

NewSumm Lightning Talk

AGGSHAP: Analyzing Multi-Sentence Aggregation in Abstractive Summarization via the Shapley Value

Motivation

Jingyi He, Meng Cao, Jackie Chi Kit Cheung

- How do we measure multi-sentence aggregation in text summarization?
- Abstractiveness vs Aggregation? Go beyond word overlap!



White House weighing whether Obama should meet with Raul Castro.

[...] one of the big questions is whether Obama will [...] have a face-to-face meeting with Raul Castro. [...]

[...] And so what the White House is going to be weighing is whether this meeting would be a way to generate more progress.[...]

Both summary sentences use few novel n-grams..

(a) Multi-sentence Fusion



Experts question if packed out planes are putting passengers at risk.

[...] some experts are questioning if having such packed out planes is putting passengers at risk.[...]

(b) Single-sentence Simplification

AGGSHAP: Analyzing Multi-Sentence Aggregation in Abstractive Summarization via the Shapley Value

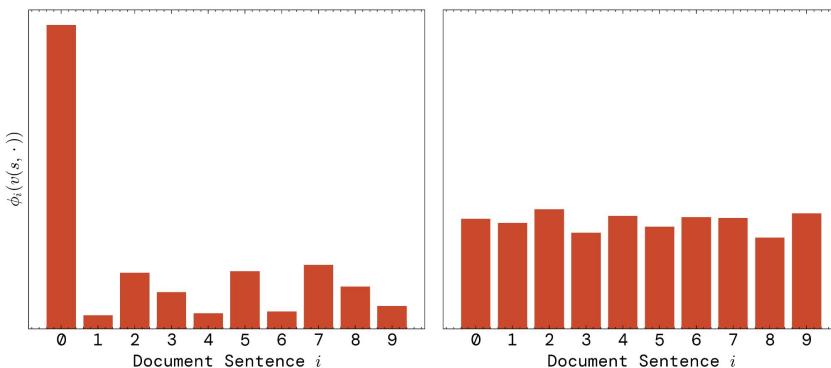
Motivation

- How do we measure multi-sentence aggregation in text summarization?
- Abstractiveness vs Aggregation? Go beyond word overlap!

Proposal - AGGSHAP

The more document sentences cover information in a summary sentence, the higher the degree of aggregation of the summary sentence

Low vs high degree of aggregation



Findings

- AGGSHAP effectively distinguishes **multi-sentence fusion** from **extraction or paraphrasing**.
- Abstractive summarization models **rarely perform semantic aggregation**.
- Human **evaluation metrics** mostly **ignore aggregation**.

More in the poster session...

DebateKG: Automatic Policy Debate Case Creation with Semantic Knowledge Graphs



Allen Roush David Mezzetti

Plailabs NeuML

allen@plailabs.com david.mezzetti@neuml.com

Link to Demo: <https://huggingface.co/spaces/Hellisotherpeople/DebateKG>



- Competitive Debate in the USA is an educational activity held at thousands of high schools and universities.
 - Policy Debate, the type we target, is the most intense format, which keeps a full year long resolution/topic. The affirmative team argues for any plan which topically implements the resolution. The negative team argues for why the affirmative teams plan is a bad idea and should be rejected (or why their plan is “untopical” as in doesn’t topically affirm the resolution). Note the infinite number of potential plans, and the infinite number of potential reasons why they’re bad.
 - Quality of speech act doesn’t matter compared to the quality of evidence/argumentation. Debaters usually “speed read” AKA “spread” their arguments to maximize the number of pieces of evidence presented and strengthen their cases.
- (Roush and Ballaji, 2020) introduced the “**DebateSum**” dataset of ~190K pieces of Policy debate evidence with corresponding biased abstractive summaries, biased word-level extractive summaries, citations, and metadata.
- Our follow-up work introduces **DebateKG**, which leverages Argumentative “Semantic Knowledge Graphs” to construct effective debate cases. We enhance the DebateSum dataset with over 53,000 new examples (updating from 2019 cutoff with 2020-2023 years of evidence), further metadata for every example, and we create 9 semantic knowledge graphs using this data
 - We define a “Semantic Knowledge Graph” as a Knowledge Graph where vertices represent some granularity of text (full document, summary, or even sentences within the document), and edges are drawn between each vertex and its nearest neighbors with semantic similarity higher than a specified cutoff amount. We limit the number of edges to 100
 - We used the “**txtai**” vectorDB from **NeuML** to create a synchronized unified SQL index, Graph index, and vector index over each of the rows in the dataset

DebateKG: Automatic Policy Debate Case Creation with Semantic Knowledge Graphs



Allen Roush David Mezzetti

Plailabs NeuML

allen@plailabs.com david.mezzetti@neuml.com

Link to Demo: <https://huggingface.co/spaces/Hellisotherpeople/DebateKG>



- We find that a constrained shortest path traversal on these kind of Semantic Knowledge Graphs creates high quality debate cases.
 - By “constrained” we mean that we constrain the retrieved vertices to follow certain properties. The most important constraint is to make sure that we are only drawing evidence from the corresponding side that we are on (i.e. we don’t want negative evidence included in an affirmative debate case)
 - Lots of potential for value from other graph algorithms. In this context, Something like pagerank would find the most “generic/central” evidence which would be highly applicable to a diverse range of arguments. Lots of further work to do here
 - No good way to automatically evaluate this. We evaluate on the domain specific heuristic that shorter debate cases (by average number of words) are better as they give the debater further speaking time to add any additional arguments they want at the end. Better evaluation is left for future work.
- Contributed semantic graphs vary based on granularity (either indexing the full document, the abstractive summary, or each sentence within the document) and based on the underlying language model (Legalbert, longformer, mpnet)
 - Much of this research was performed in late 2022, many many better models have come out since then. Performance is expected to significantly improve with SOTA models.
- **Final Takeaway/Call-to-Action:**
 - Just as we regularly contribute pre-trained LLMs, we should also regularly contribute pre-built vector indexes over datasets (making them the extractive analogy to an LLM) and pre-built Semantic Knowledge Graphs. We could find very few examples of open-sourced Semantic Knowledge Graphs. This is problematic since indexing is time-consuming and expensive for those without good GPUs.

Extract, Select and Rewrite: A Modular Sentence Summarization Method

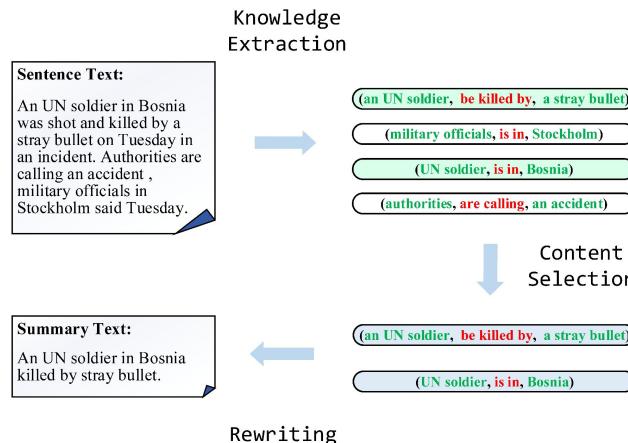
Shuo Guan, Vishakh Padmakumar

Motivations

- Modular approaches have greater controllability and better low-resource performance than end-to-end approaches.
- The existing "modularity" rely on extra encoders which is an implicit modular method and may limit their interpretability.

Structure

We decompose summarization into three stages, with knowledge triples as the granularity. The structure of ESR is shown in the following figure using a sentence example from Gigaword.



Performance

Model	R-1	R-2	R-L
BART (2020)	24.19	8.12	21.31
PEGASUS+Sum (2022)	29.83	9.50	23.47
BART-R3F (2021)	30.31	10.98	24.74
ESR			
$S_R + R_G$	30.63	10.82	24.78
$S_R + R_R$	29.92	10.51	24.26
$S_{RIK} + R_{G1k}$	29.67	10.09	24.00
$S_{RIK} + R_{RIK}$	29.38	10.02	23.90
$S_{RIK} + R_G$	29.09	10.07	23.86

Case Study

ST: Zairean president Mobutu Sese Seko will stay at his French Riviera residence until at least the middle of the week because of an increase in diplomatic activity, a Mobutu aide said on Sunday.

