GLIMPSE: Pragmatically Informative Multi-Document Summarization of Scholarly Reviews

*Maxime DARRIN[1,2,3,4] *Ines AROUS[2,3]
Pablo PIANTANIDA[1,2,4,5] Jackie Chi Kit CHEUNG[2,3,6]
[1]International Laboratory on Learning Systems, [2]MILA - Quebec AI Institute
[3]McGill University [4]Université Paris-Saclay
[5]CNRS, CentraleSupélec, [6]Canada CIFAR AI Chair
maxime.darrin@mila.quebec ines.arous@mila.quebec
pablo.piantanida@mila.quebec jackie.cheung@mcgill.ca

Abstract

Scientific peer review is essential for the quality of academic publications. However, the increasing number of paper submissions to conferences has strained the reviewing process. This surge poses a burden on area chairs who have to carefully read an ever-growing volume of reviews and discern each reviewer's main arguments as part of their decision process. In this paper, we introduce GLIMPSE, a summarization method designed to offer a concise yet comprehensive overview of scholarly reviews. Unlike traditional consensus-based methods, GLIMPSE extracts both common and unique opinions from the reviews. We introduce novel uniqueness scores based on the Rational Speech Act framework to identify relevant sentences in the reviews. Our method aims to provide a pragmatic glimpse into all reviews, offering a balanced perspective on their opinions. Our experimental results with both automatic metrics and human evaluation show that GLIMPSE generates more discriminative summaries than baseline methods in terms of human evaluation while achieving comparable performance with these methods in terms of automatic metrics.

1 Introduction

Peer review is the standard process for evaluating researchers' work submitted to conferences or academic journals across all fields. Its primary function is to maintain quality standards for academic publications and provide authors with constructive feedback on their work. Its effectiveness is currently being challenged by a significant surge in the number of submissions. Conferences in computer science, such as the International Conference on Learning Representations (ICLR) and the Association for Computational Linguistics (ACL), among others, regularly receive thousands of submissions. For instance, the number of submissions to ICLR and ACL has increased fivefold since 2017, reaching 3407 and 3378 submissions, respectively, in 2022.

The reviewing process is inherently a tool for scientific communication with at least two target audiences: the authors of a paper and the area chair. The former should get feedback on their work, whereas the latter has to synthesize the salient points from all the reviews as part of their decision process. The area chair has to extract from the reviews the overall sentiment about the paper while gathering the common ideas and unique arguments raised by the reviewers. An efficient highlighting mechanism could help area chairs in their decision-making process, thus reducing their workload.

Several methods have been developed to generate summaries from reviews in various domains (Bražinskas et al., 2020; Chu and Liu, 2019). Many of these methods identify salient segments in reviews based on the discussed topics and the sentiment polarity (Zhao and Chaturvedi, 2020; Li et al., 2023a; Amplayo et al., 2021a), or using centrality-based metrics (Ge et al., 2023; Liang et al., 2021). However, these techniques fall short in the peer review domain. Indeed, they are designed to generate a *consensus* opinion summary, reflecting common opinions without identifying divergent and unique ones. This poses a challenge for the peer review domain since area chairs are concerned with both common and divergent opinions among reviewers.

In this paper, we recognize one of the underlying communication goals in the reviewing/meta-reviewing process: to convey the review's main points to the area chair concisely. The distillation of such salient information in a concise message has been a long-standing focus of study within the pragmatics domain. One of the most influential probabilistic approaches to pragmatics is the **Rational Speech Act** (RSA). It formulates the communication problem akin to a "reference game", with the

*Equal contribution.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 12737–12752

goal of associating each item in a set with the most informative yet concise utterance that distinguishes it from others.

We take inspiration from this reference-game scenario to introduce the discriminative summarization task. The goal is to generate a summary for each review that highlights commonalities, differences, and unique perspectives, thereby distinguishing one review from others of the same submission. This framework closely aligns with the challenge of summarizing academic reviews, distilling key discriminative insights from lengthy reviews, thus contextualizing each review in relation to others. To this end, we map the discriminative summarization task to a reference game and propose **GLIMPSE, a novel pragmatically informative summarization method for scholarly reviews.** At the technical level, we leverage the RSA model, a framework for pragmatic modeling rooted in Bayesian inference that solves the reference game setting (Frank and Goodman, 2012). We define two novel RSA-based scores that measure the *informativeness* and the *uniqueness* of opinions in scholarly reviews. We use these scores to rank utterances describing a review and aggregate them to compose a "glimpse" of all reviews.

We conducted extensive experiments on a real-world peer review dataset from the ICLR conference collected over a four-year time period. We compare GLIMPSE performance to state-of-the-art methods in multi-document summarization. We design extractive and abstractive variants of our framework and shed light on their properties. Our results show that GLIMPSE generates informative and concise summaries. To the best of our knowledge, we are the first to cast the multi-document summarization problem as a reference game and adopt RSA to identify common and divergent opinions in reviews.

