ReflectSumm: A Benchmark for Course Reflection Summarization

Yang Zhong*, Mohamed Elaraby*, Diane Litman
Ahmed Ashraf Butt, Muhsin Menekse
*Equal Contribution

University of Pittsburgh

Carnegie Mellon University

PURDUE UNIVERSITY®

1

Text Summarization

ML models performed well on standard
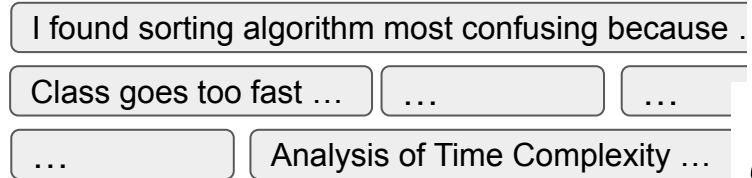   benchmarks with ample data

Text Summarization

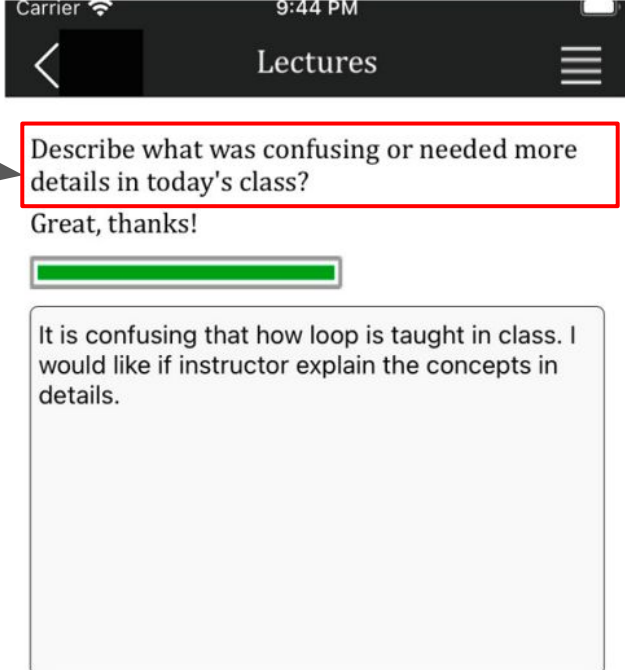ML models performed well on standard benchmarks with ample data

Yet, little is known about how models work on low-resource real-life applications

CourseMIRROR

I found sorting algorithm most confusing because .

Class goes too fast … | … | …

… | Analysis of Time Complexity …

Summary of Reflections

CourseMIRROR Data Collection

- **Step 1:** Students are prompted to write a reflection about what they found confusing or interesting

Carrier 9:44 PM

< Lectures ☰

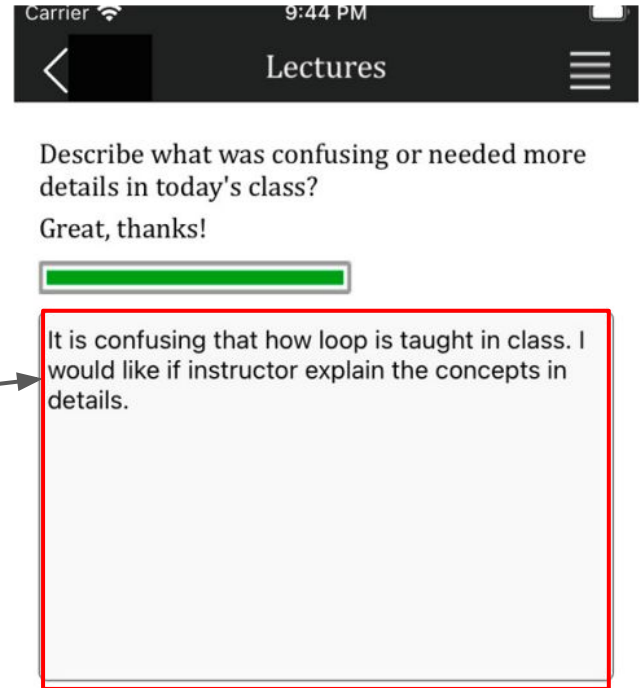Describe what was confusing or needed more details in today's class?