Selected Triples:

(Zairean president Mobutu Sese Seko, will stay at, his French Riviera residence)

(Zairean president Mobutu Sese Seko, will stay until, the middle of the week)

Ref: Zairean president Mobutu to stay in France till mid-week

BART: Tanzania's Mobutu to stay at **Riviera residence** until middle of week

ESR (Gigaword content selector):

- Gigaword rewriter:** **Zairean president Mobutu** will stay at his **French Riviera residence** until the middle of week
- Reddit-TIFU rewriter:** **Zairean Mobutu** will stay at his **French Riviera president residence**... it's said that he will stay until the middle of week

From Sparse to Dense: GPT-4 Summarization with Chain of Density Prompting

Generating Extractive and Abstractive Summaries in Parallel from Scientific Articles Incorporating Citing Statements

Three Types of summaries:

- i) Gold standard: Abstracts and Human-created
- ii) Silver standard: T5-Generated

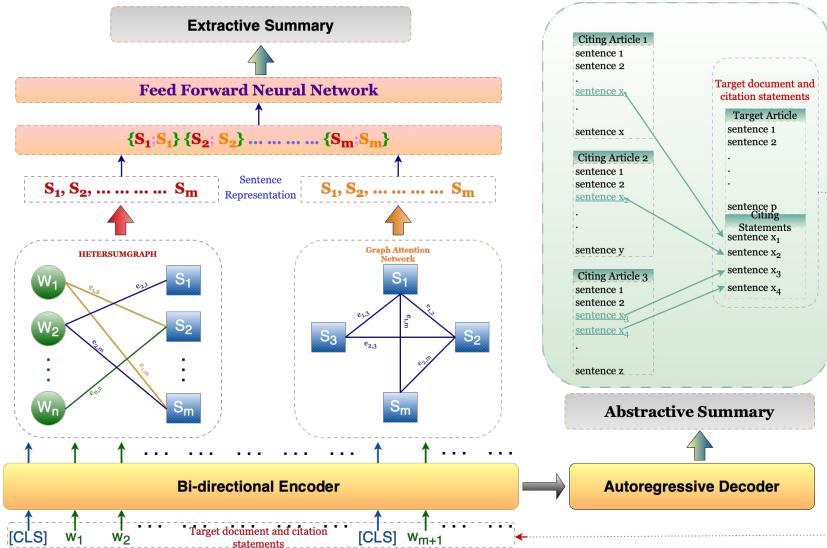


Table 1: Results on the modified SSN corpus. The results consider both the abstracts and the T5-generated summaries incorporating citation statements as the reference summaries. The best results are boldfaced.

Models	On Abstracts as Summaries				On T5-Generated Summaries			
	R-1	R-2	R-L	METEOR	R-1	R-2	R-L	METEOR
Extractive								
BERTSumExt	42.92	14.19	39.01	33.09	43.11	14.21	39.12	33.07
HeterSumGraph	44.27	14.52	39.73	33.18	44.30	14.53	39.74	33.18
GRETEL	45.22	15.19	40.23	36.87	45.23	15.19	40.24	36.88
Proposed Model (Extractive)	45.19	15.18	40.21	36.83	45.19	15.21	40.23	36.85
Abstractive								
PTGen+Cov	41.66	13.08	36.95	32.44	41.60	13.10	36.72	32.40
BERTSumAbs	42.06	14.52	38.17	32.49	42.04	14.56	38.17	32.49
BERT+CopyTransformer	42.43	15.01	39.03	32.88	42.44	15.05	39.04	32.91
Proposed Model (Abstractive)	44.82	15.19	39.31	36.50	44.83	15.19	39.30	36.51

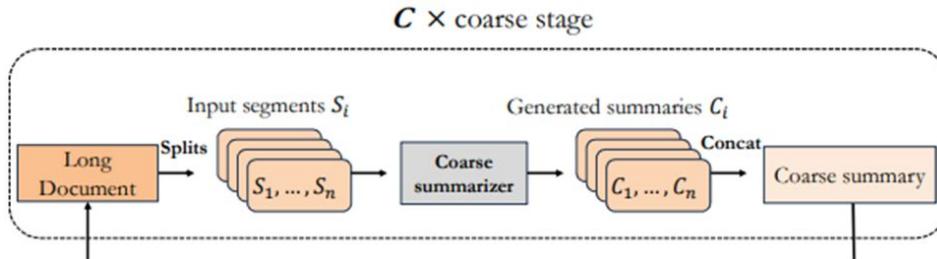
Table 2: Model performance analysis on two CL-SciSumm-2020 summary categories. All values are F-1 scores.

Models	Abstracts as Summaries			Human-created Summaries		
	R-2	R-SU4	METEOR	R-2	R-SU4	METEOR
Jaccard-focused GCN	0.19931	0.09956	-	0.2042	0.14162	-
Clustering	0.1959	0.0962	-	0.1749	0.1169	-
MMR2	0.15067	0.07851	-	0.15073	0.10237	-
LSTM+BabelNet	0.329	0.172	-	0.241	0.171	-
Proposed Model						
Extractive Summarizer	0.43	0.266	31.12	0.42	0.249	30.18
Abstractive Summarizer	0.43	0.250	30.98	0.41	0.234	30.06

Improving Multi-Stage Long Document Summarization with Enhanced Coarse Summarizer

Multi-stage long document summarization

- Compressing long document into concise text through summarization



- Generate the final summary through the concise text



Limitation of previous approaches

- A low performing coarse summarizer adversely affects the final performance

Our Goal

- Enhancing the coarse summarizer for improve final performance

Proposal

- Generating high-quality new data for coarse summarizer
- Proposing a new objective function incorporating contrastive learning

In-context Learning of Large Language Models for Controlled Dialogue Summarization: A Holistic Benchmark and Empirical Analysis

Results & Highlights

- LLMs can generate reasonable summaries via ICL inference but perform differently.
- LLMs can achieve controlled dialogue summarization via ICL.
- Adding control signals like keywords into the prompt guides models to include relevant information.
- LLMs exhibit the bias to omit against numerical information within the dialogue.

Model	Success Rate (%)
OPT-IML-1.3B	19.0 ($\uparrow 4.2$)
LLaMA-7B	10.1 ($\uparrow 4.7$)
Alpaca-7B	7.8 ($\uparrow 3.5$)
BLOOM-7B	28.3 ($\uparrow 17.0$)

Table 9: The success rates of numerical keywords.

Model	Size	ROUGE-1	ROUGE-2	ROUGE-L	Perplexity	Factual Consistency(%)
OPT	1.3B	30.7	6.6	22.6	64.7	60.2
OPT-IML	1.3B	34.6	9.9	27.8	264.4	80.9
mT5-XL	3.7B	21.9	7.4	21.5	139.3	48.4
CEREBRAS-GPT	6.7B	31.5	7.4	22.4	28.0	66.6
LLaMA	7B	31.0	7.3	22.9	41.1	94.0
Alpaca	7B	32.0	7.1	23.7	90.8	97.3
BLOOM	7B	32.1	7.7	23.2	38.2	82.1
GPT3-davinci-003	175B	43.8	17.0	39.4	66.6	-

Table 3: Evaluation results in the uncontrolled setting. The ROUGE F-scores are reported. The optimal performance is highlighted in bold. GPT-3 serves as the factual consistency evaluator, so its factual consistency is excluded.

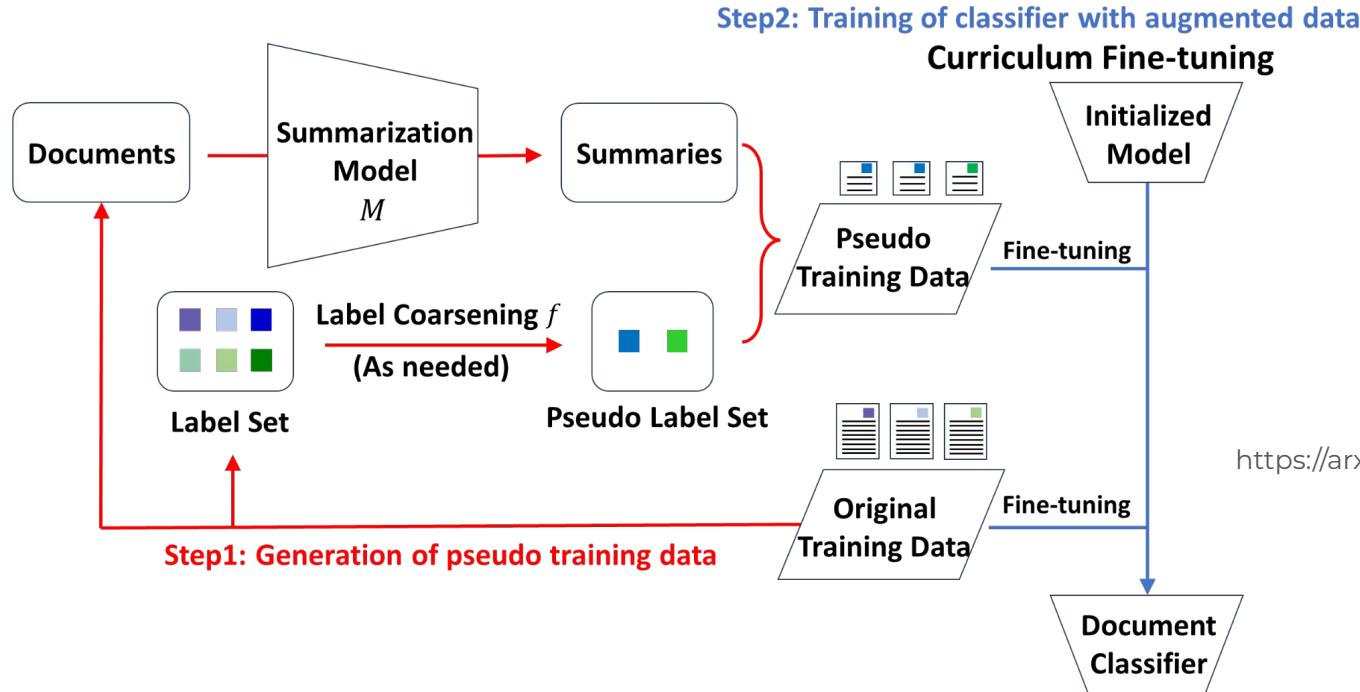
Gold Summary	Model	Uncontrolled setting	Entity control
The audition starts at 7.30 P.M. in Antena 3.	BLOOM-7b1	Jamilla and Yoyo are going to an audition for a TV show. They are waiting for the station to start broadcasting.	Jamilla and Kiki are going to an <u>audition</u> for a TV show. The station is <u>Antena3</u> .
Adam and Karen are worried that May suffers from depression. Karen will call her friend who is a psychologist and ask for advice.	mT5-XL	Melissa shows Brian what she bought and he finds it awesome.	Adam and <u>karen</u> talk to their friend about <u>her depression</u> . They talk to their friend's <u>psychologist</u> and she agrees to help.
Mike suspects he might have had an allergic reaction to something.	LLaMA-7B	Mike wakes up with a strange rash on his face. He has no clue what it is and asks his girlfriend Emma to help him figure it out.	Mike <u>suspects</u> that he is allergic to something, but he does not know what it is.
Flo cannot get an appointment at the salon until the 6th. Flo worries she's going to be gray. Flo will have to get a touch-up kit at Tesco.	CEREBRAS-6.7B	Gina and Flo are having a conversation about getting their hair done at the same time.	Flo is <u>worried</u> that she will be too late for her appointment at the hair salon. She will get a <u>kit</u> to get her hair touched up.