Contributions: Overall, we make the following key contributions:

1. We propose a new setting for multi-document summarization: discriminative summarization and cast it as a reference game problem.
2. We propose new *informativeness* and *uniqueness* scores for multi-document summarization based on the RSA model of human communication.
3. We conduct empirical evaluation to demonstrate that we extract more discriminative summaries than consensus-based summarizers in both automatic and human evaluation.

2 Related Work

Unsupervised Opinion Summarization. Our work is related to unsupervised opinion summarization, where several methods have been developed to summarize reviews of products or hotels. These methods can be divided into two categories: abstractive and extractive. Abstractive summarization aims to generate a coherent summary that reflects salient opinions in input reviews. For instance, MeanSum (Chu and Liu, 2019) relies on an auto-encoder architecture, in which an aggregated representation of the input reviews is fed to a decoder that generates review-like summaries. Another approach consists of using a hierarchical variational autoencoder model (Bražinskas et al., 2020) to generate summaries that represent dominant opinions within reviews. Other methods rely on modeling fine-grained information within reviews. This includes disentangling aspects and sentiments (Amplayo et al., 2021b; Wang and Wan, 2021; Suhara et al., 2020; Ke et al., 2022), topics (Xu and Lapata, 2020) and contrastive opinions (Iso et al., 2022; Carenini and Cheung, 2008). The absence of attributability in these methods presents a challenge in the peer review domain due to the need for transparency and context. In the peer review process, area chairs might seek to trace the source of the synthesized opinions to assess their validity and relevance. **Our method addresses the challenge of attributability** by associating a summary with each review, ensuring transparency and allowing area chairs to trace and understand the synthesized opinions in context.

In contrast to abstractive summarization, which entails generating new text, extractive summarization aims to extract significant phrases directly from the input reviews. A fundamental method in extractive summarization involves clustering review segments and iteratively extracting the central elements of these segments to form a summary. Various techniques depend on a model to acquire representations for review sentences. These representations are then utilized by an inference algorithm, often in an encoder-decoder architecture, to choose sentences for summarization (Li et al., 2023a; Amplayo et al., 2021a; Angelidis et al., 2021; Gu et al., 2022; Angelidis and Lapata, 2018). Current extractive methods concentrate on amalgamating common opinions found across reviews. However, we contend that such common opinions hold less significance in scholarly contexts. Instead,

Solid theoretical results are provided to confirm the doubly robustness of the treatment estimator and outcome estimator. This paper is well-written, however, I still have some concerns about the contributions. [...] Why the VCnet designed to be dependent of treatment information? The dependence is not theoretically discussed in this paper. The experimental results are not convincing for me. Only two baselines are compared with the proposed method. [...] But the design of VCnet needs the numerical results to confirm its effectiveness for ADRF.

This paper is to develop a varying coefficient neural network to estimate average dose-response curve (ADRF). Although this paper has several interesting results, the paper is full of many typos and small errors. The current paper needs substantial improvement. The introduction section is not well written since the logic does not flow very smooth.. In the proof of all theorems, there are some obvious mistakes inside.

This problem is well-motivated - estimating dose-response is a challenging and practically important problem. The paper is well written. It explained complex ideas in the semi- parametric literature clearly. The comparison against existing works is clear. The theory, as far as I can tell, is solid. It improves the existing results in targeted regularization and can be adapted to analyze one step TMLE.

Figure 1: Illustration of our proposed RSA-based scores applied to real-world scholarly reviews. We consider each sentence in a review as a candidate summary. The most common opinions in the reviews are highlighted in blue whereas the unique ones are highlighted in red using our RSA-based scores.

we advocate for a pragmatic approach that situates each review relative to others, emphasizing unique, common, and divergent ideas (Mani and Bloedorn, 1999; Wan et al., 2011).

Summarization in Scientific Peer Review The task most closely related to ours in the scholarly domain is the meta-review generation task. Various strategies have been developed for this task. For instance, MetaGen (Bhatia et al., 2020) generates an extractive summary draft of reviews, then uses a fine-tuned model for decision prediction, and generates an abstractive meta-review based on both the draft and the predicted decision. Similarly, (Li et al., 2023b) leverage the hierarchical relationships within reviews and metadata, such as reviewers' confidence and rating, to generate a meta-review along with the acceptance decision. Recently, (Zeng et al., 2023) proposed to prompt a large language model in a guided and iterative manner to generate a meta-review. Our task is different because our goal is to support area chairs by generating a summary of reviewers' opinions and highlighting areas of divergence. Predicting the meta-reviewer's decision and generating a meta-review based on the predicted decision is out of the scope of our work as we believe it is not desirable to do so. The metareview evaluation should ultimately remain in the hands of a human expert as it involves scientific expertise and reasoning about the evaluated paper.