Great, thanks!

It is confusing that how loop is taught in class. I would like if instructor explain the concepts in details.

CourseMIRROR Data Collection

- Step 1: Students are prompted to write a reflection about what they found confusing or interesting
- **Step 2**: Reflections are collected through the CourseMIRROR app

Carrier 9:44 PM

Lectures

Describe what was confusing or needed more details in today's class?

Great, thanks!
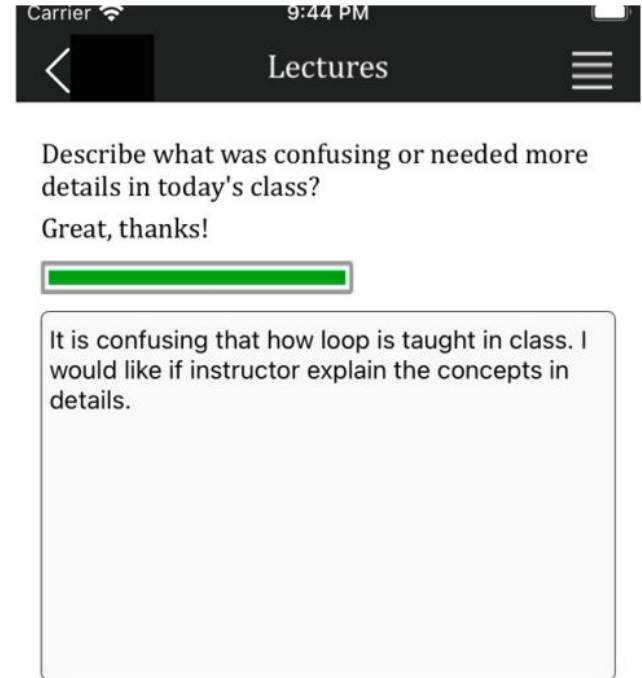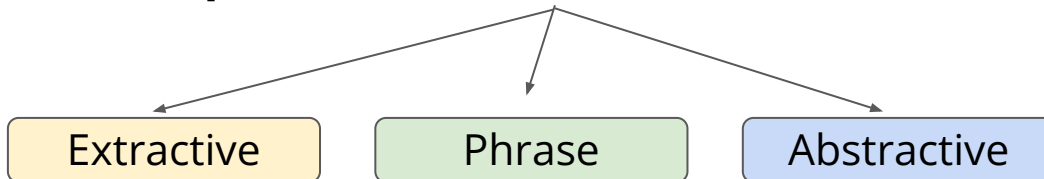
It is confusing that how loop is taught in class. I would like if instructor explain the concepts in details.

CourseMIRROR Data Collection

- Step 1: Students are prompted to write a reflection about what they found confusing or interesting
- Step 2: Reflections are collected through the CourseMIRROR app
- **Step 3:** Annotate reflections' summaries

Extractive   Phrase   Abstractive

Course Reflection Summarization

1. Offering reflection summaries can aid instructors in **enhancing lecture preparation and supporting students**

Reflection Prompt
Describe what you found most interesting in today's class

Student Reflections
- Nothing in particular today
- Despite the confusion, I did find setting up these problems to be very interesting and rewarding.
- Equipotentials
- i thought the breakout room questions were interesting because i learned how to do questions
- I found the last problem in class the most interesting because it was proven we can derive almost anything.
- The most interesting thing was that finding electric potential doesn't require a path, but only the magnitude of the charge and it's distance from the point of interest.
- I really enjoy line integrals and I can tell that we're moving towards using them to calculate potential.
- Collection of point charges (pairing them)
- How we can calculate something so complicated as electrons passing through an area is very cool.
- I found equipotentials to be the most interesting thing, especially drawing a equipotentials for a dipole!
- I thought it was interesting that Vnet is equal to all Vs added together
- I found how conductors act to be interesting.