Table 8: Comparisons of generated summaries between controlled and uncontrolled settings. Underlined keywords are the control signals included in the prompts.

Summarization-based Data Augmentation for Document Classification

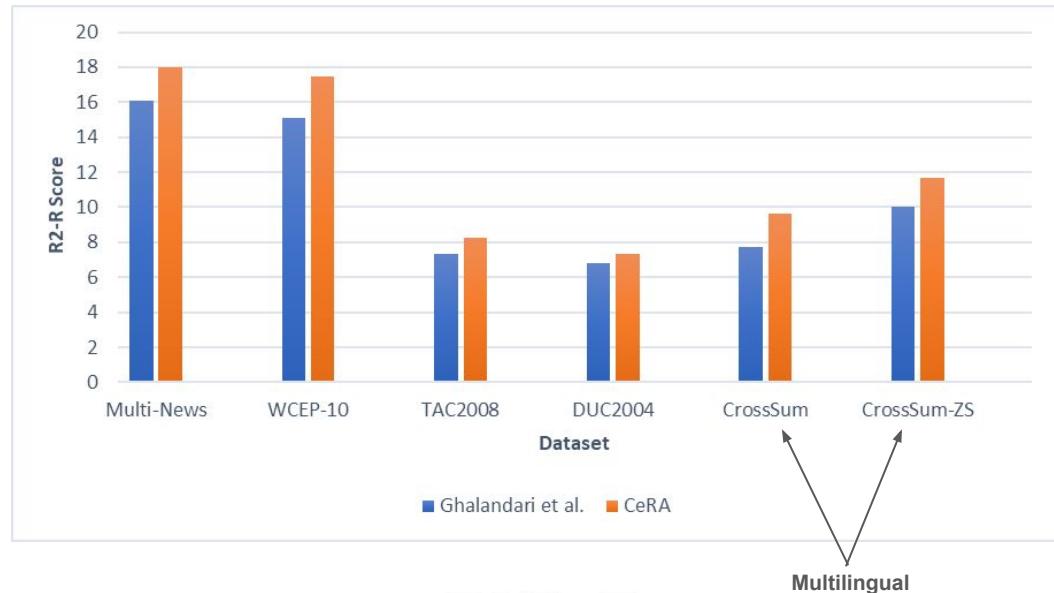
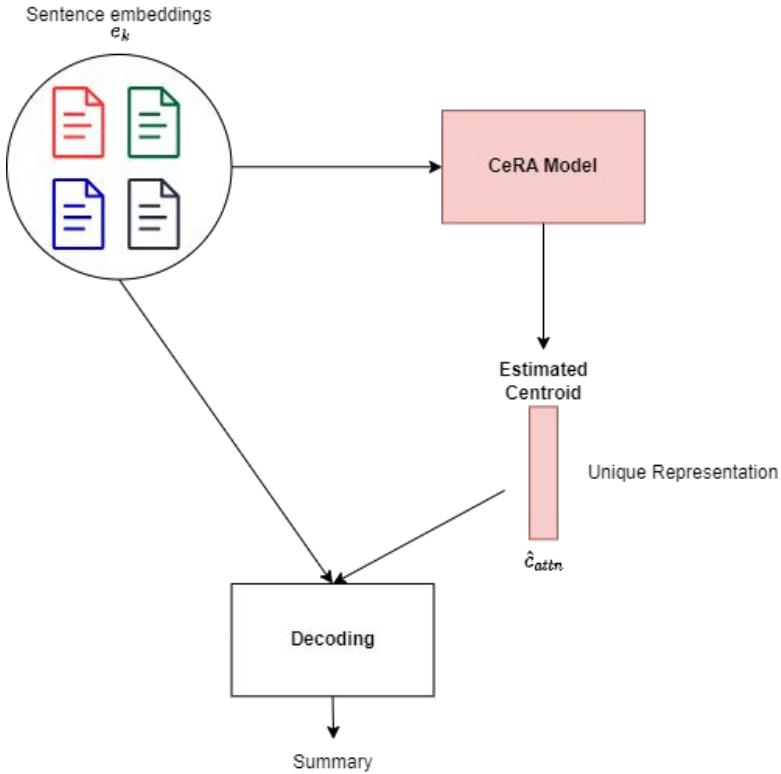
Motivation: Can LMs become to understand lengthy text by learning short text first, as humans do?

Proposal: We leverage summarization to apply data augmentation for document classification



We hope that automatic summarization can help more NLP tasks in the future!

Supervising the Centroid Baseline for Extractive Multi-Document Summarization

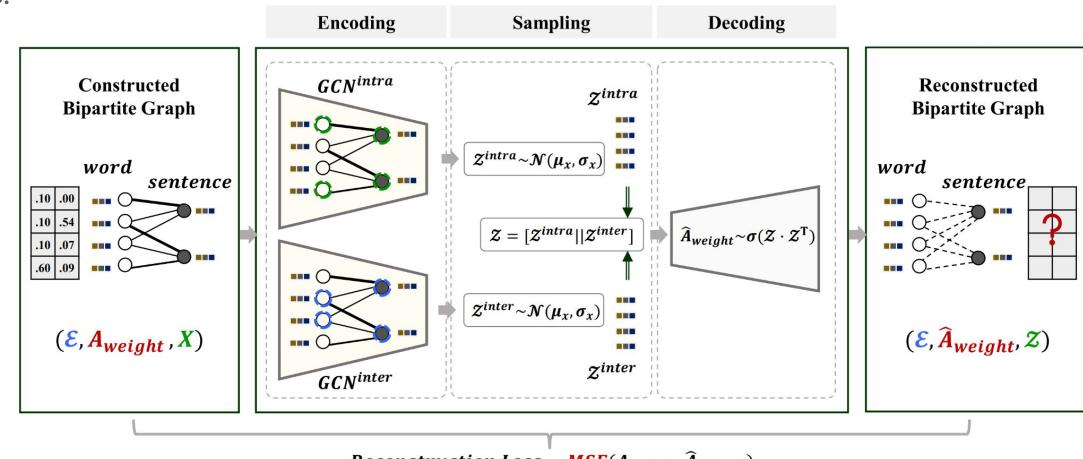
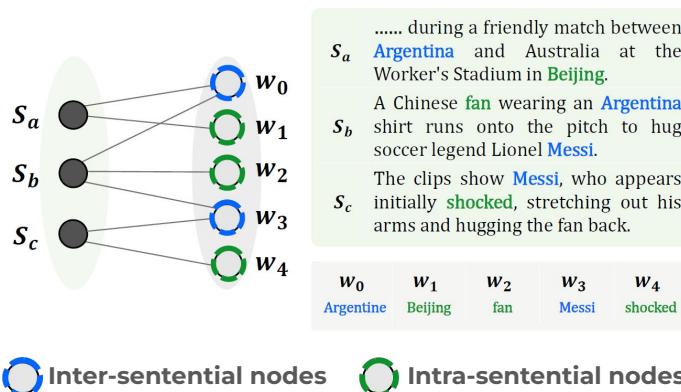


Multilingual
→
↑

Bipartite Graph Pre-training for Unsupervised Extractive Summarization with Graph Convolutional Auto-Encoders

Motivation: We contend that **pre-training informative and distinctive sentence representations**, aids in ranking important sentences in downstream summarization.

Proposal: We propose a novel **graph pre-training autoencoder** to obtain sentence embeddings by explicitly **modelling intra-sentential distinctive features and inter-sentential cohesive features** through sentence-word bipartite graphs.