Rational Speech Act (RSA) framework. The Rational Speech Act theory (Frank and Goodman, 2012) is the most influential probabilistic approach to pragmatics (Qing and Franke, 2015; Degen, 2023). It formulates communication between a pragmatic speaker and a listener as probabilistic reasoning, where the speaker's goal is to choose the utterance that is both short and informative with respect to the speaker's intended referent (Degen, 2023). As a result, the RSA framework aims to effectively identify the most informative utterance from multiple potential options in a given context. It has been applied to various tasks, including image captioning (Ou et al., 2023; Cohn-Gordon et al., 2018), translation (Cohn-Gordon and Goodman, 2019), dialogue (Kim et al., 2020, 2021), and text generation (Shen et al., 2019), among others. However, there is little research on applying the RSA framework to multi-document summarization. In our study, we tailor the RSA framework for multi-document summarization tasks, aiming to produce summaries that capture various perspectives while minimizing redundancy.

3 Discriminative Multi-Document Summarization (D-MDS)

Most approaches to multi-document summarization prioritize generating consensus-based summaries by emphasizing redundant opinions across reviews. However, in domains like peer review, where including unique and divergent opinions is crucial for a comprehensive summary, the conventional focus on common opinions becomes less relevant. This shift in emphasis is particularly pertinent in scholarly review summarization, where the objective is to convey the various opinions and distinctive viewpoints expressed by reviewers. With this objective in mind, we introduce the discriminative

multi-document summarization task, inspired by the reference game setting. **Our aim is to furnish the meta-reviewer with a summary for each review, enabling them to swiftly identify the source review based on the summary content.**

Formally, we define the discriminative multi-document summarization problem as follows:

Definition 1 (The discriminative multi-document summarization problem)**.** Let $\mathcal{N} = \{d_1, \cdots, d_N\}$ be a set of documents. For each document d_i, we suppose we have K candidate summaries and we form $\mathcal{K} = \{s_{i,j}\}_{1 \leqslant i \leqslant N, 1 \leqslant j \leqslant K}$ the set of all candidates. The goal is to select the most informative summary from that set for each document $d_i \in \mathcal{N}$

These candidates can be generated using various summarization methods (see Sec. 5.1).

4 Problem Formulation and Pragmatic Summarization

In this section, we formulate the discriminative multi-document summarization problem as a reference game. We then apply the Rational Speech Act (RSA) framework, a probabilistic approach to pragmatics that tackles reference games, to address this summarization problem.

4.1 D-MDS Problem as a Reference Game

In a reference game setting (Frank and Goodman, 2012), a speaker and a listener are given a set of objects. The speaker provides a description of a target among the set of objects, and a listener aims to select the correct target given the speaker's description. The speaker's goal is to describe the target using one of its properties in an informative yet concise manner. Similarly, in discriminative multi-document summarization, our goal is to develop a summarizer that provides a concise yet informative description of each review within a set of reviews. Given this similarity, we formally map the D-MDS problem to a reference game setting as follows.

Definition 2 (D-MDS as a reference game)**.** Let $\mathcal{O} \triangleq \{d_1, \cdots, d_N\}$ be the set of documents and $\mathcal{C} \triangleq \{s_{i,j}\}_{1 \leqslant i \leqslant N, 1 \leqslant j \leqslant K}$ the set of candidate summaries. Let $M : \mathcal{O} \times \mathcal{C} \rightarrow [0, 1]$ be a truth matrix that indicates the likelihood of a candidate summary s being a summary of a document d. In standard reference games, the truth-matrix is boolean as the properties for an object are either true or false. In our setting, we approximate the truth-matrix using a pre-trained language model on summarization LM to score the likelihood of each candidate summary s to be associated with a document d: $M(t, s) \approx \mathrm{LM}(s|d)$.

4.2 The Pragmatic Summarizer

One popular framework to efficiently tackle reference games is the Rational Speech Act (RSA) model of human communication Frank and Goodman (2012), which models optimal communication between pragmatic agents and provides a formal framework to build pragmatic speakers and listeners. Therefore, we propose two novel RSA-based scores to select informative and unique opinions for review summarization, which we discuss in detail in Sec. 5.2. In what follows, we present how we adapt RSA to our summarization setting.

The RSA framework posits that both the speaker and listener maximize the utility of their communication and they both assume the other is rational. They iteratively adapt to each other to reach a common understanding of the context. Formally, we define the utility of communicating to a listener L, the summary s for a document d as:

$$V_L(s, d) = \log L(d|s) - \mathrm{Cost}(d), \qquad (1)$$

where $L(d|s)$ denotes the conditional probability of guessing the object d — the document/review in our case — upon receiving s — a short summary— according to the listener L, $\mathrm{Cost}(s)$ is the cost of transmitting the summary s. In most cases, the cost is assumed to be 0, and we measure the informativeness, which is defined with the conditional probability $L(d|s)$.