Course Reflection Summarization

1. Offering reflection summaries can aid instructors in enhancing lecture preparation and supporting students

2. Course reflections **vary in the length and structure** (different from opinion summarization)

Reflection Prompt
Describe what you found most interesting in today's class

Student Reflections
● Nothing in particular today
● Despite the confusion, I did find setting up these problems to be very interesting and rewarding.
● Equipotentials
● I thought the breakout room questions were interesting because i learned how to do questions
● I found the last problem in class the most interesting because it was proven we can derive almost anything.
● The most interesting thing was that finding electric potential doesn't require a path, but only the magnitude of the charge and it's distance from the point of interest.
● I really enjoy line integrals and I can tell that we're moving towards using them to calculate potential.
● Collection of point charges (pairing them)
● How we can calculate something so complicated as electrons passing through an area is very cool.
● I found equipotentials to be the most interesting thing, especially drawing a equipotentials for a dipole!
● I thought it was interesting that Vnet is equal to all Vs added together
● I found how conductors act to be interesting.

Course Reflection Summarization

1. Offering reflection summaries can aid instructors in enhancing lecture preparation and supporting students

2. Course reflections vary in the length and structure (different from opinion summarization)

3. **Automatic summarization** can help **scale the use** of reflections in educational practice

Lack of Benchmarks!

Reflection Prompt
Describe what you found most interesting in today's class

Student Reflections
● Nothing in particular today
● Despite the confusion, I did find setting up these problems to be very interesting and rewarding.
● Equipotentials
● i thought the breakout room questions were interesting because i learned how to do questions
● I found the last problem in class the most interesting because it was proven we can derive almost anything.
● The most interesting thing was that finding electric potential doesn't require a path, but only the magnitude of the charge and it's distance from the point of interest.
● I really enjoy line integrals and I can tell that we're moving towards using them to calculate potential.
● Collection of point charges (pairing them)
● How we can calculate something so complicated as electrons passing through an area is very cool.
● I found equipotentials to be the most interesting thing, especially drawing a equipotentials for a dipole!
● I thought it was interesting that Vnet is equal to all Vs added together
● I found how conductors act to be interesting.

Our Work

- Manually annotated corpus with different types of summaries and rich metadata

- A comprehensive benchmarking of models for varied summarization tasks

Our Work

- **Manually annotated corpus with different types of summaries and rich metadata**

- A comprehensive benchmarking of models for varied summarization tasks

ReflectSumm Dataset

Metadata
Gender; Ethnicity; Age

Reflection Prompt
Describe what you found most interesting in today's class

Student Reflections
• Nothing in particular today -> 1.0
• Despite the confusion, I did find setting up these problems to be very interesting and rewarding. -> 3.0
• Equipotentials -> 2.0
• i thought the breakout room questions were interesting because i learned how to do questions -> 4.0
• I found the last problem in class the most interesting because it was proven we can derive almost anything. -> 4.0
• The most interesting thing was that finding electric potential doesn't require a path, but only the magnitude of the charge and it's distance from the point of interest. -> 4.0
• I really enjoy line integrals and I can tell that we're moving towards using them to calculate potential. -> 4.0
• Collection of point charges (pairing them) -> 2.0
• How we can calculate something so complicated as electrons passing through an area is very cool. -> 3.0
• I found equipotentials to be the most interesting thing, especially drawing a equipotentials for a dipole! -> 4.0
• I thought it was interesting that Vnet is equal to all Vs added together -> 4.0
• I found how conductors act to be interesting. -> 3.0

Annotated Specificity Score

⭐ Extractive Summary

• I found equipotentials to be the most interesting thing, especially …
• The most interesting thing was that finding electric potential doesn't require a path …
(three more reflections selected from input)

