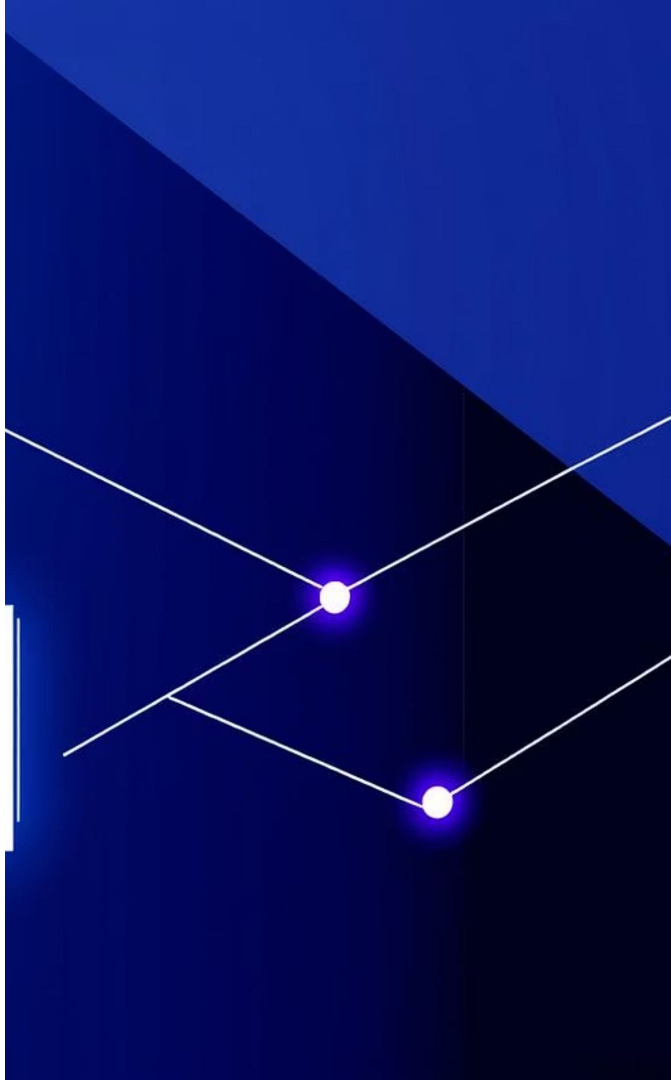
GLIMPS: Summarizing Scholarly Review

A novel method for multi-document summarization of scholarly reviews, extracting both common and unique opinions.

Mozhdeh Hajiani 330007
Vida Ahmadi 301905

Politecnico di Torino



The Challenge of Peer Review

Increasing Submissions

Conferences like ICLR and ACL have been a fivefold increase in submissions since 2017,straining the review process.

Area Chair Burden

Area chair must read a growing volume of reviews to discern main arguments and make decisions.

Limitations of Traditional Summarization

Consensus-Based

Existing methods focus on common opinion, failing to identify divergent or unique arguments crucial for peer review.

Lack of Attributability

Many methods don’t allow tracing synthesized opinions back to their source, hindering transparency.

Introducing GLIMPSE

- GLIMPSE is a summarization method designed to offer a concise yet comprehensive overview of scholarly reviews.
- The goal is to generate a summary for each review that highlights commonalities, differences, and unique perspectives, and distinguishing one review from others of the same submission.

Novel Approach

- Novel summarization method for scholarly reviews and cast it as a reference game problem.
- Extracts both common and unique opinions.

RSA Framework

- Uses Rational Speech Act (RSA) framework to identify unique and common opinions.
- Measure the informativeness of opinions in scholarly reviews.
- Framework for pragmatic modeling rooted in Bayesian inference that solves the reference game setting

Uniqueness Scores

- Introduces novel scores to identify relevant sentences.
- Measure the uniqueness of opinions in scholarly reviews.

Discriminative Multi-Document Summarization (D-MDS)

They introduce D-MDS, a new task where the goal is to provide a summary for each review, enabling quick identification of the source review

$$V_L(s,d) = \log L(d|s) - \text{Cost}(d), \qquad L_0(d|s) = \frac{\text{LM}(s|d)}{\sum_{d \in \mathcal{O}} \text{LM}(s|d)}.$$

$$S_t(s|d) = \frac{\exp(V_{L_{t-1}}(d,s))}{\sum_{s'} \exp(V_{L_{t-1}}(d,s'))}. \qquad L_t(d|s) = \frac{S_t(s|d)}{\sum_{d'} S_t(s|d')}.$$

This problem is formally mapped to a “reference game” setting, where a speaker describes a target object and a listener identifies it.

