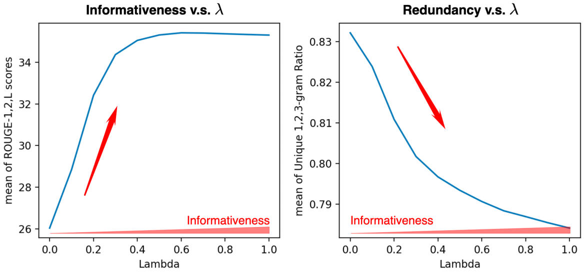


Recall:

$$\text{MMR-score}_i = \lambda P(y_i) - (1-\lambda)\max_{s_j \in \hat{S}} Sim(s_i, s_j)$$

To explore the balance between **informativeness** and **non-redundancy**, we finetune λ in MMR-Select on the validation set.

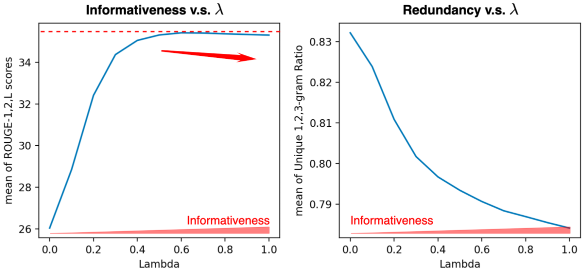
Informativeness & Redundancy With MMR-Select



Findings:

- Consistent with previous work[JKMH19], there is a trade-off between informativeness and non-redundancy.

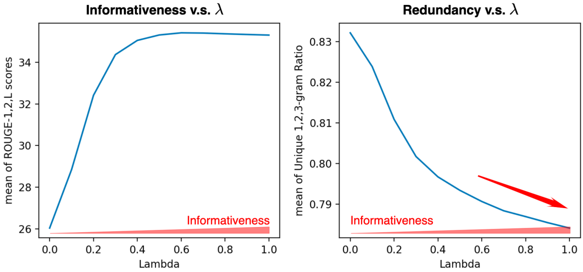
Informativeness & Redundancy With MMR-Select



Findings:

- Consistent with previous work[JKMH19], there is a trade-off between informativeness and non-redundancy.
- There is an upper bound on how much the generated summary can match the ground-truth summary.

Informativeness & Redundancy With MMR-Select



Findings:

- Consistent with previous work[JKMH19], there is a trade-off between informativeness and non-redundancy.
- There is an upper bound on how much the generated summary can match the ground-truth summary.
- The redundancy in the generated summary continued to increase as the redundancy component weigh less.

Experiment Results - Redundancy

Categ.	Model	Pubmed				arXiv			
		Uni-%	Bi-%	Tri-%	NID	Uni-%	Bi-%	Tri-%	NID
-	Naive MMR	56.55	90.93	96.95	0.1881	53.01	88.82	96.28	0.1992
-	ExtSum-LG	53.02	87.29	94.37	0.2066	52.17	87.19	95.38	0.2088
A	+SR Dec.	52.88	87.17	94.32	0.2070	51.98	87.08	95.31	0.2097
A	+NeuSum Dec.	54.88 †	88.71 †	95.13 †	0.1993 †	-	-	-	-
B	**+RdLoss**	53.23 †	87.41	94.43	0.2052 †	52.17	87.20	95.36	0.2085
C	+Tri-Blocking	**57.58** †	**93.05** †	**98.56** †	**0.1818** †	**56.12** †	**92.38** †	**98.94** †	**0.1876** †
C	**+MMR-Sel.**	53.76 †	88.04 †	94.96 †	0.2022	52.80 †	87.64 †	95.40	0.2055 †
C	**+MMR-Sel.+**	53.93 †	88.32	95.14	0.2014	52.76 †	87.78 †	95.70 †	0.2055 †
-	Oracle	56.66	89.25	95.55	0.2036	56.74	90.81	96.82	0.2029
-	Reference	56.69	89.45	95.95	0.2005	58.92	90.13	97.02	0.1970

