Multi-Model Rational Speech Act Framework for Scientific Peer Reviews Summarization

Natalia Lebedeva
s309608@studenti.polito.it
Politecnico di Torino

Shakti Rathore
s328222@studenti.polito.it
Politecnico di Torino

Ali Al Housseini
s333940@studenti.polito.it
Politecnico di Torino

***Abstract*—In the academic publishing context, peer reviews play a crucial role. However, they are both essential and challenging, as major conferences receive an overwhelming number of paper submissions to process. Each submission receives multiple reviews, and area chairs are responsible for analyzing and synthesizing these reviews, a task that is becoming increasingly demanding. Our work builds upon a previous study that introduced a novel approach to summarizing reviews by identifying both the most common and the most unique aspects of each review before combining them. A key contribution of that work was the application of the Rational Speech Act (RSA) framework in the context of summarization. Using this work as a baseline, we propose two extensions. The first employs an ensemble of models for the RSA framework, using an aggregation function on the candidate summary scores. The second introduces a vectorized form of RSA, also leveraging multiple models to construct the arrays used in the process. Our experimental results demonstrate improved performance compared to baseline methods across several evaluation metrics.**
Repository: ***https://github.com/alialhousseini/glimpse-mds***

I. Introduction

Peer review is the foundation of the scientific world. To understand the scale of this process, in 2023, the Association for Computational Linguistics (ACL), one of the most well-known conferences in computer science, received **4,864** submissions.

To provide authors with meaningful feedback on their work, area chairs must summarize the reviews received for each submission, highlighting the most common opinions while also reporting the unique insights provided by individual reviewers. Several methods have been proposed for summarizing scientific reviews, but most focus primarily on identifying and reporting the dominant sentiment or most common opinions. In [1], the authors introduced a methodology that incorporates not only the most prevalent viewpoints but also the most divergent aspects of reviews. A key innovation of their approach was the introduction of the **Rational Speech Act (RSA) framework** in the context of summarization.

Building on this work, we propose two extensions aimed at improving the method. Our objective is to develop a multi-document summarization pipeline capable of combining different reviews for a paper while effectively capturing both the **common perspectives** of the reviewers and the **unique insights** that could be highly valuable for final feedback.

For each set of reviews, we generate candidate summaries. By leveraging a **large language model (LLM)**, we obtain **unnormalized scores** that reflect the representativeness of each candidate with respect to the reference reviews. Iterating over these scores using the RSA framework allows us to compute final values that capture both the **uniqueness** and **informativeness** of each candidate.

Our main contributions can be divided into two key approaches:

1) **Model Ensemble for RSA** – In our first extension, we studied and selected different models to compute the initial probabilities. After this selection, we computed the probabilities using a combination of these models, aggregating the scores they produced in order to exploit their varying strengths—all before applying RSA.
2) **Vectorized RSA** – In our second approach, we implemented a vectorized version of RSA. Instead of aggregating model outputs at the beginning of the process, we preserved the strengths of different models throughout the pipeline, performing aggregation only at the final stage before generating the summary.

We conducted extensive experiments on both approaches and compared the results with those obtained in [1], as well as with state-of-the-art multi-document summarization methods. Our experiments were performed on the same dataset used in [1], which includes review data from the **ICLR** conference collected between **2017 and 2021**.

II. Related Work

The task of **multi-document summarization (MDS)** has been addressed by many researchers. In the survey of deep learning-based methods for MDS [2], authors study the variety of existing approaches for extractive, abstractive and hybrid MDS. They distinguish different network design strategies: from *naive networks* where documents are input into one deep NN to extract features and generate summaries, to *ensemble networks* leveraging several models to obtain better results. The authors of the survey highlight the need to develop MDS models that capture cross-document relations. Another promising development approach concerns exploring LMs that specifically target long input sequences, such as Longformer [3] or BigBird [4]. We take these findings as inspiration to use an *ensemble of LMs* to solve the problem at hand.