<https://github.com/OpenSUM/BiGAE>

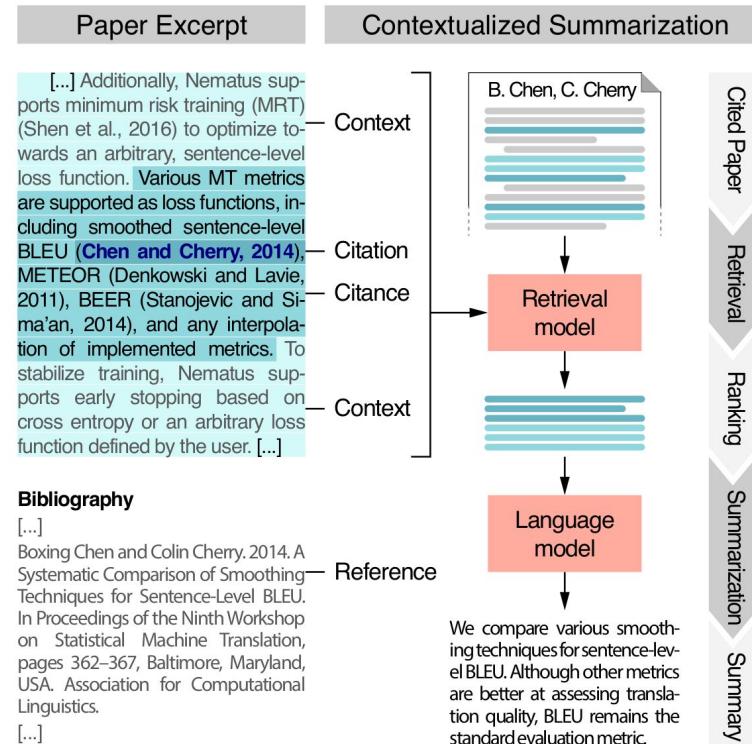
arXiv <https://arxiv.org/abs/2310.18992>



Can you Summarize my learnings? Towards
Perspective-based Educational Dialogue
Summarization

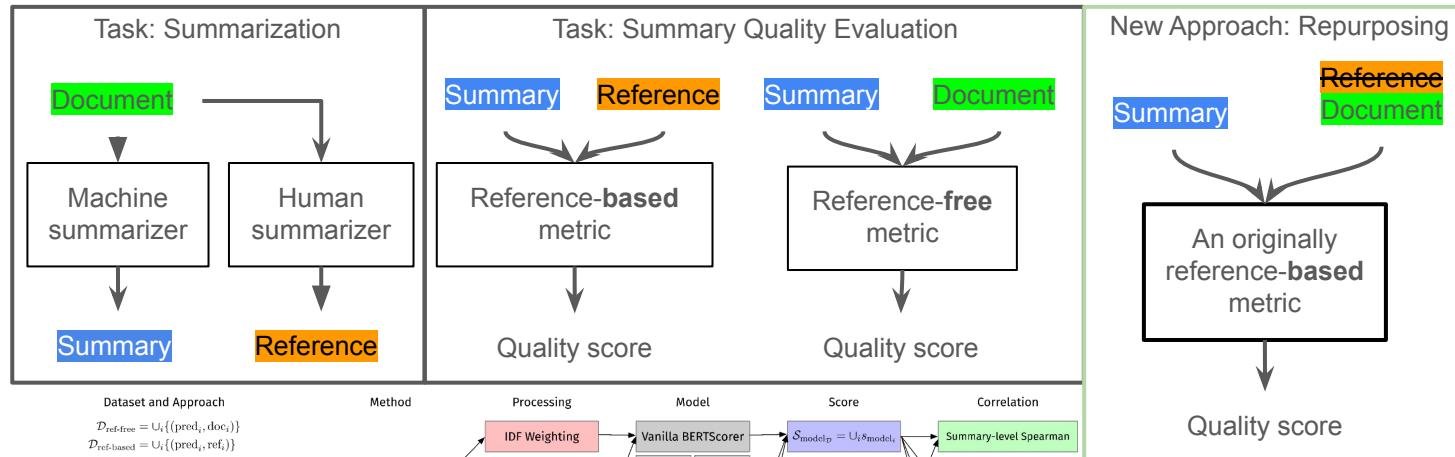
Citance-Contextualized Summarization of Scientific Papers

- Abstracts are conventionally used as summaries of scientific papers.
- They are insufficient/incomplete for providing sufficient context for understanding citations.
- We propose leveraging citation texts for retrieving relevant information from a cited paper for summarization.
- We present **Webis-Context-Scisumm**, a new corpus with 540K papers and 4.6M citation contexts for contextualized summarization.



DocAsRef: An Empirical Study on Repurposing Reference-based Summary Quality Metrics as Reference-free Metrics

“Reference-based BERTScore performs better as a reference-free metric.”



[https://github.com/
SigmaWe/DocAsRef](https://github.com/SigmaWe/DocAsRef)

arXiv [https://arxiv.org/
abs/2212.10013](https://arxiv.org/abs/2212.10013)

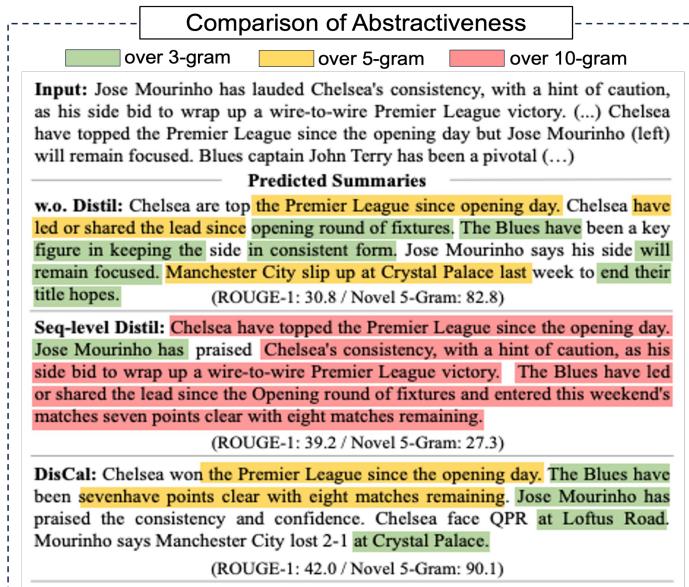


Enhancing abstractiveness of summarization models through calibrated distillation

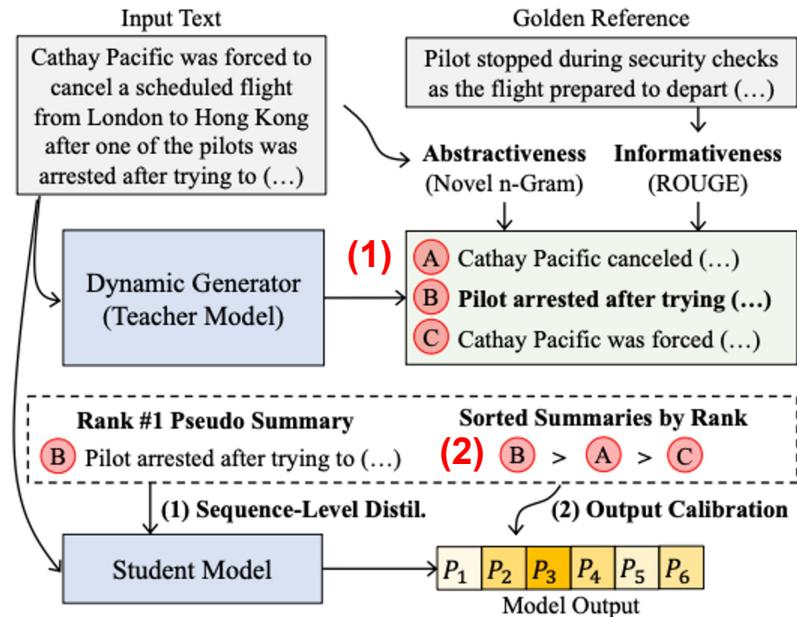


Motivation: existing distillation methods make abstractive summarization to be extractive-like summarization

Our Goal: improving both informativeness (Rouge scores) and abstractiveness (novel n-gram scores) together

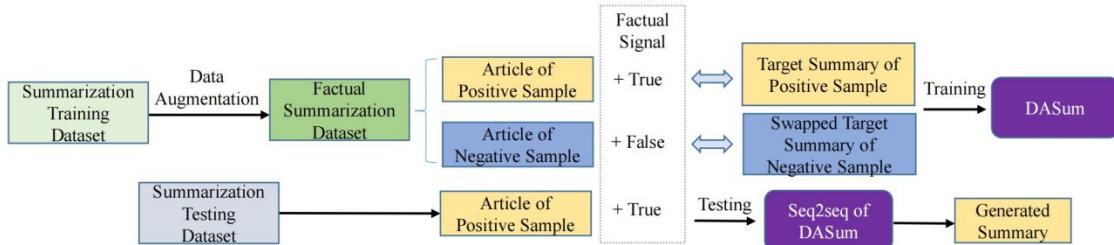


Proposed Calibrated Distillation: (1) exposing diverse pseudo summaries and utilizing (2) rank information derived from them w.r.t informativeness and abstractiveness



Factual Relation Discrimination for Factuality-oriented Abstractive Summarization

Our motivation is to enhance the model's attention to the factual nature of summarization by constructing both factual (positive samples) and counterfactual summaries (negative samples).