Experiment Results - Redundancy

Categ.	Model	Pubmed				arXiv			
		Uni-%	Bi-%	Tri-%	NID	Uni-%	Bi-%	Tri-%	NID
-	Naive MMR	56.55	90.93	96.95	0.1881	53.01	88.82	96.28	0.1992
-	ExtSum-LG	53.02	87.29	94.37	0.2066	52.17	87.19	95.38	0.2088
A	+SR Dec.	52.88	87.17	94.32	0.2070	51.98	87.08	95.31	0.2097
A	+NeuSum Dec.	54.88 †	88.71 †	95.13 †	0.1993 †	-	-	-	-
B	**+RdLoss**	53.23 †	87.41	94.43	0.2052 †	52.17	87.20	95.36	0.2085
C	+Tri-Blocking	**57.58** †	**93.05** †	**98.56** †	**0.1818** †	**56.12** †	**92.38** †	**98.94** †	**0.1876** †
C	**+MMR-Sel.**	53.76 †	88.04 †	94.96 †	0.2022	52.80 †	87.64 †	95.40	0.2055 †
C	**+MMR-Sel.+**	53.93 †	88.32	95.14	0.2014	52.76 †	87.78 †	95.70 †	0.2055 †
-	Oracle	56.66	89.25	95.55	0.2036	56.74	90.81	96.82	0.2029
-	Reference	56.69	89.45	95.95	0.2005	58.92	90.13	97.02	0.1970

Findings:

- Trigram Blocking makes the largest improvement on redundancy reduction

Experiment Results - Redundancy

Categ.	Model	Pubmed				arXiv			
		Uni-%	Bi-%	Tri-%	NID	Uni-%	Bi-%	Tri-%	NID
-	Naive MMR	56.55	90.93	96.95	0.1881	53.01	88.82	96.28	0.1992
-	ExtSum-LG	53.02	87.29	94.37	0.2066	52.17	87.19	95.38	0.2088
A	+SR Dec.	52.88	87.17	94.32	0.2070	51.98	87.08	95.31	0.2097
A	+NeuSum Dec.	54.88 †	88.71 †	95.13 †	0.1993 †	-	-	-	-
B	**+RdLoss**	53.23 †	87.41	94.43	0.2052 †	52.17	87.20	95.36	0.2085
C	+Tri-Blocking	**57.58** †	**93.05** †	**98.56** †	**0.1818** †	**56.12** †	**92.38** †	**98.94** †	**0.1876** †
C	**+MMR-Sel.**	53.76 †	88.04 †	94.96 †	0.2022	52.80 †	87.64 †	95.40	0.2055 †
C	**+MMR-Sel.+**	53.93 †	88.32	95.14	0.2014	52.76 †	87.78 †	95.70 †	0.2055 †
-	Oracle	56.66	89.25	95.55	0.2036	56.74	90.81	96.82	0.2029
-	Reference	56.69	89.45	95.95	0.2005	58.92	90.13	97.02	0.1970

Findings:

- Trigram Blocking makes the largest improvement on redundancy reduction
- Almost all the methods can effectively reduce redundancy except for SR Decoder.