Pragmatic reasoning is modeled as an iterative process that starts from a *literal* listener, *i.e.* a listener who has no assumption about the speaker. We denote it with $L_0(d|s)$ and define it using a pre-trained language model LM:

Definition 3 (Literal listener)**.**

$$L_0(d|s) = \frac{\mathrm{LM}(s|d)}{\sum_{d \in \mathcal{O}} \mathrm{LM}(s|d)}. \qquad (2)$$

The pragmatic speaker adapts to a listener under the cost constraint by maximizing the communication utility (Zaslavsky et al., 2021). We define it as follows:

Definition 4 (Pragmatic speaker)**.** For a given target document d, the speaker's distribution over the summaries is defined as the softmax of the utility scores with a listener L_{t-1}, where L_0 is the literal

listener given in Eq. 2.

$$S_t(s|d) = \frac{\exp(V_{L_{t-1}}(d, s))}{\sum_{s'} \exp(V_{L_{t-1}}(d, s')}. \quad (3)$$

The pragmatic listener evaluates the summary provided by the pragmatic speaker and identifies the document that is most likely to be its source.

Definition 5 (Pragmatic listener)**.** For a given target summary s, the pragmatic listener's distribution over the documents is defined using the pragmatic speaker S_t from Eq. 3 as follows:

$$L_t(d|s) = \frac{S_t(s|d)}{\sum_{d'} S_t(s|d')}. \quad (4)$$

We iterate between the pragmatic speaker (Eq. 3) and the listener (Eq. 4), recursively adapting the listener to the speaker for an empirically defined number T of iterations.

5 GLIMPSE Framework

Our framework comprises three steps: 1) generating candidates through either abstractive or extractive summarization techniques, 2) selecting candidates using RSA-based scores defined in Sec. 5.2, and 3) composing a summary using the selected candidates and a template.

5.1 Candidate Generation

The candidate summaries can be extracted sentences from the reviews, abstractive summaries sampled from a summarization model conditioned to review, or any other generated summary. We consider both the extractive setting, where the candidates are extracted sentences from the reviews, and the abstractive setting, where a language model generates the candidate summaries. We specifically focus on the extractive setting, as we want to ensure attribution of the summary to the source review.

5.2 Informativeness and Uniqueness Scores

We propose two RSA-based scores derived from the RSA speaker and listener in our setting Sec. 4.2: a *pragmatic-speaker-based score* and a *uniqueness score*. The *pragmatic-speaker-based score* identifies the most discriminative utterance to refer to a source document. The *uniqueness score* measures the extent to which a candidate summary represents a common — or unique — idea in the source documents.

Pragmatic-speaker-based score. We directly score the summaries for a document according to the pragmatic speaker's distribution. In our case, this process assesses the informativeness of a summary within the set of candidates. Hence, we define the *pragmatic-speaker-based score* to be the argmax of the speaker: Eq. 3:

$$\text{RSA-Speaker}(d) \triangleq \arg\max_{s \in \mathcal{C}} S_t(s|d). \quad (5)$$

In our experiments, we refer to the summarizer leveraging this score as GLIMPSE-Speaker.

Uniqueness score. In our context, the RSA listener is designed to pragmatically infer a document given a summary. We propose quantifying the listener's uncertainty regarding a particular summary given the following intuition: if the listener is uncertain, then the summary could apply to many documents; conversely, if the listener is confident, the summary can only be associated with a single document.

Example 1. Consider the following reviews.

- Review 1: *This paper is well-written. However, the theoretical part lacks clarification.*
- Review 2: *This paper is well-written. I believe it should be accepted.*

Given the sentence "*The paper is well-written,*" a listener cannot accurately deduce the source review. However, when provided with the sentence "*I believe it should be accepted,*" the listener can easily identify that the source review is Review 2.

We propose to measure this uncertainty by comparing the listener's probability distribution defined in Eq. 4 with the uniform distribution $\mathcal{U}$ over the documents:

$$\text{Unique} \triangleq D_{\text{KL}}(L(\cdot|\ s) \| \mathcal{U}). \quad (6)$$

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. Intuitively, it measures how far the listener distribution conditioned to a summary is far from the uniform distribution. The GLIMPSE variant using this score is referred to as GLIMPSE-Unique.

These two scores can be used to identify unique or common opinions across the different documents, as shown in Figure 1, or to select the most informative summary among a set of candidates. While, in this work, we use the scores to compose overall summaries, they can be used as standalone tools for content highlighting.

5.3 GLIMPSE for Standard Multi-Document Summarization (MDS)

The standard Multi-Document Summarization (MDS) task aims to generate a single summary for multiple documents. Generally, this summary is built to reflect the consensus among the documents. In order to provide a comparison with previous work, we generate summaries by concatenating the three most common with the three most unique candidate summaries identified through our RSA-based scores defined in Eq. 5 and Eq. 6. The most unique candidates are selected either using the RSA speaker score (Eq. 5) or using the uniqueness score (Eq. 6). Summaries utilizing the former are referred to as GLIMPSE-Speaker, while those utilizing the latter are referred to as GLIMPSE-Unique. This method ensures a balanced representation of both common and unique opinions.