12

ReflectSumm Dataset

Metadata
Gender
Ethnicity
Age

Reflection Prompt
Describe what you found most interesting in today's class

Student Reflections
- Nothing in particular today -> 1.0
- Despite the confusion, I did find setting up these problems to be very interesting and rewarding. -> 3.0
- Equipotentials -> 2.0
- i thought the breakout room questions were interesting because i learned how to do questions -> 4.0
- I found the last problem in class the most interesting because it was proven we can derive almost anything. -> 4.0
- The most interesting thing was that finding electric potential doesn't require a path, but only the magnitude of the charge and it's distance from the point of interest. -> 4.0
- I really enjoy line integrals and I can tell that we're moving towards using them to calculate potential. -> 4.0
- Collection of point charges (pairing them) -> 2.0
- How we can calculate something so complicated as electrons passing through an area is very cool. -> 3.0
- I found equipotentials to be the most interesting thing, especially drawing a equipotentials for a dipole! -> 4.0
- I thought it was interesting that Vnet is equal to all Vs added together -> 4.0
- I found how conductors act to be interesting. -> 3.0

Annotated
Specificity Score

Extractive Summary

- I found equipotentials to be the most interesting thing, especially …
- The most interesting thing was that finding electric potential doesn't require a path …
(three more reflections selected from input)

Abstractive Summary

The students today found calculations and relationships to other concepts that they have learned in this and other classes interesting. They also found potential energy and equipotentials very interesting, as well as some integration concepts.

Phrase Summary

- equipotentials
- calculations
- relations to old concepts
- potential
- integration

13

ReflectSumm Dataset

- Student reflections on lectures collected from real-world lectures
- **782** input-summary pairs from 24 university courses, spanning four STEM subjects

*We only included data from students who explicitly provided consents, and carefully removed all private information

ReflectSumm Dataset

- Student reflections on lectures collected from real-world lectures
- **782** input-summary pairs from 24 university courses, spanning four STEM subjects
- **Manual annotations** for each pair of input-summary
  - Individual input reflection has annotated specificity score and metadata of the author (demographic information*)
  - Three types of summaries are manually annotated by high-quality in-house annotators.

*We only included data from students who explicitly provided consents, and carefully removed all private information

Dataset Analysis

Lecture Distribution

COMPINF 4.6%
PHYS 18.2%
ENGR 56.9%
CS 20.3%
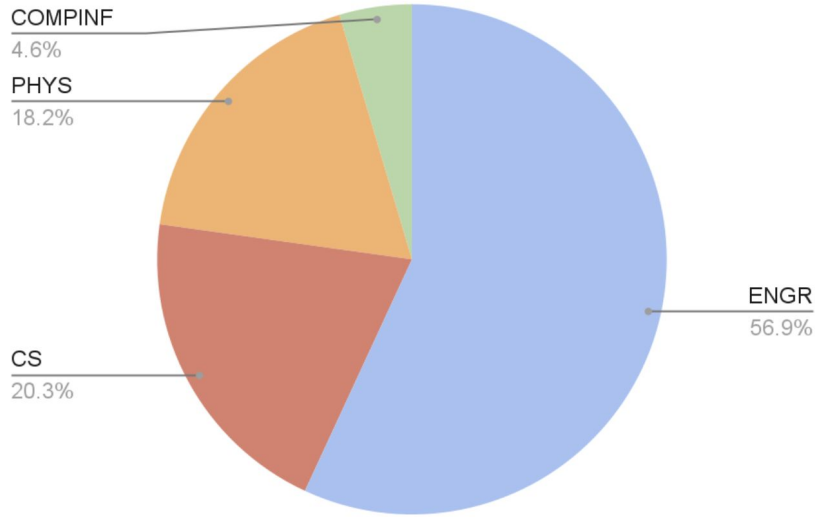
Our dataset covers diverse subjects