Counterfactual Summarization Construction

(1) Pronoun swapping

E1: **Original summary sentence:** Her sister, shaneah, was dating lloyd.

Modified summary sentence: His sister, shaneah, was dating lloyd.

(2) Sentence negation

E2: **Original summary sentence:** Peter may have enough to spread around.

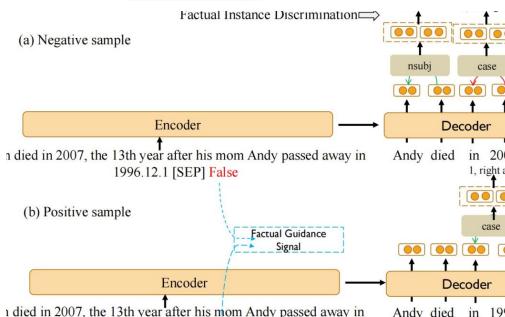
Modified summary sentence: Peter may not have enough to spread around.

(3) Time and date entity, quantifier, and named entity swapping

E3: **Original summary sentence:** Andy died in 1996.12.1.

Set of DATA type entities in the source document: {1996.12.1, 2007, this week, Friday}

Modified summary sentence: Andy died in 2007.



For positive samples, as shown in (b), the correct factual signal (True) is added to the end of the source document.

For negative samples, as shown in (a), the counterfactual signal (False) is added, and there are both factual and counterfactual relations in the summary.

FREDSum: A Dialogue Summarization Corpus for French Political Debates

	Dataset	Lang.	#Transcripts	#Words/trans.	#Turns/trans.	#Speakers/trans.	#Words/sum.
dialog	SAMSum	EN	16369	93.8	11.2	2.4	20.3
	DialogSum	EN	13460	131.0	-	-	23.6
	MediaSum	EN	463596	1553.7	30.0	6.5	14.4
meeting	AMI	EN	137	6007.7	535.6	4.0	296.6
	ICSI	EN	59	13317.3	819.0	6.3	488.5
	MeetingBank	EN	6892	3800.3	146.9	3.2	87.2
	VCSum	CN	239	14106.9	73.1	5.6	231.9
	ELITR	EN/CS	179	7549.9	884.5	6.5	327.9
debate	FREDSum	FR	142	2595.5	54.2	4.2	238.9
	FREDSum _{preS}	FR	740	71386.8	685.0	53.4	-
	FREDSum _{preA}	FR	4563	28619.8	445.8	56.0	-

Table 1: Comparison between FREDSUM and other news, dialogue or meeting summarization datasets. # stands for the average result. FREDSum_{preS} and FREDSum_{preA} represent the pretraining datasets released along FREDSum.

- Motivations :
 - Address the scarcity of summarization data
 - Utilize the rich source of political debates to offer unique challenges and opportunities
 - Drive innovation in the field of discourse analysis
- Contributions :
 - Introducing FREDSum, the first large scale French multi-party summarization resource
 - Multi-level abstractive summaries
 - Extensive experiments and human evaluation, a taxonomy of common hallucination

Improving the Robustness of Summarization Models by Detecting and Removing Input Noise

Google Research

Kundan Krishna, Yao Zhao, Jie Ren, Balaji Lakshminarayanan, Jiaming Luo, Mohammad Saleh, Peter J. Liu

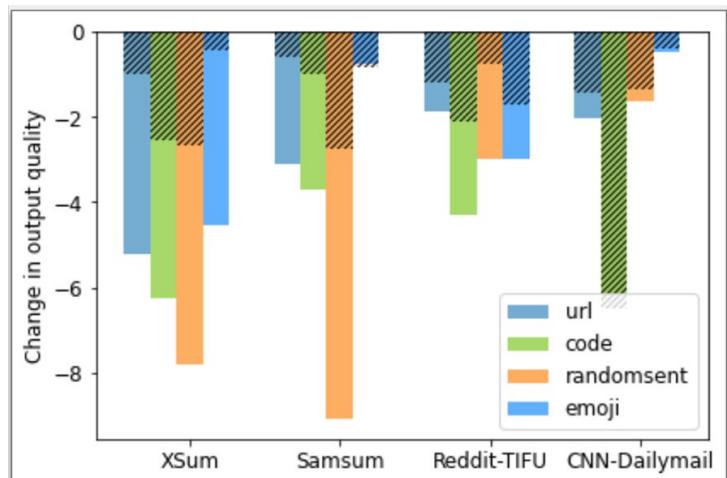
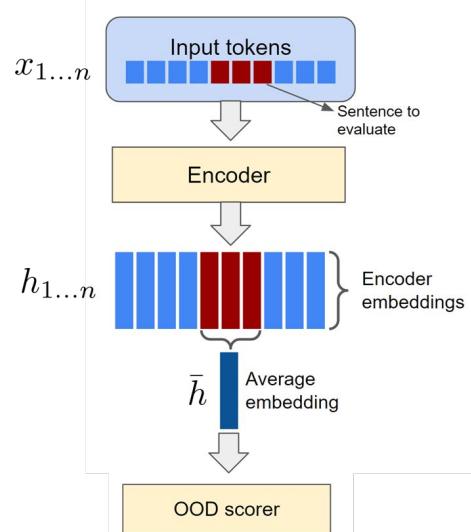
Carnegie Mellon University

Problem: Data extracted from web pages might contain noise of unknown types (e.g. ads, code ...) which can hurt summary quality.

Solution: Compute input embeddings using the summarization model and use a Mahalanobis distance based OOD detector to identify noisy spans.

Result: We show that addition of different kinds of noise can lead to large drops in output quality, and our proposed approach to filter out noise can recover a large percentage of that performance drop across datasets and noise types (shaded region below)

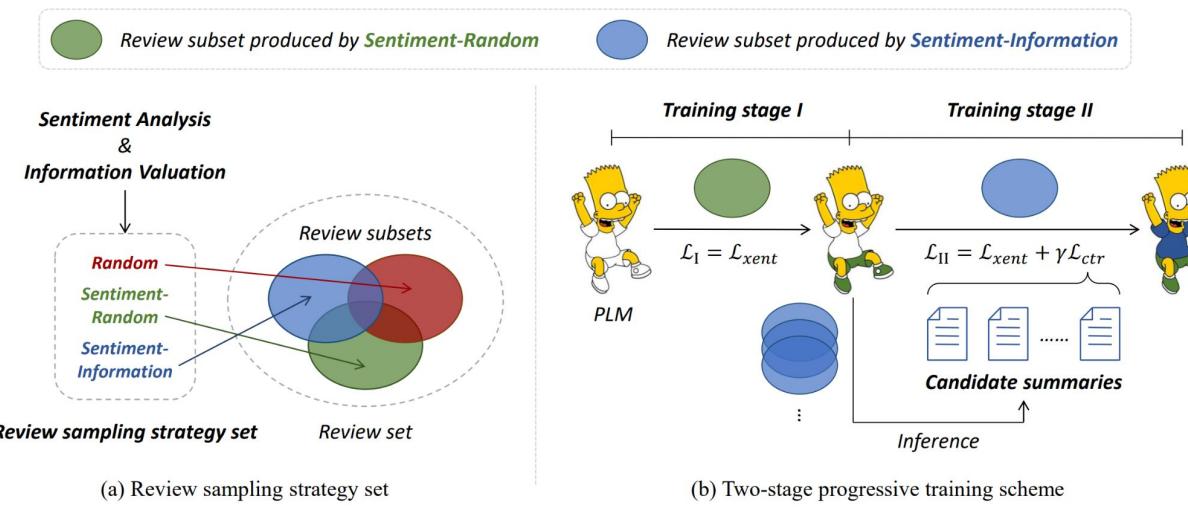
Graham participated in the Government of Canada's Defence Review, as one of four members of a Minister's Advisory Panel, providing input for Defence Minister Harjit Sajjan.
The review aimed to consult with Canadians across the country in order to develop a future road map for Canada's defence policy. In June 2017, it was released as "Strong, Secure, Engaged."
In 2016 Graham published an autobiography, "Call of the World: A Political Memoir", reprinted in paperback in 2018.
External links.
! colspan="3" | Cabinet post
! colspan="3" | Cabinet posts (2)



Large-Scale and Multi-Perspective Opinion Summarization with Diverse Review Subsets

Motivation

- (1) process larger sets of reviews (from 8-10 reviews to hundreds of reviews)
- (2) provide summaries from different perspectives
- (3) work with limited computational resources* (GPU, context length, etc.)



Proposal (SubSumm)

- supervised summarization framework for large-scale multi-perspective opinion summarization
- review sampling strategy set regarding sentiment orientation and information value
- two-stage training scheme where contrastive learning with candidate summaries is extra performed

Medical Text Simplification: Optimizing for Readability with Unlikelihood Training and Reranked Beam Search Decoding

We want to simplify medical texts.
However, existing work tends to copy, rather than simplify.