The feasibility of using summarization and other NLP techniques to produce **scientific paper reviews** was explored in [5]. While NLP-based systems can assist in the peer review process by highlighting key contributions and identifying

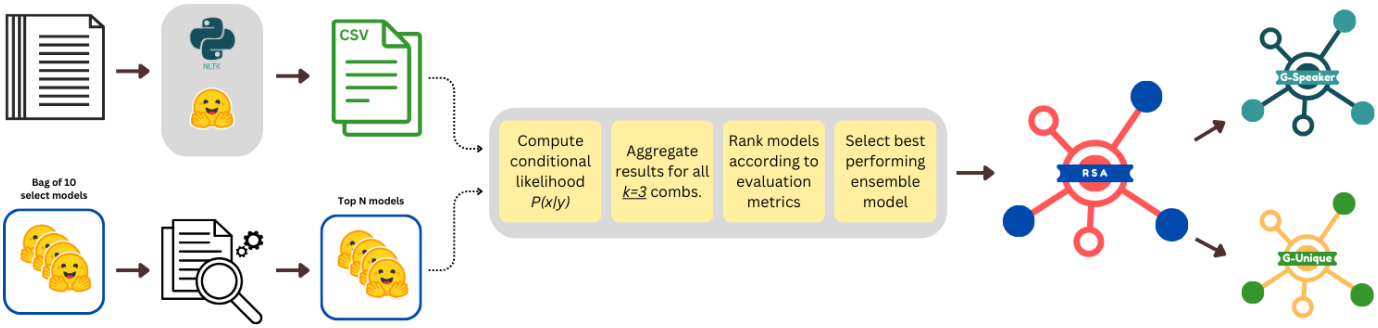

Fig. 1. Overview of the main architecture

minor issues, they struggle with deeper evaluations of theoretical accuracy and novelty. The authors outline the essential qualities of a good scientific review, emphasizing criteria such as decisiveness, comprehensiveness, and justification. This highlights the necessity of human judgment in evaluating AI-generated meta-reviews.

Our research is based on the GLIMPSE method [1]. The key idea is to generate a meta-review for the area chair, combining both *most common* and *divergent* opinions extracted from the reviews by applying the RSA framework [6]. RSA is a probabilistic model of communication, incorporating Bayesian inference to model how speakers and listeners interact in a cooperative manner. In the extensions we propose, we try to address the limitations of the GLIMPSE such as low fluency of the generated summaries, sensitivity to the sentence segmentation process in the extractive setting, and codebooking phenomena in GLIMPSE-Speaker.

III. Methodology

The general setting for multi-document summarization is as follows: Let $\mathbf{D} = \{d_1, ..., d_N\}$ be the set of documents (articles) and $\mathbf{C} = \{s_{i,j}\}_{1 \leq i \leq N, 1 \leq j \leq K}$ set of candidates summaries. The generation of candidate summaries is done through one of these approaches:

- **Extractive Approach** – Identifying and refining candidate summaries by segmenting and preprocessing sentences from the reviews associated with each paper.
- **Abstractive Approach** – Leveraging a LLM to generate a predefined number of candidate summaries per review. For this, we employed three models:
  1) BART [7] Large CNN
  2) Pegasus Arxiv [8]
  3) Pegasus Large [8]

As a result, we obtain **four** datasets with different set of candidates for each document. We denote them as $\mathbf{X} = \{x_{extr}, x_{BART}, x_{pgl}, x_{pg-arxiv}\}$.

In the proposed pipeline, the multi-document summarization problem is framed as a reference game, leveraging the RSA framework as described in [1]. The process starts with a likelihood matrix, where each candidate is assigned a probability of representing one of the source documents (reviews) of the same article.

After iterating over these values using RSA, we obtain two separate tables containing listener and speaker probabilities. These tables allow us to extract both the **most common** and the **most unique candidates**, which are then used to generate the final summary.

Utilizing the methods and newly introduced metrics from [1], we generate two distinct types of summaries. When relying solely on the uniqueness score, which emphasizes the most common and unique candidates, we obtain the *GUnique* summary. Alternatively, by incorporating the RSA-speaker score alongside the uniqueness score, we produce the *GSpeaker* summary.

A. Extension 1: Model Ensemble for RSA

There are several LLMs, both specialized and general-purpose, for summarization, as well as various metrics to evaluate the quality of a summary. Therefore, our first approach is to leverage the strengths of different models instead of relying on a single one.

In [1], the initial likelihood matrix is computed using a single model that is fine-tuned for the summarization task. In contrast, in our approach, we compute the initial probabilities using multiple models. Then, for each candidate, once we obtain the initial scores returned by each model, we aggregate them.

In this way, the likelihood matrix before applying the RSA process contains scores obtained through an ensemble of models, capturing diverse perspectives.

B. Extension 2: Vectorized RSA

In our second approach, we continue to leverage multiple models but adopt a different strategy to preserve and exploit their individual characteristics.