Reflection Specificity Distribution

Score of 1 10.7%
Score of 4 22.6%
Score of 2 14.2%
Score of 3 52.5%
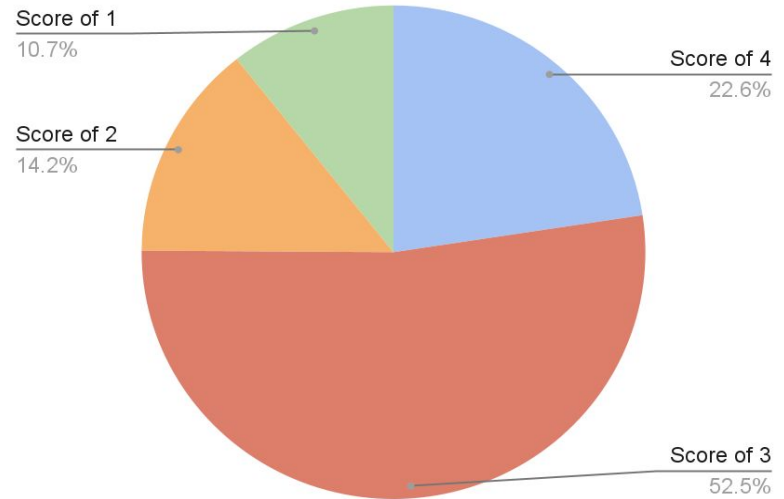
Students write high quality reflections with details

Our Work

- Manually annotated corpus with different types of summaries and rich metadata

- **A comprehensive benchmarking of models for varied summarization tasks**

Extractive	Phrase	Abstractive

Extractive Summarization

Task: Select 5 reflections as the summary

Extractive   Phrase   Abstractive

Extractive Summarization

Task: Select 5 reflections as the summary

Models:

- **BERTSUM-EXT** (Liu and Lapta, 2019)
    - Fine-tuned on CNN/DM (287K) or ReflectSumm (less than 800)
- **MatchSUM** (Zhong et al., 2020)
- **ChatGPT (GPT)** with different prompts
    - Reflections only
    - Reflections + Specificity

Eval Metrics: ROUGE (R-1/R-2/R-L); BERTScore; Partial Match F1

Extractive	Phrase	Abstractive

Extractive Summarization

Legend: R-L (F1 score), BERTScore (F1 score), Partial Match (F1 score)

	R-L (F1 score)	BERTScore (F1 score)	Partial Match (F1 score)
BERT-EXT (cnndm)	53.39	89.91	37.74
BERT-EXT (ReflectSumm)	54.25	89.94	37.49
MatchSum	56.7	90.57	38.94
GPT-reflect	58.26	89.98	37.68
GPT-reflect + specificity	56.85	90.29	36.45
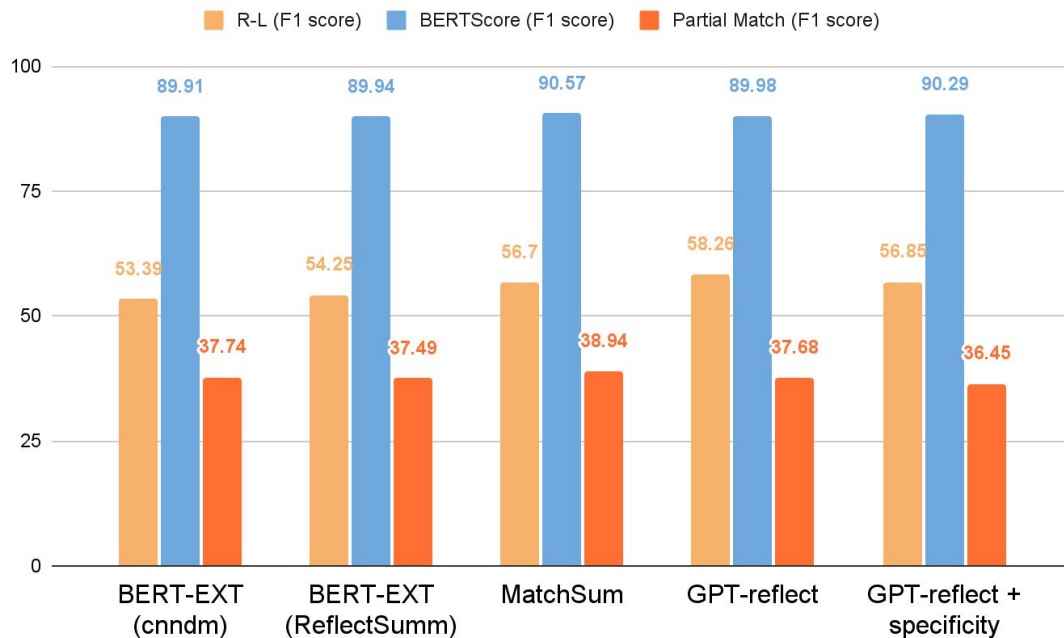
Extractive   Phrase   Abstractive