Source

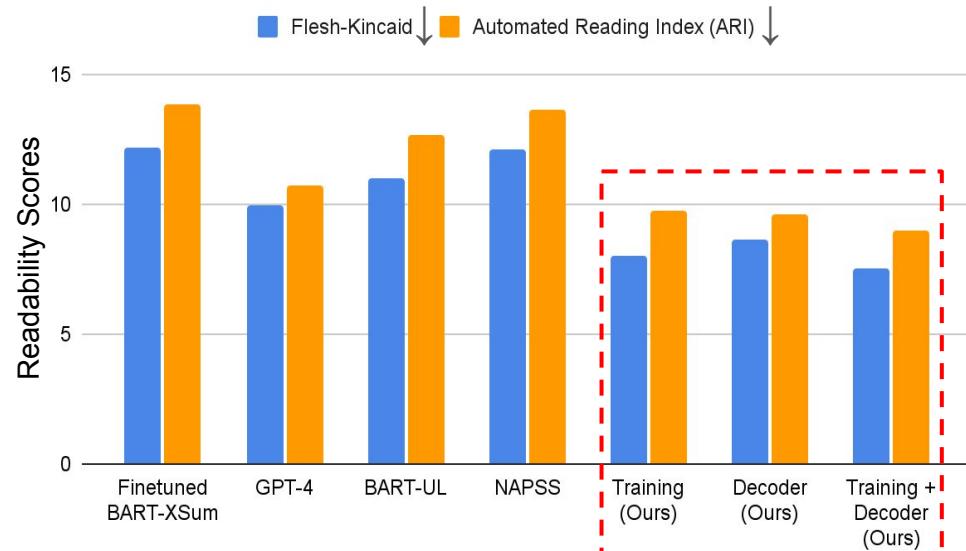
A total of 38 studies involving 7843 children were included... **Very few data were available for other outcomes ... no statistically significant difference between education and control.** Asthma education aimed at children and their carers **who present to the emergency department for acute exacerbations can result in lower risk of future emergency department presentation and hospital admission.**

Simplified Version (NAPSS, Lu et al., 2023)

A total of 38 studies involving 7843 children were included...
Asthma education aimed at children and their carers **who present to the emergency department for acute exacerbations can result in lower risk of future emergency department presentation and hospital admission... Very few data were available for other outcomes ... no statistically significant difference between education and control.**

We propose training and decoding strategies to improve readability, which produce simpler text on two datasets.

Cochrane



Open Domain Multi-document Summarization: A Comprehensive Study of Model Brittleness under Retrieval

John Giorgi, Luca Soldaini, Bo Wang, Gary Bader, Kyle Lo, Lucy Lu Wang, Arman Cohan

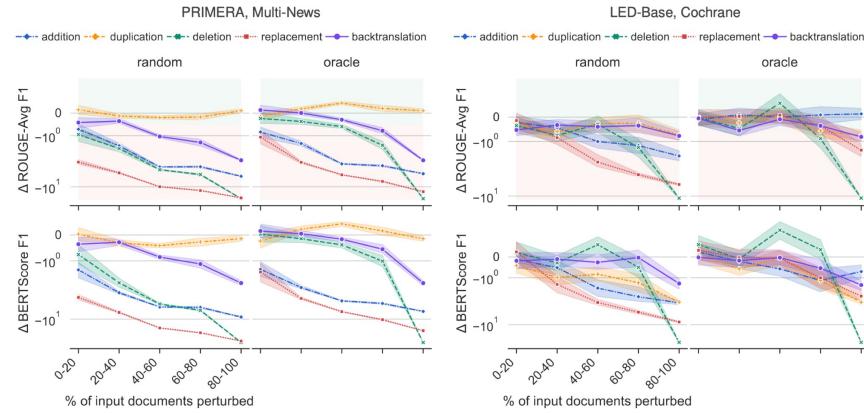
💡 Open domain MDS: documents must be first retrieved given a user information need

🧐 Input documents for most existing MDS datasets are often given (assumes no retrieval)

🧠 We bootstrap this new task using reference summaries as queries

📊 We evaluate how existing SoTA summarizers and retrievers behave

📈 We investigate impact of different retrieval errors



Dataset	Model	P@K	R@K	ROUGE-Avg F1	Δ ROUGE-Avg F1
Multi-News	PRIMERA	0.22	0.82	31.66	-7.39
	PEGASUS	—	—	31.23	-8.49
	LSG-BART-base	—	—	30.05	-6.44
	GPT-3.5-turbo	—	—	23.86	-2.46
WCEP-10	PRIMERA	0.63	0.67	35.50	-1.02
	LSG-BART-base	—	—	35.76	-1.15
	GPT-3.5-turbo	—	—	26.36	-0.22
Multi-XScience	PRIMERA	0.06	0.40	18.31	-0.57
	MS ²	0.16	0.22	19.66	-0.14
	Cochrane	0.17	0.57	17.39	-0.28

PMIndiaSum: Multilingual and Cross-lingual Headline Summarization for Languages in India

What did we do?

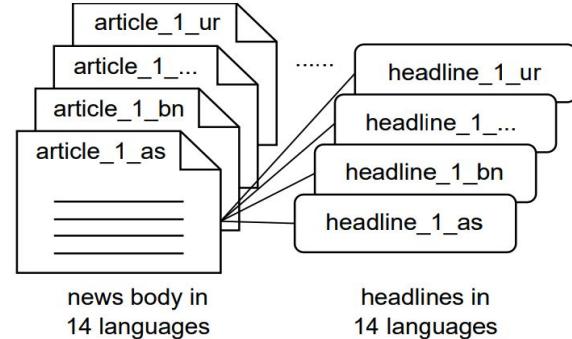
- crawl from the Prime Minister of India website
- 14 languages, 4 families:
 - as, bn, gu, hi, kn, ml, mni, mr, or, pa, ta, te, ur, en
- 196 language directions:
 - monolingual, cross-lingual, multilingual

But is it good?

- from a govt website
- rule-based cleaning
- human evaluation
- LaBSE cross-lingual scores

What did the GPUs do?

- extractions
- translate+summarize,
- fine-tuning mBART and IndicBART:
 - monolingual, cross-lingual (ok-ish)
 - multilingual (needs work)
- prompting LLMs
- and more ...



Re-Examining Summarization Evaluation across Multiple Quality Criteria

Ori Ernst, Ori Shapira, Ido Dagan, and Ran Levy

Ranking summarization evaluation metrics by correlation to human ratings doesn't make sense!

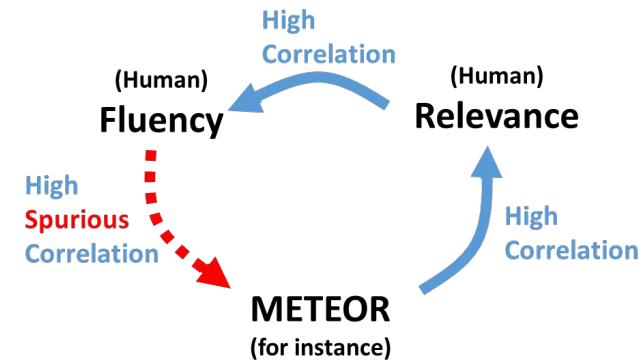
Some of the correlations are spurious!

**Want to know the reason?
How can we deal with it?**

Come to our poster!

Metric correlation to human annotations (SummEval)

Metric	Coherence	Consistency	Fluency	Relevance
ROUGE-3	0.2206	0.7059	0.5092	0.3529
CHRF	0.3971	0.5294	0.4649	0.5882
METEOR	0.2353	0.6324	0.6126	0.4265
ROUGE-1	0.2500	0.5294	0.5240	0.4118



Responsible AI Considerations in Text Summarization Research: A Review of Current Practices

Yu Lu Liu, Meng Cao, Su Lin Blodgett, Jackie Chi Kit Cheung, Alexandra Olteanu, Adam Trischler

We investigate how RAI issues are covered in the contemporary text summarization literature:

We conduct a **systematic review** of >300 summarization papers

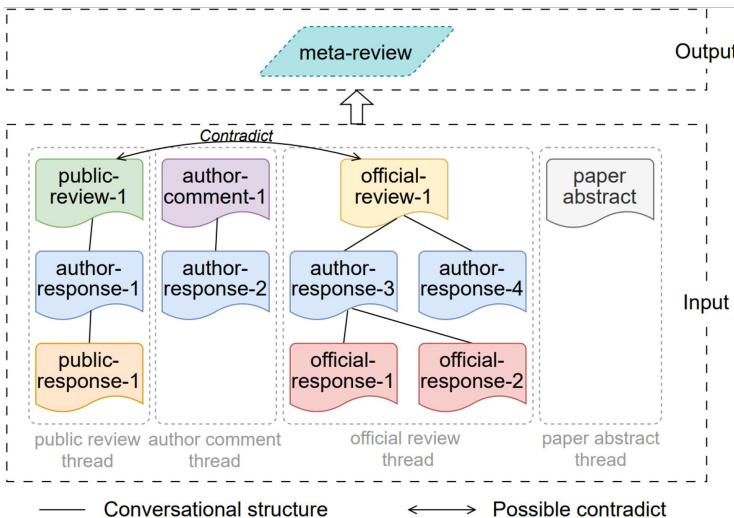
Relatively few papers engage with possible stakeholders or **contexts of use**, which limits their consideration of RAI issues.