Example 2. Given the reviews in Figure 1, we can compose two summaries using our simple template:

- **GLIMPSE-Speaker**: This paper is well-written, however, I still have some concerns about the contributions. The experimental results are not convincing I really enjoyed reading it. The introduction section is not well written since the logic does not flow very smooth. Why the VCnet designed to be dependent of treatment information?

- **GLIMPSE-Unique**: This paper is well-written, however, I still have some concerns about the contributions. The experimental results are not convincing I really enjoyed reading it. The introduction section is not well written since the logic does not flow very smooth. This paper is to develop a varying coefficient neural network to estimate average dose-response curve (ADRF). Why the VCnet designed to be dependent of treatment information?

6 Experimental Setup

In this section, we outline the experimental setup used to evaluate the performance of GLIMPSE. We present the datasets, baselines, evaluation metrics, and implementation details.

Datasets. We collect data from the ICLR conference, which provides open access to reviews and meta-reviews for all submissions through OpenReview[1]. Our dataset contains 28062 reviews for 8428 submissions to the ICLR conference from 2017 until 2021.

Evaluation. The task of collecting reference summaries of scholarly reviews is challenging due to their length and domain specificity. We identify an alternative method to collect scholarly summaries by considering meta-reviews. Meta-reviews convey mainly the decision of the meta-reviewer and, depending on their style, may summarize the reviews. We only use these "summary-like" meta-reviews instead of those that solely convey area chairs' decisions. We apply heuristic rules to extract "summary-like" meta-reviews and perform manual verification. We set criteria for identifying them, including length, reference to at least one reviewer, and semantic similarity measured by cosine similarity with all reviews. We obtained 226 summary-like meta-reviews that we use for evaluation.

6.1 Evaluation Metrics

Evaluation against gold standards. For comprehensiveness and following common practices, we include ROUGE evaluations of the summaries against a gold standard. In our case, we assume that the area chair's motivations for their decision provide a reasonable comparison. However, this evaluation may be limited, because ROUGE scores have a low correlation with human judgments according to various studies (Kryscinski et al., 2020; Kocmi et al., 2021).

Discrimininativeness. Following (Ou et al., 2023; Cohn-Gordon et al., 2018), we measure the discriminativeness of a summary, i.e., whether an evaluative listener can identify the source review based on the summary content. We construct our evaluative listener by comparing the summary and the reviews using cosine similarity of their paraphrase embeddings (Reimers and Gurevych, 2019).

Learned metrics. We evaluated the generated summaries using the SEAHORSE metrics trained on human judgment (Clark et al., 2023). They assess the summaries along six axes: coverage, attribution, comprehensibility, grammar, repetition and conciseness. Coverage and attribution, respectively, measure if the main ideas of the source text are present in the summary and verify the accuracy of information attribution. Comprehensibility and grammar assess the summary's overall fluency and

[1] https://openreview.net/

grammatical correctness, while repetition and conciseness evaluate the absence of redundant utterances and the summary conciseness. In the multi-document summarization setting, we evaluate the summary generated as discussed in Sec. 5.3. High coverage across all reviews suggests that the summary effectively captures the main ideas of all the reviews.