Extractive Summarization

The performance slightly improves when utilizing the BERT-EXT model trained on our dataset, which <u>is much smaller</u> than the CNN/DM dataset

Extractive	Phrase	Abstractive

Extractive Summarization

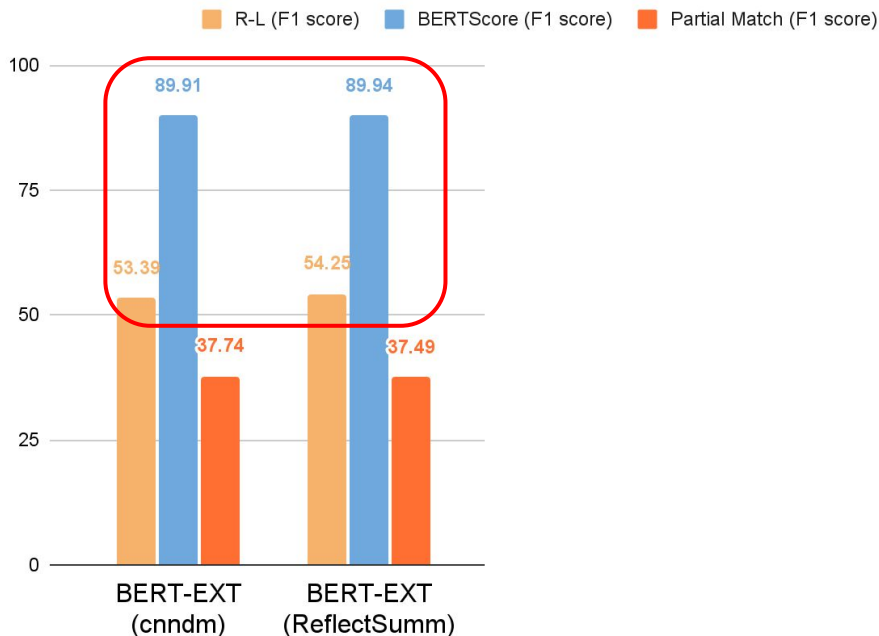The zero-shot GPT-based model achieves comparable or superior performance compared to robust baselines, but it may occasionally struggle to faithfully extract complete reflections

R-L (F1 score)   BERTScore (F1 score)   Partial Match (F1 score)

	MatchSum	GPT-reflect	GPT-reflect + specificity
R-L (F1 score)	56.7	58.26	56.85
BERTScore (F1 score)	90.57	89.98	90.29
Partial Match (F1 score)	38.94	37.68	36.45
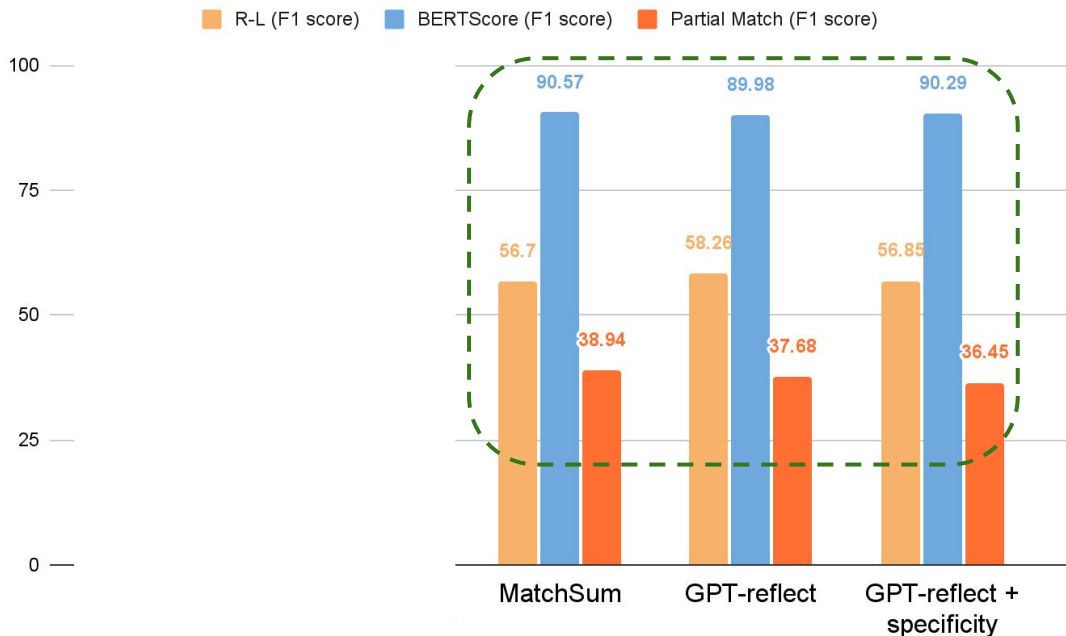
Extractive   Phrase   Abstractive