GLIMPSE Framework: Key Scores

Pragmatic-Speaker-Based Score (GLIMPSE-Speaker)

- Informativeness (RSA speaker)
- Identifies the most discriminative utterance to refer to a source document

$$\text{RSA-Speaker}(d) \triangleq \arg\max_{s \in \mathcal{C}} S_t(s|d)$$

Uniqueness Score (GLIMPSE-Unique)

- Uniqueness (KL divergence from uniform)
- Measures how far the listener distribution conditioned to summary is far from the uniform distribution.
- Measures how unique a candidate summary is, or its divergence from others.
- High values of uniqueness score indicate the uniqueness of the candidate summary or its divergence from other candidates while lower values indicate that the candidate summary is common to multiple source documents.

$$\text{Unique} \triangleq D_{\text{KL}}(L(\cdot|\ s) \| \mathcal{U})$$

These scores help identify unique or common opinion and select the most informative summary.

Summary Generation Steps

- 1. Generate candidate summaries (extractive/abstractive).
- 2. Score using RSA-based informativeness and uniqueness.
- 3. Compose final summary with template (common + unique insights).

Dataset

📚 Dataset Source: (ICLR submissions 2017–2021)

- 28,062 reviews from 8,428 submissions.
- Includes full text of peer reviews and corresponding meta-reviews.
- but only the 2017 reviews were used for evaluation in our work.

📌 Summary-like Meta-Reviews:

- Used as proxy gold summaries for evaluation.
- Selected based on:
  - Length and structure
  - Mention of specific reviewer points
  - Semantic similarity to reviews (cosine embedding)

✅ Provides real-world, domain-specific benchmark for summarization tasks.

Evaluation Metrics

The main evaluation approaches were used:

1. ROUGE Scores

- Measures overlap with meta-review text (gold summaries).
- Limitation: low correlation with human judgment.

2. SEAHORSE Metrics

- Human-aligned scores for assessing the summaries along six axes: Coverage, Attribution, Grammar, Conciseness, Repetition.
- High coverage across all reviews suggests that the summary effectively captures the main ideas of all the reviews.
- More reflective of summary quality beyond surface text overlap.

Summary Quality and Human Evaluation

Conciseness & Coverage

GLIMPSE generates more concise summaries with significantly better coverage of main ideas than baselines, and help identify source reviews.

Human Judgment

Human evaluators found GLIMPSE summaries more informative and unique than those from LLM.

Method	Discriminativeness
GLIMPSE	93.75%
Llama	85.18%
MDS	0%

1

Replicate GLIMPSE

On ICLR 2017 reviews to confirm trend alignment.

2

Adapt to PubMed

For biomedical text summarization.

3

Compare Models

BART, PEGASUS, and T5 for abstractive summarization.

Methodology - Reproducing GLIMPSE

We replicated the GLIMPSE framework, a discriminative multi-document summarization (D-MDS) approach for scholarly peer reviews.

Candidate Summary Generation

Extractive and abstractive methods (using BART).

Pragmatic Scoring & Selection

RSA-based scores: Pragmatic-Speaker (informativeness) and Uniqueness (opinion divergence).

Summary Composition

Sentences ranked by RSA scores, filtered, and structured.

Methodology - Reproducing GLIMPSE

Libraries:

PyTorch 2.0
Transformers 4.28.0

Hardware:

NVIDIA A100 (40 GB)

Dataset:

Due to resource constraints, we used a smaller ICLR 2017 dataset (490 reviews) instead of the full 2017–2021 dataset (28,062 reviews).

Challenges:

- GPU memory constraints
- Batch size = 1
- Gradient checkpointing
- Input truncation to 1024 tokens

Replication Results: ICLR 2017

Our ROUGE-1 scores (0.10–0.19) were lower than the original paper's (0.22–0.39) due to a smaller dataset and resource constraints.