Instead of aggregating the scores before applying RSA, we maintain a **vectorized representation** of the likelihood scores for each candidate. Each dimension of this vector corresponds to a likelihood score computed by a different model. This approach allows us to retain the unique behavior and strengths of each model rather than merging them prematurely into a single value.

Finally, after the RSA inference step, we aggregate the refined scores to compute the final uniqueness and informativeness values.

IV. Experimental Setup

A. Data Preparation

The authors of the original paper used the dataset described in Section I, however, for evaluation, they restricted the number of articles to **226**. As described in [1], the most promising set of paper-review pairs was selected based on a set of sequential stages: data preprocessing, manual verification, and cosine similarity between meta-reviews (named as 'gold') and reviews. We follow the same approach to obtain our top-226 articles used for evaluation.

B. Extension 1

A crucial aspect of our work involved selecting appropriate models. Following the pipeline from Figure 1, first, we establish a set of models capable of generating summaries relevant to our context. To achieve this, we tested multiple models and performed a preliminary manual assessment of their outputs. Based on these evaluations, we refined our selection to a set of 10 models, which are detailed in the Table A1.

Separately, for each dataset in $\mathbf{X}$, our goal is to select Top-N (N=5) models for generating likelihood scores. For this reason, we conducted an extensive experimental study to investigate evaluation results based on BERTScore [9], and UniEval [10], separately for GLIMPSE-Speaker and GLIMPSE-Unique settings from [1].

By now, we generate all the possible $k = 3$ combinations out of the N models depicted. More precisely, we aggregate conditional likelihoods by averaging them.

Afterward, we rank ensemble models based on their evaluation results, then, we select the best-performing one. Finally, we use the RSA framework to produce summaries combining the most common and the most unique (GUnique) or informative (GSpeaker) candidates.

C. Extension 2

While defining a set of models suitable for the summarization task, we observed that each model excels in certain aspects while underperforming in others. To capitalize on these strengths, we adopted a selection strategy based on key evaluation criteria. Specifically, we addressed the primary challenges of MDS by focusing on fluency, attribution, conciseness, consistency, and the preservation of main ideas in the generated summaries. These aspects were assessed using SEAHORSE [11] and UniEval [10].

By evaluating model performance across these metrics, we identified a set of optimal model-metric pairs, as summarized in Table A3. Our approach preserves the advantages of multiple models by representing the likelihood scores for each candidate-document pair as a vector, as illustrated in Figure 2. We extended the RSA framework to accommodate vectorized inputs and generate vectorized output scores. Finally, we aggregated these scores to construct summaries, following the methodology described in the previous section.

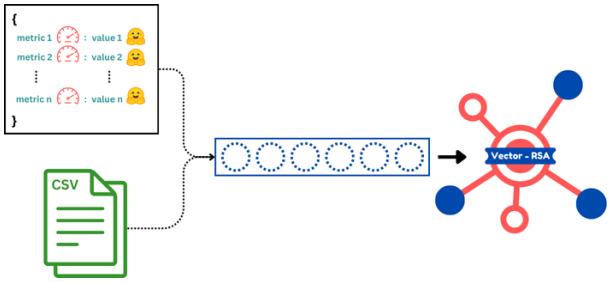

Fig. 2. Vectorized-RSA Framework

V. Results

To evaluate the quality of the summaries, we compared the generated summaries against the gold summary-like meta-reviews. Apart from discriminativeness, where we used the concatenated version of the reviews as a reference against the generated summary, we maintained this distinction in our evaluation to remain as consistent as possible with the reference implementation from the original paper. All the results are reported in Table I. Our evaluation primarily focuses on comparing the performance of the two proposed techniques against the results of GLIMPSE, as reported in [1]. In the abstractive setting, conducting a detailed comparative analysis presents challenges, as the original paper does not specify which model was used for candidate summary generation. This lack of transparency limits direct comparisons in this aspect.

Regarding the ROUGE scores, they overall remain relatively low, which is expected given that the gold meta-reviews often include information that does not directly overlap with the original reviews. In the extractive setting, both of our proposed extensions outperform the original GLIMPSE framework. However, in the abstractive case, our ROUGE scores are lower.

Furthermore, we investigated the absence of BERTScore in the evaluations reported in [1]. Our findings indicate that BERTScore yields consistently low results across all model variations, which likely explains why it was omitted from their evaluation criteria.