Most papers do not discuss the **limitations** of their own work, and rarely include any **ethical reflections**.

... and more findings at the poster session!

Summarizing Multiple Documents with Conversational Structure for Meta-Review Generation

Miao Li, Eduard Hovy, and Jey Han Lau



Model(#Params)	Test Data	R-L↑	BERTS↑	UniEval-Con↑	UniEval-Rel↑	ACC↑
BART (406M)	CF	26.84	14.89	71.85	78.74	0.683
PEGASUS (568M)	CF	26.77	13.66	73.12	79.49	0.649
PRIMERA (447M)	CF	29.13	12.33	66.85	81.70	0.639
LED (459M)	CF	29.19	15.32	70.04	82.82	0.698
PegasusX (568M)	CF	29.30	15.69	71.33	81.30	0.707
RAMMER (459M)	CF	29.19	15.88*	73.21*	83.15*	0.724

- RAMMER (our model) based on sparse attention performs better than other baseline models
- However, all fail to recognise and resolve these conflicts in its meta-reviews

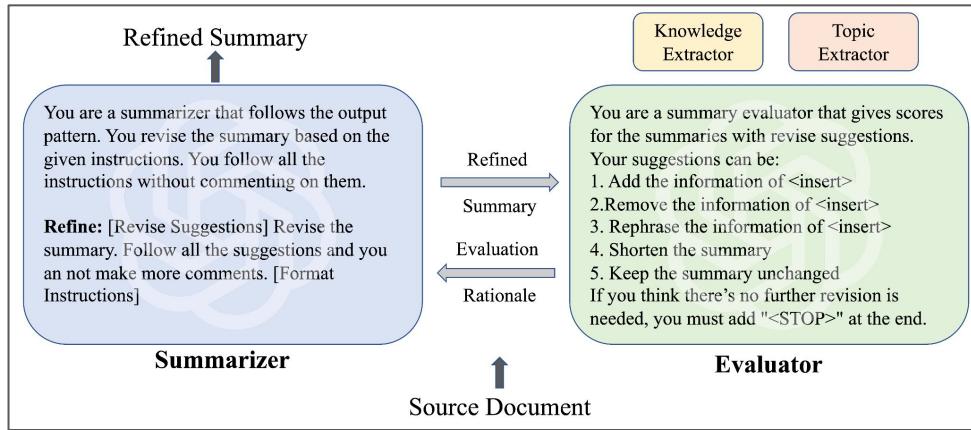
- Meta-review generation is formulated to abstractive multi-document summarization (11,995/1,499/1,499)
 - Explicit conversational structure
 - Source documents feature conflicts (13.6%)
 - A rich set of metadata, such as review ratings

Model	Recognition	Resolution
PRIMERA	3/23	2/23
LED	4/23	4/23
PegasusX	5/23	5/23
RAMMER	8/23	3/23

Table 9: Performances of summarization models on conflict recognition and resolution for CF samples.

SummIt: Iterative Text Summarization via ChatGPT

User	Summarize the Document: Chelsea's on loan midfielder Oriol Romeu goes up against Sportsmail's Martin Keown as the pair predict the scores for the weekend's Premier League and FA Cup matches.....
ChatGPT Summarizer	Summary: Chelsea's on loan midfielder Oriol Romeu and Sportsmail's Martin Keown predict the scores for the weekend's Premier League and FA Cup matches, including the standout fixture between Chelsea and Manchester United.....
ChatGPT Evaluator	Revision Comments: The summary is too long and contains unnecessary details. It should be shortened to only include the main point, which is Oriol Romeu and Martin Keown predicting the scores for the weekend's Premier League and FA Cup matches. Evaluation Score: [0.1, 0.6, 0.2, 0.1, 0]
ChatGPT Evaluator	Summary: Chelsea's on loan midfielder Oriol Romeu and Sportsmail's Martin Keown predict the scores for the weekend's Premier League and FA Cup matches.
ChatGPT Evaluator	Revision Comments: The summary is concise and includes the main point of the document. There's no need to further revise.<STOP> Evaluation Score: [0.9, 0.1, 0, 0, 0]
ChatGPT Evaluator	Chelsea's on loan midfielder Oriol Romeu and Sportsmail's Martin Keown predict the scores for the weekend's Premier League and FA Cup matches.



Model	Coherence	Fluency	Relevance	Consistency	Conciseness	Overall	Human Pref
CNN/DM							
BART	3.92	4.16	4.00	3.12	3.64	3.24	0.04
T5	3.72	4.24	4.32	3.52	3.84	3.68	0.10
PEGASUS	3.20	3.53	3.33	2.87	1.85	1.63	0.00
ChatGPT	4.20	4.36	4.28	4.01	3.92	4.01	0.34
SummIt	4.24	4.50	4.29	4.12	3.84	4.09	0.52
XSum							
BART	3.97	4.30	4.13	3.30	3.93	3.84	0.30
T5	3.84	4.32	4.02	3.63	3.84	3.25	0.08
PEGASUS	3.13	4.10	3.52	2.87	2.03	2.41	0.00
ChatGPT	4.03	4.40	4.30	3.93	3.87	3.92	0.24
SummIt	4.04	4.35	4.28	4.05	3.72	3.96	0.38

Synthesize, if you do not have: Effective Synthetic
Dataset Creation Strategies for Self-Supervised
Opinion Summarization in E-commerce

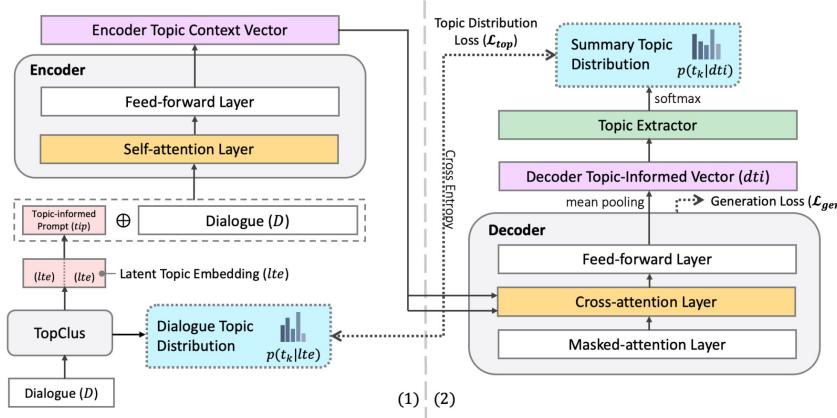
Topic-Informed Dialogue Summarization using Topic Distribution and Prompt-based Modeling

Motivation

A dialogue can easily shift topics due to the intentions of multiple speakers, so various topics can emerge within the dialogue.

Method

We propose **Topic-Informed Dialogue Summarizer (TIDSum)** that generates a comprehensive summary by considering the dialogue topic distribution, and the topic information from topic-informed prompt.



Dialogue

...*(brief)* Tina: How about pasta for dinner?

Steve: Sounds great!

Tina: With broccoli, ham, cheese and cream?

Steve: Scrumptious.

Tina: Your favourite.

Steve: Indeed.

1

Tina: But there is a snag.

Steve: Too perfect to be true?

Tina: It's not about that. We'd need to do some shopping after work. Can you handle it yourself?

Steve: Can we handle it together? You know how scatterbrained I am when it comes to shopping lists.

Tina: I do know!

Steve: Together?

Tina: Fine. Will you be leaving work on time?

Steve: Guess so. I don't expect any problems.

Tina: Ok. Let's meet in the car park, shall we?

Steve: Sure. ...*(brief)*

2

BART generation

Tina and Steve will meet in the car park to do some shopping after work.

3

TIDSum generation

Tina and Steve will have pasta for dinner. They will meet in the car park and do the shopping after work together.

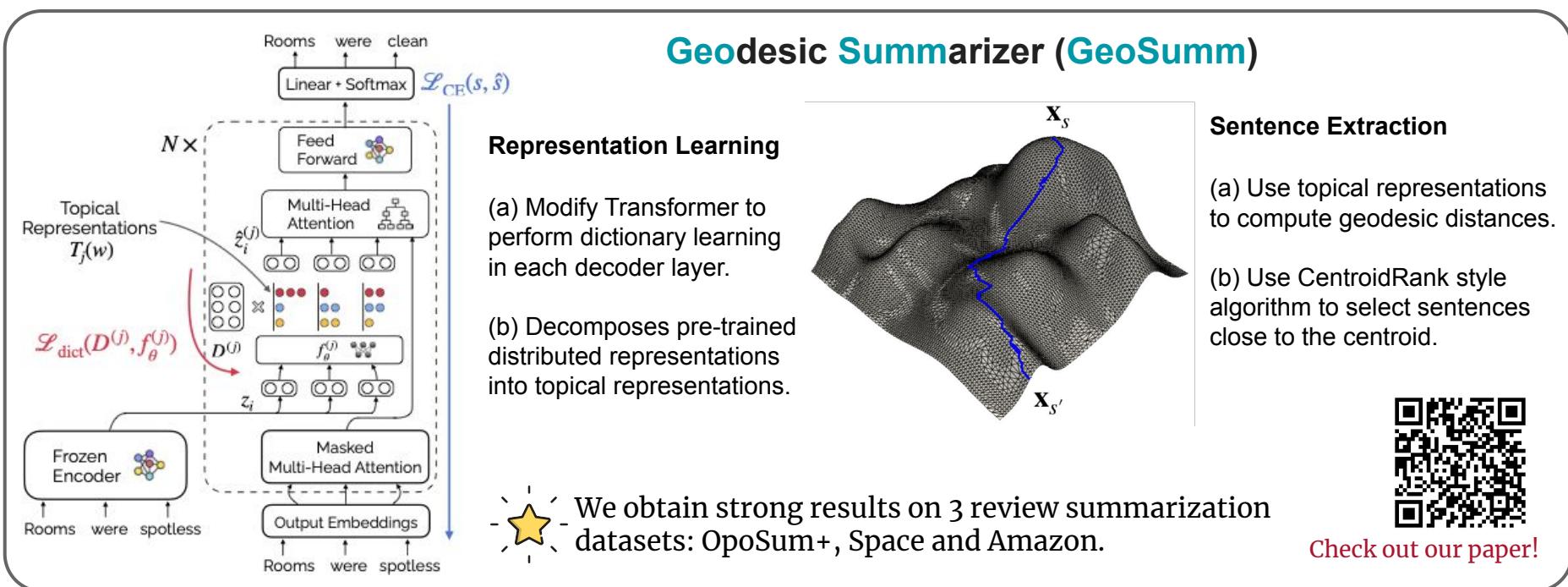
Unsupervised Opinion Summarization Using Approximate Geodesics