ICLR 2017 Replication Results

Method	R-1	R-2	R-L	Cov.	Attr.	Gram.	Conc.	Repet.
Extractive								
Random	0.14±0.08	0.02±0.02	0.10±0.05	0.19±0.22	0.90±0.10	0.77±0.19	0.27±0.16	0.95±0.13
LSA	0.19±0.08	0.02±0.03	0.12±0.04	0.26±0.23	0.93±0.04	0.75±0.19	0.34±0.15	0.97±0.10
LexRank	0.17±0.09	0.02±0.03	0.11±0.05	0.26±0.23	0.93±0.06	0.77±0.20	0.38±0.17	0.97±0.09
GLIMPSE-Speaker	0.17±0.07	0.02±0.02	0.12±0.05	0.05±0.10	0.08±0.08	0.42±0.29	0.05±0.05	0.90±0.19
GLIMPSE-Unique	0.18±0.07	0.03±0.03	0.12±0.04	0.06±0.11	0.09±0.08	0.66±0.23	0.06±0.05	0.94±0.10
Abstractive								
Llama7bInstruct	0.10±0.07	0.03±0.02	0.06±0.03	0.51±0.16	0.85±0.04	0.16±0.10	0.23±0.06	0.87±0.11
GLIMPSE-Speaker	0.19±0.07	0.03±0.02	0.13±0.04	0.19±0.06	0.62±0.03	0.64±0.29	0.25±0.04	0.94±0.05
GLIMPSE-Unique	0.18±0.07	0.03±0.02	0.12±0.04	0.05±0.10	0.08±0.08	0.77±0.21	0.06±0.06	0.97±0.04

Extension 1 - PubMed Adaptation

To assess GLIMPSE's domain adaptability, we fine-tuned BART on a PubMed subset (10,000 articles).

Challenges Addressed

- Specialized terminology
- Structured abstracts
- High information density

Candidate Generation

- Abstractive summaries via beam search
- Extractive summaries via TF-IDF

RSA reranking applied GLIMPSE-Speaker and GLIMPSE-Unique scores, with evaluation using ROUGE and SEAHORSE metrics.

Extension 1 - PubMed Adaptation

Dataset:

- PubMed
- Used ~2,500 articles (25% of 10,000)

Model:

- Fine-tuned BART (facebook/bart-large-cnn)

Settings:

- 3 epochs
- Learning rate = 5e-5
- Input truncation = 1024 tokens

Extension 1 - Results

Extension 1 (PubMed) showed extractive methods with lower ROUGE but improved SEAHORSE metrics, indicating enhanced factual accuracy.

EXTENSION 1: DOMAIN ADAPTATION RESULTS

Method	R-1	R-2	R-L	Cov.	Attr.	Gram.	Conc.	Repet.
Extractive								
GLIMPSE-Speaker	0.14±0.08	0.05±0.06	0.10±0.06	0.10±0.08	0.39±0.09	0.59±0.21	0.16±0.06	0.96±0.07
GLIMPSE-Unique	0.13±0.07	0.04±0.05	0.09±0.05	0.08±0.08	0.38±0.07	0.53±0.22	0.14±0.05	0.95±0.10
Abstractive								
GLIMPSE-Speaker	0.20±0.09	0.08±0.07	0.14±0.07	0.15±0.09	0.41±0.10	0.29±0.19	0.18±0.07	0.91±0.14
GLIMPSE-Unique	0.19±0.08	0.08±0.07	0.14±0.07	0.15±0.09	0.41±0.10	0.51±0.24	0.18±0.07	0.93±0.13

Extension 2 - Model Comparison

We compared BART, PEGASUS, and T5 on the ICLR 2017 dataset to identify the most effective abstractive summarization model.

1	2	3
PEGASUS Summarization-focused model with Gap-Sentence Generation pretraining.	**T5** General-purpose text-to-text transformer, reformulates all NLP tasks.	**Fine-tuning** All models fine-tuned with consistent settings.

- Aim:See if pretrained models outperform BART for peer review summarization

Extension 2 - Results

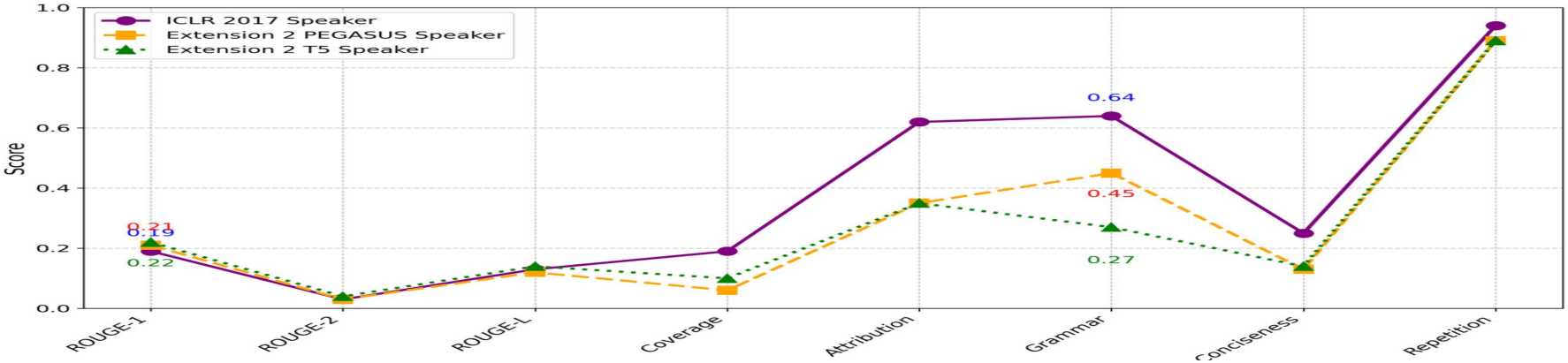
Extension 2 (PEGASUS, T5) showed T5 achieving the highest ROUGE scores (R-1: 0.22), surpassing ICLR 2017 abstractive results. However, both models had lower SEAHORSE scores, particularly in Grammar and Coverage.