One of the key advantages of our approach is reflected in the results obtained using SEAHORSE [11] metrics. Notably, we achieved significant improvements in attribution and conciseness, along with generally higher scores for the preservation of main ideas. However, an unexpected drop in grammar scores was observed, particularly in the abstractive setting. In terms of repetition, our extractive summarization results are comparable to the baseline, whereas in the abstractive case, the scores are generally lower and highly dependent on the candidate generation model.

To further enhance our evaluation, we introduced UniEval [10] metrics, which assess coherence, consistency, and fluency between a reference text and a generated summary. Our approach achieves reasonable performance across these metrics aligned with human judgment.

Regarding the results obtained in our experimental setup, our two proposed extensions can be seen as comparable

		Method	Rouge-1	Rouge-2	Rouge-L	BERTScore	Coherence	Consis.	Fluency	Disc.	Attrib.	Concis.	Grammar	Cov.	Rept.
Extractive		Random	0.32	0.06	0.15	-	0.40	0.47	0.46	**0.11**	0.48	0.15	0.58	0.09	0.84
		LSA	0.38	0.08	0.16	-	0.41	0.42	0.36	0.04	0.57	0.19	0.59	0.14	0.84
		LexRank	0.38	0.09	0.17	-	0.34	0.44	0.36	**0.11**	0.56	0.21	**0.60**	0.18	0.85
		GLIMPSE-Speaker [1]	0.22	0.04	0.11	-	-	-	-	-	0.36	0.13	0.36	0.09	0.89
		GLIMPSE-Unique [1]	0.27	0.06	0.13	-	-	-	-	-	0.39	0.15	0.38	0.13	0.89
		GSpeaker (Ours)	0.29	0.04	0.15	0.01	0.53	0.69	0.59	0.01	0.89	0.36	0.29	0.14	**0.90**
		GUnique (Ours)	0.29	0.04	0.15	**0.02**	0.52	0.68	0.61	0.01	0.89	0.34	0.28	0.13	0.89
		Vect-GSpeaker (Ours)	0.29	0.04	0.14	0.01	0.56	0.70	0.63	0.01	**0.90**	0.36	0.30	0.15	0.89
		Vect-GUnique (Ours)	0.28	0.04	0.15	0.01	0.53	0.70	0.62	0.01	0.88	0.34	0.28	0.13	**0.90**
Abstractive	**Baselines**	PlanSum	0.25	0.06	0.14	-	-	-	-	-	0.21	0.13	0.32	0.07	0.37
		Llama 7b Instruct	**0.39**	**0.09**	**0.18**	-	-	-	-	-	0.63	0.25	0.49	0.23	0.79
		GLIMPSE-Speaker [1]	0.33	0.07	0.15	-	-	-	-	-	0.44	0.27	0.53	0.32	**0.90**
		GLIMPSE-Unique [1]	0.34	0.07	0.16	-	-	-	-	-	0.44	0.27	0.54	0.33	0.84
	BART	GSpeaker (Ours)	0.30	0.06	0.15	**0.02**	0.46	0.73	0.71	0.01	0.76	**0.43**	0.35	**0.42**	0.84
		GUnique (Ours)	0.29	0.06	0.15	0.01	0.46	0.72	0.70	0.01	0.76	0.41	0.33	0.41	0.68
		Vect-GSpeaker (Ours)	0.28	0.06	0.14	-0.01	0.49	0.74	0.71	0.01	0.75	0.42	0.33	0.39	0.76
		Vect-GUnique (Ours)	0.29	0.06	0.14	0.01	0.45	0.71	0.71	0.01	0.73	0.41	0.33	0.40	0.69
	Pegasus Arxiv	GSpeaker (Ours)	0.17	0.04	0.10	-0.13	0.55	0.67	**0.72**	0.01	0.64	0.37	0.28	0.35	0.57
		GUnique (Ours)	0.16	0.04	0.11	-0.13	0.59	0.68	**0.72**	0.01	0.62	0.33	0.30	0.31	0.46
		Vect-GSpeaker (Ours)	0.15	0.04	0.01	-0.14	0.61	0.65	0.71	0.01	0.61	0.33	0.29	0.29	0.52
		Vect-GUnique (Ours)	0.16	0.04	0.01	-0.14	0.54	0.65	0.71	0.01	0.61	0.34	0.31	0.33	0.49
	Pegasus Large	GSpeaker (Ours)	0.19	0.03	0.11	-0.13	0.50	**0.75**	0.66	0.01	0.87	0.27	0.27	0.15	0.13
		GUnique (Ours)	0.18	0.03	0.10	-0.13	0.54	**0.75**	0.66	0.03	0.86	0.26	0.28	0.16	0.10
		Vect-GSpeaker (Ours)	0.16	0.03	0.10	-0.16	**0.65**	**0.75**	0.66	0.01	0.84	0.25	0.29	0.14	0.10
		Vect-GUnique (Ours)	0.18	0.03	0.11	-0.14	0.52	0.74	0.65	0.01	0.85	0.26	0.28	0.14	0.10