Somnath Basu Roy Chowdhury, Nicholas Monath, Avinava Dubey, Amr Ahmed, and Snigdha Chaturvedi



Motivation: We aim to identify the good text representations and algorithms for extractive unsupervised opinion summarization.

Proposal: We show that topical text representations (topic model based) can be a good way to represent text and present a approximate geodesic-based CentroidRank algorithm to perform sentence extraction.



Using LLM for Improving Key Event Discovery: Temporal-Guided News Stream Clustering with Event Summaries

Nishanth Nakshatri, Siyi Liu, Sihao Chen, Daniel Hopkins, Dan Roth, Dan Goldwasser

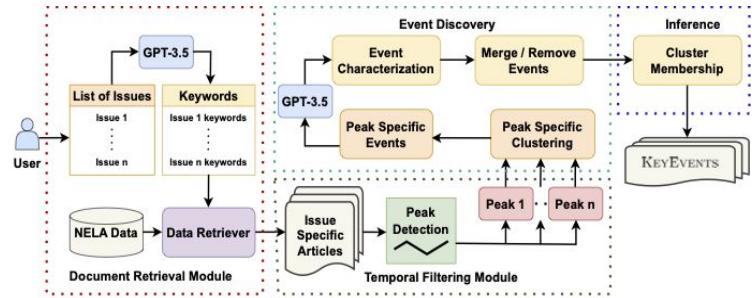


Task: We propose a principled approach to identify real-world news events, without any human intervention.

Birds Eye View: The process is related to past work on *interactive clustering*. We use LLM inference instead of human feedback for cluster refinement.

Our Method:

- Retrieve event candidates using non-parametric methods (such as HDBSCAN).
- Use LLM to characterize event candidates, and reason about their validity.
 - **Event Characterization** is viewed as a multi-document summarization
 - **Merge/Remove Events** to refine the events (viewed as an entailment problem)



- ★ We obtain highly coherent event clusters compared to competitive baselines
- ★ We release the resulting event dataset on 11 contemporary issues

Mitigating Framing Bias with Polarity Minimization Loss

Mitigating Framing Bias with Polarity Minimization Loss

Neutral multi-news Summarization (NeuS) with Polarity Minimization Loss

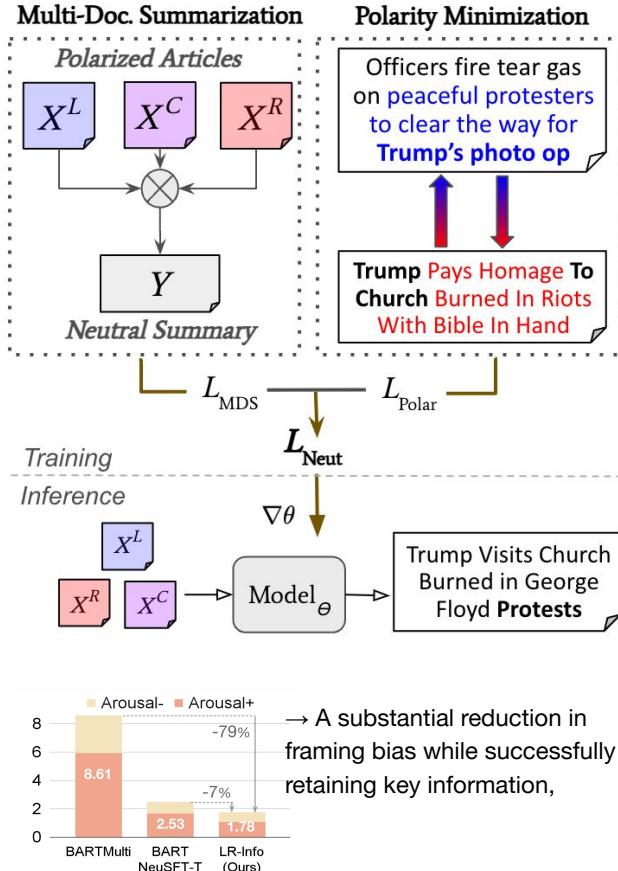
Motivation: Conventional multi-document summarization model does not address the critical issue of **framing bias** in news article summarization, a topic that remains largely under-explored.

Task: NeuS (Lee et al., NAACL 2022) focuses on **neutrality** of the summary out of news articles with varying degrees and orientations of political bias.

Proposal:

- Propose a polarity minimization loss that teaches the model to minimize the polarity difference between **polarized input articles**.
- L_{polar} is designed to jointly optimize the model to map arbitrary polarity ends bidirectionally (e.g., left \rightarrow right; right \rightarrow left)

[Ref] **NeuS**: Neutral Multi-News Summarization for Mitigating Framing Bias, N. Lee, Y. Bang, T Yu, A. Madotto, P. Fung, NAACL 2022



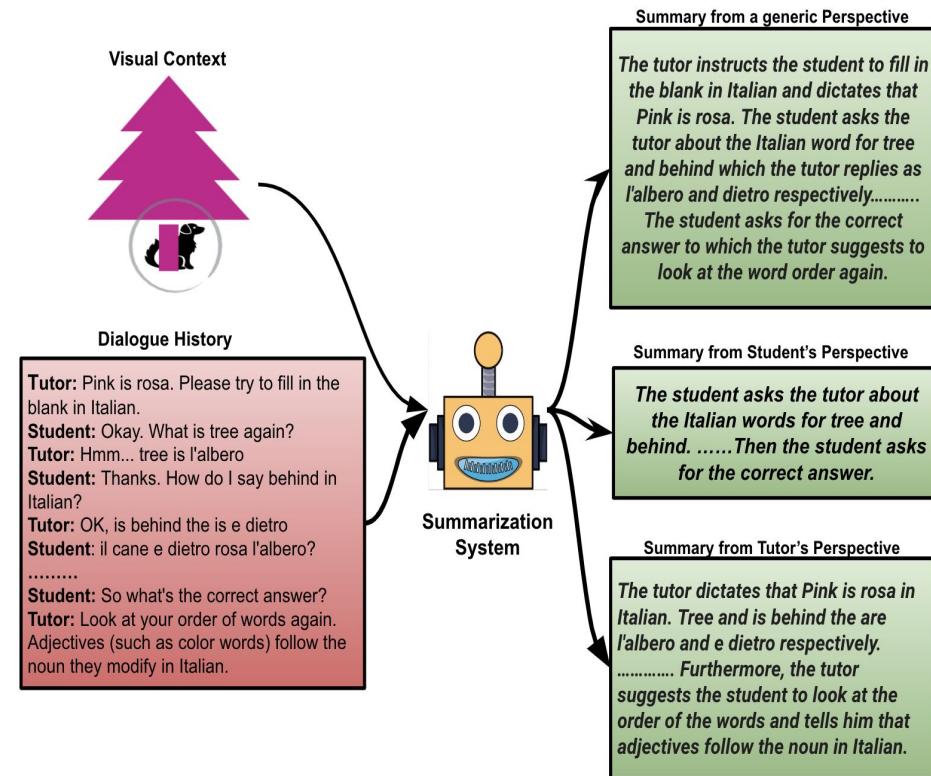
Can you Summarize my learnings? Towards Perspective-based Educational Dialogue Summarization

Raghav Jain, Tulika Saha, Jhagrut Lalwani, Sriparna Saha

Motivation

The MM-PerSumm task, or Multi-modal Perspective based Dialogue Summarization, is a new task proposed in this study in the field of educational dialogue analysis.

It focuses on summarizing educational dialogues from three unique perspectives: the student, the tutor, and a generic viewpoint.



Can you Summarize my learnings? Towards Perspective-based Educational Dialogue Summarization

Raghav Jain, Tulika Saha, Jhagrut Lalwani, Sriparna Saha