EXTENSION 2: PEGASUS AND T5 RESULTS

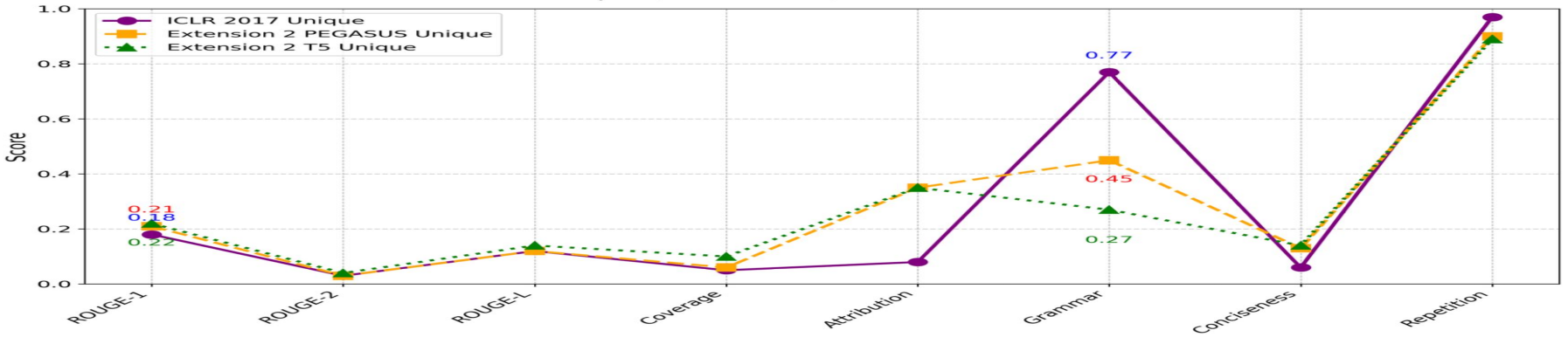
Method	R-1	R-2	R-L	Cov.	Attr.	Gram.	Conc.	Repet.
PEGASUS (Abstractive)								
GLIMPSE-Speaker	0.21±0.08	0.03±0.03	0.12±0.04	0.06±0.06	0.35±0.02	0.45±0.24	0.13±0.03	0.89±0.12
GLIMPSE-Unique	0.21±0.08	0.03±0.03	0.12±0.04	0.06±0.06	0.35±0.02	0.45±0.24	0.13±0.03	0.90±0.12
T5 (Abstractive)								
GLIMPSE-Speaker	0.22±0.08	0.04±0.04	0.14±0.06	0.10±0.07	0.35±0.03	0.27±0.14	0.14±0.04	0.89±0.13
GLIMPSE-Unique	0.22±0.08	0.04±0.04	0.14±0.06	0.10±0.07	0.35±0.03	0.27±0.14	0.14±0.04	0.89±0.13

Extension 2 - Results

GLIMPSE speaker(Abstractive): Extension 2 vs ICLR 2017



GLIMPSE Unique(Abstractive): Extension 2 vs ICLR 2017



Conclusion

GLIMPSE replication validated its effectiveness, outperforming Random and rivaling baselines. PEGASUS and T5 further improved ROUGE scores, with T5 showing superior bigram and sequence overlap.

However, Grammar and Attribution scores declined, suggesting limitations due to dataset size, computational constraints, and minimal fine-tuning.

GLIMPSE reproducible but resource-constrained.

RSA scoring remains effective.

PubMed adaptation:

- Slight drop in ROUGE
- Better quality in human-aligned metrics

T5 shows promise:

- Highest ROUGE
- Needs fluency and coverage improvements

Thank you for your attention!