Experiment Results - Redundancy

Categ.	Model	Pubmed				arXiv			
		Uni-%	Bi-%	Tri-%	NID	Uni-%	Bi-%	Tri-%	NID
-	Naive MMR	56.55	90.93	96.95	0.1881	53.01	88.82	96.28	0.1992
-	ExtSum-LG	53.02	87.29	94.37	0.2066	52.17	87.19	95.38	0.2088
A	+SR Dec.	52.88	87.17	94.32	0.2070	51.98	87.08	95.31	0.2097
A	+NeuSum Dec.	54.88 †	88.71 †	95.13 †	0.1993 †	-	-	-	-
B	**+RdLoss**	53.23 †	87.41	94.43	0.2052 †	52.17	87.20	95.36	0.2085
C	+Tri-Blocking	**57.58** †	**93.05** †	**98.56** †	**0.1818** †	**56.12** †	**92.38** †	**98.94** †	**0.1876** †
C	**+MMR-Sel.**	53.76 †	88.04 †	94.96 †	0.2022	52.80 †	87.64 †	95.40	0.2055 †
C	**+MMR-Sel.+**	53.93 †	88.32	95.14	0.2014	52.76 †	87.78 †	95.70 †	0.2055 †
-	Oracle	56.66	89.25	95.55	0.2036	56.74	90.81	96.82	0.2029
-	Reference	56.69	89.45	95.95	0.2005	58.92	90.13	97.02	0.1970

Findings:

- Trigram Blocking makes the largest improvement on redundancy reduction
- Almost all the methods can effectively reduce redundancy except for SR Decoder.
- By injecting the RL mechanism, the MMR-Select+ works better than MMR-Select, especally on the Pubmed dataset.

Experiment Results - Informativeness

Categ.	Model	Pubmed			arXiv		
		ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
-	Naive MMR	37.46	11.25	32.22	33.74	8.50	28.36
-	ExtSum-LG	45.18	20.20	40.72	43.77	17.50	38.71
A	+SR Dec.	45.18	20.16	40.69	43.92	17.65	38.83
A	+NeuSum Dec.	44.54	19.66	40.42	-	-	-
B	**+RdLoss**	45.30 †	**20.42** †	40.95 †	**44.01** †	**17.79** †	**39.09** †
C	+Tri-Blocking	43.33	17.67	39.01	42.75	15.73	37.85
C	**+MMR-Sel.**	45.29 †	20.30 †	40.90 †	43.81	17.41	38.94
C	**+MMR-Sel.+**	**45.39** †	20.37 †	**40.99** †	43.87 †	17.50	38.97 †
-	Oracle	55.05	27.48	49.11	53.89	23.07	46.54

Experiment Results - Informativeness

Categ.	Model	Pubmed			arXiv		
		ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
-	Naive MMR	37.46	11.25	32.22	33.74	8.50	28.36
-	ExtSum-LG	45.18	20.20	40.72	43.77	17.50	38.71
A	+SR Dec.	45.18	20.16	40.69	43.92	17.65	38.83
A	+NeuSum Dec.	44.54	19.66	40.42	-	-	-
B	**+RdLoss**	45.30 †	**20.42** †	40.95 †	**44.01** †	**17.79** †	**39.09** †
C	+Tri-Blocking	43.33	17.67	39.01	42.75	15.73	37.85
C	**+MMR-Sel.**	45.29 †	20.30 †	40.90 †	43.81	17.41	38.94
C	**+MMR-Sel.+**	**45.39** †	20.37 †	**40.99** †	43.87 †	17.50	38.97 †
-	Oracle	55.05	27.48	49.11	53.89	23.07	46.54

Findings:

- The three new methods can reduce redundancy significantly while also improving the informativeness significantly.

Experiment Results - Informativeness

Categ.	Model	Pubmed			arXiv		
		ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
-	Naive MMR	37.46	11.25	32.22	33.74	8.50	28.36
-	ExtSum-LG	45.18	20.20	40.72	43.77	17.50	38.71
A	+SR Dec.	45.18	20.16	40.69	43.92	17.65	38.83
A	+NeuSum Dec.	44.54	19.66	40.42	-	-	-
B	**+RdLoss**	45.30 †	**20.42** †	40.95 †	**44.01** †	**17.79** †	**39.09** †
C	+Tri-Blocking	43.33	17.67	39.01	42.75	15.73	37.85
C	**+MMR-Sel.**	45.29 †	20.30 †	40.90 †	43.81	17.41	38.94
C	**+MMR-Sel.+**	**45.39** †	20.37 †	**40.99** †	43.87 †	17.50	38.97 †
-	Oracle	55.05	27.48	49.11	53.89	23.07	46.54