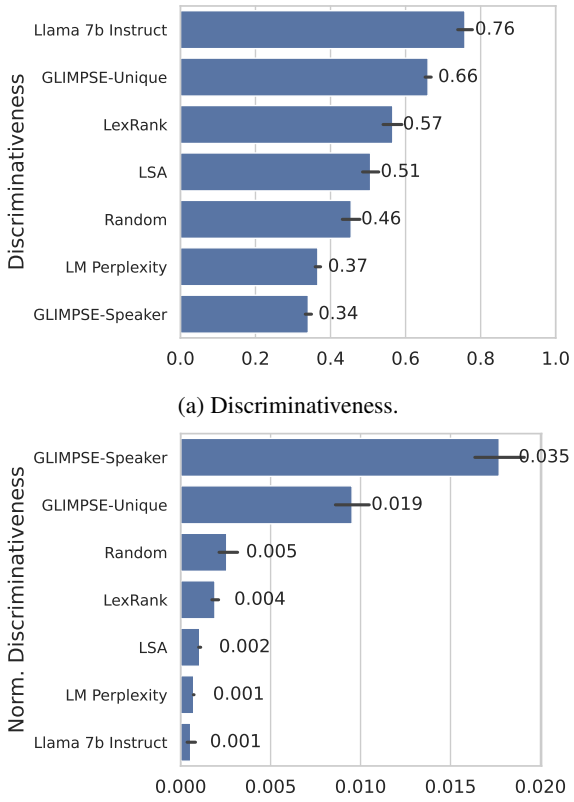
6.2 Comparison Methods

We compare our framework with baseline summarization methods: 1) Latent Semantic Analysis (LSA) (Steinberger and Jezek, 2004) extracts sentences from a document based on latent topics. 2) LexRank (Erkan and Radev, 2004), is a graph algorithm that uses TF-IDF to calculate weights for sentence segments and selects segments near the center as the summary[2]. 3) Quantized Transformer space (QT) proposes a clustering interpretation of the quantized space for extractive summarization (Angelidis et al., 2021). 4) PlanSum (Amplayo and Lapata, 2021) is an abstractive summarization method that incorporates content planning in a summarization model. 5) Convex Aggregation for Opinion Summarization (COOP) (Iso et al., 2021) is an abstractive summarization method that leverages embedding representation to condition summary generation. 6) Llama 7b Instruct[3], is a Large Language Model that generates abstractive summaries via prompting. In addition, we report the performance of the random selection of sentences from the documents. We also compare our method with generative models in terms of discriminativeness by evaluating the perplexity of candidate summaries conditioned on the source text.

Candidate summary generation. We generate candidate summaries by extracting sentences from reviews in the extractive setting. For the abstractive setting, we use pre-trained language models, such as PEGASUS (Zhang et al., 2020) or BART summarizers (Lewis et al., 2020), to generate summaries for each review.

7 Experimental Results

We present the results for both the discriminative and the standard multi-document summarization tasks. We compare our method with the baselines and evaluate the quality of the generated summaries using automatic metrics and a human evaluation.



(a) Discriminativeness.

(b) Discriminativeness per character.

Figure 2: Discriminativeness for all the baselines and our methods in extractive mode (GLIMPSE-Unique (Extr.), GLIMPSE-Speaker (Extr.)), and a strong abstractive method (Llama 7b Instruct).

7.1 Discriminative Summarization

In this setting, our goal is to evaluate the **discriminativeness** of the generated summaries by gauging a listener's ability to identify the source review from the generated summary.

Discriminativeness evaluation. In Figure 2, we report the discriminativeness scores for extractive methods alongside Llama 7b Instruct, a strong abstractive summarization baseline. We find that Llama 7b Instruct achieves the highest performance compared with other methods (76% discriminativeness); followed by GLIMPSE-Unique in the extractive setting (66% discriminativeness). These results suggest that our method achieves high discriminativeness while being more cost-effective compared to recent large language models in terms of memory footprint and runtime. Surprisingly, we find that selecting candidates using the perplexity of common abstractive summarizers, such as BART (LM

[2]LSA and LexRank are implemented using github.com/miso-belica/sumy

[3]https://huggingface.co/togethercomputer/Llama-2-7B-32K-Instruct

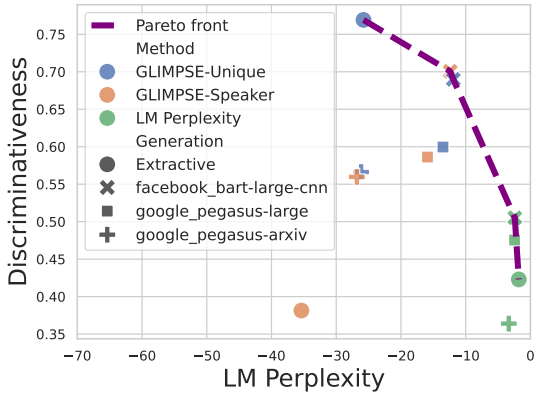


Figure 3: Trade-off between discriminativeness of the generated summaries and their fluency (measured as the log-likehood of the summaries under the generative model). The Pareto frontier shows the best trade-off between the two metrics.

Perplexity), yields results worse than a Random selection (37% vs 46%). Contrary to our expectations, GLIMPSE-Speaker has the lowest performance compared with baseline methods in terms of discriminativeness. We hypothesize that this result can be attributed to a phenomenon known as codebooking, where the RSA speaker overly tailors its output to the RSA listener, disregarding the semantic load of the summary (Arumugam et al., 2022; Wang et al., 2021).

Conciseness and information density. When evaluating properties of the generated summaries, we observed a significant variation in the length of the generated summaries ranging from 76 on average for GLIMPSE-Speaker to around 2000 for Llama. We account for these differences by evaluating the "Discriminativeness per character" in Figure 2b. Interestingly, we find that the advantage of Llama vanishes completely, while GLIMPSE-Speaker and GLIMPSE-Unique notably outperform all baseline methods. This result suggests that GLIMPSE methods produce very succinct yet very discriminative summaries. This is key for the task at hand since the goal is to alleviate the strain on area chairs.