Phrase Summarization

Task: Generate 5 phrases to summarize the reflections

Models:

- **PhraseSum**: deployed baseline (Luo and Litman 2015)
- **ChatGPT**
    - GPT-noun: prompt to generate noun phrases
    - GPT-Human: A more intricate prompt with human guidelines

Eval Metrics: ROUGE (R-1/R-2/R-L); BERTScore

Extractive	Phrase	Abstractive

23
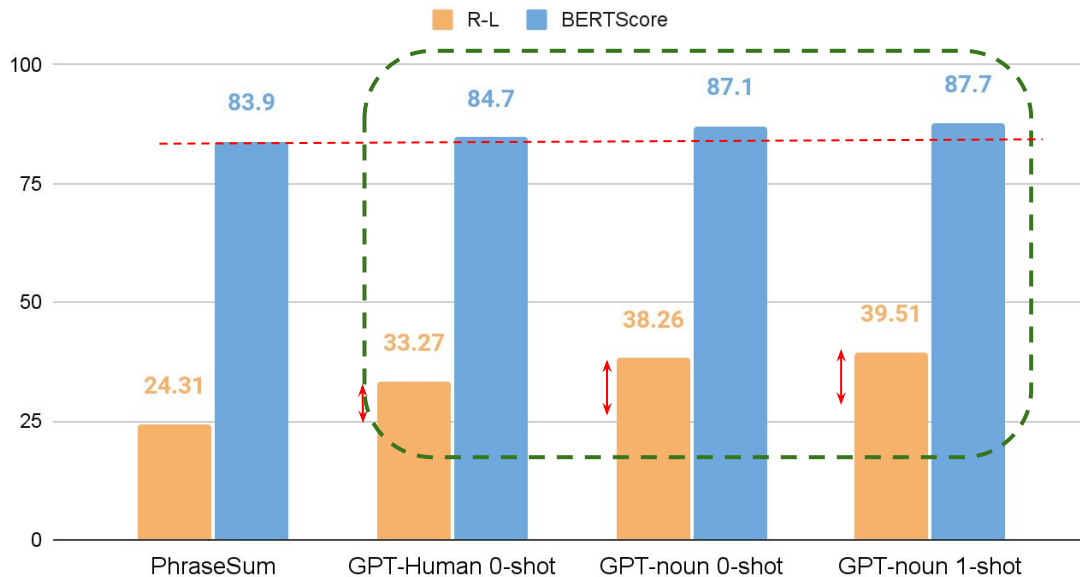
Phrase Summarization

GPT models outperform the baseline by large margin

GPT-based models are sensitive to the prompt, and few-shot could help produce higher-quality phrases

| Extractive | **Phrase** | Abstractive |

Abstractive Summarization

Finetune pretrained encoder-decoder

- Vanilla fine-tuning
- Fine-tuning with reflection specificity

Prompt GPT-3.5-turbo

- Zeroshot prompting
- Zeroshot prompting reflection specificity

Extractive   Phrase   **Abstractive**

Abstractive Summarization

Evaluation

- **Reference-based metrics**
  ROUGE (R-1/R-2/R-L); and BERTScore

- **Factuality metrics**
  Entailment based approaches — SummaC (Laban et al., 2022)