Findings:

- The three new methods can reduce redundancy significantly while also improving the informativeness significantly.
- Both Trigram Blocking and NeuSum Decoder effectively reduce redundancy, but hurt the informativeness, contrast with the exp. on news. [LL19][ZYW+18]

Experiment Results - Informativeness

Categ.	Model	Pubmed			arXiv		
		ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
-	Naive MMR	37.46	11.25	32.22	33.74	8.50	28.36
-	ExtSum-LG	45.18	20.20	40.72	43.77	17.50	38.71
A	+SR Dec.	45.18	20.16	40.69	43.92	17.65	38.83
A	+NeuSum Dec.	44.54	19.66	40.42	-	-	-
B	**+RdLoss**	45.30 †	**20.42** †	40.95 †	**44.01** †	**17.79** †	**39.09** †
C	+Tri-Blocking	43.33	17.67	39.01	42.75	15.73	37.85
C	**+MMR-Sel.**	45.29 †	20.30 †	40.90 †	43.81	17.41	38.94
C	**+MMR-Sel.+**	**45.39** †	20.37 †	**40.99** †	43.87 †	17.50	38.97 †
-	Oracle	55.05	27.48	49.11	53.89	23.07	46.54

Findings:

- The three new methods can reduce redundancy significantly while also improving the informativeness significantly.
- Both Trigram Blocking and NeuSum Decoder effectively reduce redundancy, but hurt the informativeness, contrast with the exp. on news. [LL19][ZYW+18]
- Compared with MMR-Select, MMR-Select+ works better on both redundancy and informativeness aspects.

Conclusion

- We find that longer documents tend to be more redundant, by examining large-scale summarization datasets
- We systematically explore and compare existing and newly proposed redundancy reduction methods in extractive summarization for long documents
- With the new redundancy reduction methods, the new model beats the original SOTA model on both informativeness and redundancy

Future Work

- Do experiments with generating summaries at **finer granularity** than sentences (sub-sentences, EDUs, etc.)
- Explore the methods on short documents, i.e. news articles.
- When considering redundancy in the loss function, use a pre-trained neural model to compute the similarity between sentences, instead of cosine similarity
- Human evaluation

Thanks!

References I

- Guy Feigenblat, Haggai Roitman, Odellia Boni, and David Konopnicki, *Unsupervised query-focused multi-document summarization using the cross entropy method*, Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (New York, NY, USA), SIGIR '17, Association for Computing Machinery, 2017, p. 961–964.
- Taehee Jung, Dongyeop Kang, Lucas Mentch, and Eduard Hovy, *Earlier Isn't Always Better: Sub-aspect Analysis on Corpus and System Biases in Summarization*, 3322–3333.
- Yang Liu and Mirella Lapata, *Text summarization with pretrained encoders*, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (Hong Kong, China), Association for Computational Linguistics, November 2019, pp. 3730–3740.

References II

- Ramesh Nallapati, Feifei Zhai, and Bowen Zhou, *Summarunner: A recurrent neural network based sequence model for extractive summarization of documents*, Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI'17, AAAI Press, 2017, pp. 3075–3081.
- Romain Paulus, Caiming Xiong, and Richard Socher, *A deep reinforced model for abstractive summarization*, CoRR **abs/1705.04304** (2017).
- ———, *A deep reinforced model for abstractive summarization*, CoRR **abs/1705.04304** (2017).

References III

- Wen Xiao and Giuseppe Carenini, *Extractive summarization of long documents by combining global and local context*, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (Hong Kong, China), Association for Computational Linguistics, November 2019, pp. 3011–3021.
- Qingyu Zhou, Nan Yang, Furu Wei, Shaohan Huang, Ming Zhou, and Tiejun Zhao, *Neural document summarization by jointly learning to score and select sentences*, Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (Melbourne, Australia), Association for Computational Linguistics, July 2018, pp. 654–663.