Discriminativenes versus fluency. In Figure 3, we plot the discriminativeness (in terms of discriminativenes) against the language model perplexity (as a proxy to fluency). We highlight the Pareto frontier, *i.e.* the points for which we cannot gain on a given axis without losing on the other. Along this frontier, we observe a trade-off between discriminativeness and fluency (Figure 3). This outcome is expected, as the RSA scores are designed to select less common utterances to improve discriminativeness.

Human evaluation of uniqueness. Since the main goal of the D-MDS task is to highlight the most unique information from each review, we asked human evaluators to identify the source review for each summary (See Appendix A for details about the human evaluation task). We show that GLIMPSE selects more informative and unique ideas than other summarization methods such as LLMs (Llama 7b Instruct) or abstractive multi-document summarizers as illustrated in Tab. 2. These results are consistent with the evaluation reported in Figure 2. They indicate that utterances selected with GLIMPSE-Unique are shorter and more discriminative compared with those selected with baseline methods, suggesting its potential as an effective highlighting mechanism for peer review.

7.2 Overall Summary Quality

Using ROUGE scores, we compare the overlap between the generated summaries and the summary-like metareviews. We also evaluate them using the SEAHORSE metrics. We present the results in Tab. 1.

Comparison to gold standards. Overall, the summaries generated using both our methods and the baselines exhibit minimal overlap with the metareviews. We find that our method achieves comparable performance with baseline methods in terms of ROUGE score. This result is likely influenced by the nature of metareviews, which extends beyond synthesizing reviews to justify a paper's acceptance decision.

Coverage and conciseness. We evaluate the coverage and conciseness of our generated summaries with the baseline methods. We measure the coverage by assessing if the main ideas of a review are present in the summary using the SEAHORSE metrics (cf. Tab. 1). We observe that the summaries generated with GLIMPSE are more concise and yield a significantly better coverage of the main ideas of the documents than baseline methods. This result is consistent with the discriminativeness results presented in (Figure 2b and Sec. 7.1). Using GLIMPSE to select excerpts leads to denser summaries than baseline methods.

Fluency. Similarly to our findings in the discriminative summarization setting, we find that summaries generated by our method tend to be less fluent than those generated by the baselines. We posit that this is due to our method's tendency to

	Method	Rouge-1		Rouge-2		Rouge-L		Cov.		Attr.		Gram.		Conc.		Repet.	
Extractive	Random	0.32	±0.05	0.06	±0.02	0.15	±0.03	0.09	±0.04	0.48	±0.07	0.58	±0.12	0.15	±0.04	0.94	±0.04
	LSA	0.38	±0.06	0.08	±0.03	0.16	±0.03	0.14	±0.09	0.57	±0.09	0.59	±0.15	0.19	±0.07	0.94	±0.07
	LexRank	0.38	±0.06	**0.09**	±0.03	0.17	±0.03	0.18	±0.12	0.56	±0.09	**0.60**	±0.16	0.21	±0.08	**0.95**	±0.06
	QT	0.21	±0.06	0.04	±0.02	0.12	±0.03	0.09	±0.08	0.35	±0.11	0.13	±0.05	0.12	±0.07	0.80	±0.14
	GLIMPSE-Speaker	0.22	±0.06	0.04	±0.02	0.11	±0.03	0.09	±0.05	0.36	±0.09	0.36	±0.13	0.13	±0.05	0.89	±0.08
	GLIMPSE-Unique	0.27	±0.06	0.06	±0.03	0.13	±0.03	0.13	±0.07	0.39	±0.09	0.38	±0.13	0.15	±0.06	0.89	±0.07
Abstractive	PlanSum	0.25	±0.08	0.06	±0.02	0.14	±0.04	0.07	±0.04	0.21	±0.07	0.32	±0.11	0.13	±0.07	0.37	±0.36
	Coop	0.36	±0.05	0.08	±0.02	**0.19**	±0.02	0.08	±0.09	0.26	±0.12	0.51	±0.23	0.09	±0.07	0.26	±0.31
	Llama 7b Instruct	**0.39**	±0.06	0.09	±0.03	0.18	±0.02	0.23	±0.12	**0.63**	±0.11	0.49	±0.16	0.25	±0.09	0.79	±0.26
	GLIMPSE-Speaker	0.33	±0.05	0.07	±0.03	0.15	±0.02	0.32	±0.10	0.44	±0.06	0.53	±0.10	0.27	±0.06	0.90	±0.08
	GLIMPSE-Unique	0.34	±0.04	0.07	±0.02	0.16	±0.02	**0.33**	±0.10	0.44	±0.07	0.54	±0.10	**0.27**	±0.06	0.84	±0.11

Table 1: 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 Repet. for Repetition. The best value in each column is in bold.

Method	Discriminativeness	Avg. Summary Length
GLIMPSE	93.75%	111
Llama	85.18%	1920
MDS	0%	268

Table 2: Accuracy of annotators guessing the review from which a summary is generated for our method and two baselines and the average of summary lengths generated by each method.

select rarer utterances, aiming to enhance discriminativeness, and thus follows the same trade-off observed in Figure 3.