TABLE I

COMPARISON TO METAREVIEW MOTIVATIONS USING ROUGE SCORES AND ESTIMATED HUMAN JUDGMENT USING THE SEAHORSE METRICS FOR ALL BASELINES AND OUR TEMPLATED SUMMARIES COMPARED AGAINST EACH DOCUMENT INDEPENDENTLY. COV. STANDS FOR MAIN IDEAS, ATTR. FOR ATTRIBUTION, GRAM. FOR GRAMMAR, COMPR. FOR COMPREHENSIBLE, CONC. FOR CONCISENESS, AND REPT. FOR REPETITION. FOR UNIEVAL SCORES, CONSIS. STAND FOR CONSISTENCY, COHERENCE AND FLUENCY ARE ALSO REPORTED.

approaches rather than directly competing alternatives for addressing the summarization task. The distinction between their performance is subtle, as both methods demonstrate strengths in different aspects. The primary variation in their outcomes stems from differences in the candidate generation process, which significantly influences the final summaries. This suggests that the choice between the two extensions depends largely on the summarization objectives and the characteristics of the generated candidates rather than a clear superiority of one approach over the other.

Furthermore, we conducted an ablation study to assess the validity of the RSA framework within the MDS context. Specifically, we generated summaries by evaluating candidate likelihoods based on language model (LM) perplexity, aggregated from three models as described in Section III-A. We then compared these results with those obtained using GUnique across all four datasets.

Our findings indicate that the metric scores between the two approaches are almost identical, with the most noticeable differences observed in abstractive candidates generated by Pegasus Arxiv, as shown in Figure A1. While GUnique consistently outperforms the LM perplexity-based approach, the improvement is marginal rather than substantial. These results suggest that, in the context of an LLM ensemble, the RSA-based refinement introduces only limited enhancements, highlighting the need for further investigation into its role in improving summarization quality.

VI. CONCLUSION

In this work, we proposed two extensions to the GLIMPSE summarization framework, introducing a multi-model **Rational Speech Act (RSA)** approach for **multi-document summarization (MDS)** in the context of scientific peer reviews. Our first extension, **Model Ensemble** for RSA, aggregates likelihood scores from multiple language models before applying RSA, ensuring robustness by leveraging diverse model perspectives. The second extension, **Vectorized RSA**, preserves the unique contributions of each model throughout the pipeline by maintaining a vectorized representation before final aggregation.

Our experimental results indicate that both extensions are **comparable alternatives** rather than directly competing approaches. While **Extension 1** offers improved consistency in candidate selection, **Extension 2** enhances informativeness by retaining the distinct characteristics of multiple models. The **trade-off** in performance highlights the influence of the candidate generation process, choosing between the two extensions depends on specific summarization goals.

In terms of **evaluation**, our approaches demonstrate improvements over the baseline in key metrics such as attribution, conciseness, and informativeness, as assessed using SEAHORSE and UniEval.

Future work could explore **adaptive weighting mechanisms** for combining model outputs, as well as incorporating **reinforcement learning** techniques to dynamically optimize summary generation. Further research into **controlling** factual consistency and **reducing** grammatical inconsistencies could also enhance the practical application of these approaches in real-world peer review summarization.

REFERENCES

[1] M. Darrin, I. Arous, P. Piantanida, and J. C. Cheung, "Glimpse: Pragmatically informative multi-document summarization for scholarly reviews," *arXiv preprint arXiv:2406.07359*, 2024.

[2] C. Ma, W. E. Zhang, M. Guo, H. Wang, and Q. Z. Sheng, "Multi-document summarization via deep learning techniques: A survey," *ACM Computing Surveys*, vol. 55, no. 5, pp. 1–37, 2022.