Extractive

Phrase

Abstractive

Abstractive Summarization

Reference-based metrics

Specificity information showed positive influence in case of fine-tune pretrained models

Model	R-1	R-2	R-L	BS
BART-Large	47.09	24.17	43.76	90.49
+ specificity	**47.70**	**24.85**	**44.41***	**90.57**
GPT-Human	35.83	9.40	31.85	88.23
+ specificity	36.73	9.13	31.64	88.27
GPT-one-shot	36.86	9.46	31.96	88.26

Extractive   Phrase   **Abstractive**

Abstractive Summarization

Factuality Metrics: SummaC

Is GPT more factual than human reference?

Model	SummaC ↑	
	Sentence	Document
Human-reference	0.25	0.22
BART-Large	0.25	0.21
+ specificity	0.25	0.22
GPT-Human	0.26	**0.31**
+ specificity	**0.27**	0.26
GPT-one-shot	0.26	0.26

Extractive	Phrase	**Abstractive**

Abstractive Summarization

Example

> - One thing I found interesting was how many categories of machine learning there are.
> - <u>Supervised and unsupervised learning</u> as well as discrete and continuous labels and how they all related to one another.
> - <u>Different categories of machine learning.</u>
> - The relationship between unsupervised and supervised deep learning.

Source Document

> Students enjoyed learning about the differences between <u>supervised and unsupervised learning</u>. Along with that, they also enjoyed learning about the <u>different categories in Machine Learning</u> and the different categorization and classification methods.

✔

Factual Summary

Extractive	Phrase	**Abstractive**

Abstractive Summarization

Example

**SummaC: 0.2
(low factuality)**

- One thing I found interesting was how many categories of machine learning there are.
- Supervised and unsupervised learning as well as discrete and continuous labels and how they all related to one another.
- Different categories of machine learning.
- The relationship between unsupervised and supervised deep learning.

Source Document

Students enjoyed learning about the differences between supervised and unsupervised learning. Along with that, they also enjoyed learning about the different categories in Machine Learning and the different categorization and classification methods.

Factual Summary

Extractive	Phrase	Abstractive

Abstractive Summarization

Example

SummaC: 0.2
(low factuality)

- One thing I found interesting was how many categories of machine learning there ar...
- Supervi... well as ... how the...
- Differen...
- The rela...nship between unsupervised and supervised deep learning.

Students enjo...rning about the ...supervised and ...g. Along with ...ed learning ...ategories in ...d the different categorization and classification methods.

We need to develop better factuality metrics for these types of documents

Source Document

Factual Summary

Extractive | Phrase | **Abstractive**

Broader Impact

- Rich metadata (demographic information of student) can be applied for studies on fairness / equity issues

- Our models can enable new downstream functionalities such as generating recommended readings and explaining confusing cor based on summary output

For instance, our dataset can enable researchers to analyze **differences in reflection submission rates** and **specificity among different student groups**

Broader Impact

- Rich metadata (demographic information of student) can be applied for studies on fairness / equity issues.

- Our models can enable new downstream functionalities such as generating recommended readings and explaining confusing concepts based on summary output.

Future Work

- Investigating improved prompting techniques.

- Extending the corpus to other subject domains (i.e. Psychology).

- Designing a dynamic system capable of adjusting the quantity of extracted summary outputs according to the size of the lecture.

Conclusion

- Manually annotated corpus with **different types of summaries** and **rich metadata** (age, gender, ethnicity)

Extractive	Phrase	Abstractive

- A **comprehensive benchmarking** of models for varied summarization tasks

- Check the prompts, system outputs, and dataset at https://github.com/EngSalem/ReflectSUMM
- Contact:
  - Yang Zhong yaz118@pitt.edu
  - Mohamed Elaraby mse30@pitt.edu

NSF

IES Institute of Education Sciences