[3] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," *arXiv preprint arXiv:2004.05150*, 2020.
[4] M. Zaheer, G. Guruganesh, K. A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang *et al.*, "Big bird: Transformers for longer sequences," *Advances in neural information processing systems*, vol. 33, pp. 17 283–17 297, 2020.
[5] W. Yuan, P. Liu, and G. Neubig, "Can we automate scientific reviewing?" *Journal of Artificial Intelligence Research*, vol. 75, pp. 171–212, 2022.
[6] M. C. Frank and N. D. Goodman, "Predicting pragmatic reasoning in language games," *Science*, vol. 336, no. 6084, pp. 998–998, 2012.
[7] M. Lewis, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.
[8] J. Zhang, Y. Zhao, M. Saleh, and P. Liu, "Pegasus: Pre-training with extracted gap-sentences for abstractive summarization," in *International conference on machine learning*. PMLR, 2020, pp. 11 328–11 339.
[9] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," *arXiv preprint arXiv:1904.09675*, 2019.
[10] M. Zhong, Y. Liu, D. Yin, Y. Mao, Y. Jiao, P. Liu, C. Zhu, H. Ji, and J. Han, "Towards a unified multi-dimensional evaluator for text generation," *arXiv preprint arXiv:2210.07197*, 2022.
[11] E. Clark, S. Rijhwani, S. Gehrmann, J. Maynez, R. Aharoni, V. Nikolaev, T. Sellam, A. Siddhant, D. Das, and A. P. Parikh, "Seahorse: A multilingual, multifaceted dataset for summarization evaluation," *arXiv preprint arXiv:2305.13194*, 2023.

APPENDIX

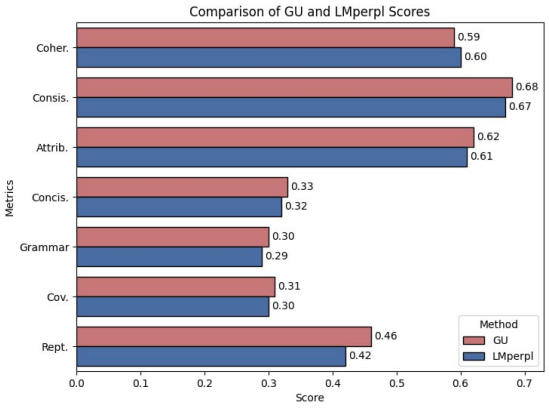

Fig. A1. Comparison of GUnique and LMPerplexity scores among different metrics

#	Model Name
1	facebook/bart-cnn-large
2	Falconsai/text_summarization
3	google/pegasus-arxiv
4	google/bigbird-pegasus-large-arxiv
5	google/pegasus-large
6	google/pegasus-xsum
7	pszemraj/led-base-book-summary
8	pszemraj/led-large-book-summary
9	mrm8488/flan-t5-large-finetuned-openai-summarize_from_feedback
10	bart-large-xsum-samsum

TABLE A1
LIST OF SELECTED MODELS FOR SUMMARIZATION

Metric	**Model**
Attribution	LED-Large
Main Ideas	Pegasus-Large
Conciseness	Pegasus-BigBird
Consistency	Pegasus-XSUM
Fluency	Flan-T5

TABLE A3
METRIC-MODEL PAIRS

Direct	**Type**	**Selections**				
		Top 1	**Top 2**	**Top 3**	**Top 4**	**Top 5**
Extractive	GSpeaker	FlanT5	Pegasus BigBird	LED-Large	Pegasus-Large	BART-XSUM
	GUnique	BART-XSUM	Flan-T5	LED-Large	Falcon	Pegasus-Arxiv
BART	GSpeaker	BART-XSUM	Pegasus-Large	Pegasus-XSUM	Pegasus-BigBird	LED-Base
	GUnique	Pegasus-Large	LED-Large	Pegasus-XSUM	FlanT5	Pegasus-Arxiv
Pegasus-Large	GSpeaker	LED-Large	Pegasus-Large	Pegasus-BigBird	BART-XSUM	Falcon
	GUnique	BART-XSUM	Falcon	LED-Large	Pegasus-BigBird	Pegasus-Large
Pegasus-Arxiv	GSpeaker	BART-XSUM	Pegasus-Large	Pegasus-BigBird	Falcon	LED-Base
	GUnique	Falcon	LED-Base	Pegasus-BigBird	Pegasus-XSUM	BART-XSUM

TABLE A2
TOP MODELS FOR EACH DATASET WITHIN TWO DIFFERENT SETTINGS.