8 Summary and Concluding Remarks

We have introduced a discriminative summarization framework for multi-document summarization and suggested a pragmatic summarization approach inspired by the Rational Speech Act model (RSA) of human communication. Our findings demonstrate the effectiveness of RSA-based scores in capturing unique, common, and divergent opinions among reviewers. This paves the way for the development of tools to aid area chairs in evaluating reviews. Our method yields more informative summaries compared to existing approaches and is suitable for extractive summarization, ensuring the interpretability of the generated summaries.

Acknowledgments

This work was granted access to the HPC resources of IDRIS under the allocation 2023-AD011013290R2 made by GENCI. Ines Arous is supported by a grant from the former Twitter, Inc. In addition, we acknowledge material support from NVIDIA Corporation in the form of computational resources provided to Mila.

9 Limitations & Ethical Considerations

We adopt summarization evaluation methods common in the research area, but the validity of these methods needs further investigation. For example, measuring the overlap between the templated summaries and the metareview motivations is limited by the appropriateness of the motivation as a gold standard and by the use of ROUGE scores, which have well-known limitations. The SEAHORSE trained metrics are also working in a somewhat out-of-distribution setting as they have not been specifically validated on scientific content. Since our method's main strength is to highlight unique and common ideas in a given text to help the reader get the most salient points in context, a natural follow-up work will be to conduct a task-based human evaluation of the highlighting it provides.

Still, extractive summarization methods are notably sensitive to the sentence segmentation process, which can occasionally result in peculiar outcomes. For instance, a brief sentence containing nonsensical content might erroneously emerge as the most informative segment of the summary simply because it is unique among the reviews. Similarly, filler sentences such as "Here are some comments" could be mistakenly identified as the most common ideas in the corpus. Although extractive summarization ensures attribution, it is constrained by these factors, unlike abstractive summarization methods, which, while prone to potential hallucinations, circumvent such pitfalls.

As with any automated tool, its deployment carries inherent risks, particularly if used without due caution in critical decision-making processes, such as determining the acceptance or rejection of scholarly work. Indeed, unexpected biases could lead

to discrimination and unfair presentation of the reviewers' points. It is important to note that our approach is not designed to fully automate decision-making or replace the review summarization process altogether. Rather, we advocate for its use as a supplementary tool, particularly in its extractive configuration, to aid in identifying salient points within reviews.

References

Reinald Kim Amplayo, Stefanos Angelidis, and Mirella Lapata. 2021a. Aspect-Controllable Opinion Summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6578–6593, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Reinald Kim Amplayo, Stefanos Angelidis, and Mirella Lapata. 2021b. Unsupervised Opinion Summarization with Content Planning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12489–12497.

Reinald Kim Amplayo and Mirella Lapata. 2021. Informative and Controllable Opinion Summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2662–2672, Online. Association for Computational Linguistics.

Stefanos Angelidis, Reinald Kim Amplayo, Yoshihiko Suhara, Xiaolan Wang, and Mirella Lapata. 2021. Extractive Opinion Summarization in Quantized Transformer Spaces. *Transactions of the Association for Computational Linguistics*, 9:277–293. Place: Cambridge, MA Publisher: MIT Press.

Stefanos Angelidis and Mirella Lapata. 2018. Summarizing Opinions: Aspect Extraction Meets Sentiment Prediction and They Are Both Weakly Supervised. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3675–3686, Brussels, Belgium. Association for Computational Linguistics.

Dilip Arumugam, Mark K. Ho, Noah D. Goodman, and Benjamin Van Roy. 2022. On Rate-Distortion Theory in Capacity-Limited Cognition & Reinforcement Learning. ArXiv:2210.16877 [cs].

Chaitanya Bhatia, Tribikram Pradhan, and Sukomal Pal. 2020. MetaGen: An academic Meta-review Generation system. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1653–1656, Virtual Event China. ACM.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020. Unsupervised Opinion Summarization as Copycat-Review Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5151–5169, Online. Association for Computational Linguistics.

Giuseppe Carenini and Jackie C. K. Cheung. 2008. Extractive vs. NLG-based Abstractive Summarization of Evaluative Text: The Effect of Corpus Controversiality. In *Proceedings of the Fifth International Natural Language Generation Conference*, pages 33–41, Salt Fork, Ohio, USA. Association for Computational Linguistics.

Eric Chu and Peter Liu. 2019. MeanSum: A Neural Model for Unsupervised Multi-Document Abstractive Summarization. In *Proceedings of the 36th International Conference on Machine Learning*, pages 1223–1232. PMLR. ISSN: 2640-3498.

Elizabeth Clark, Shruti Rijhwani, Sebastian Gehrmann, Joshua Maynez, Roee Aharoni, Vitaly Nikolaev, Thibault Sellam, Aditya Siddhant, Dipanjan Das, and Ankur Parikh. 2023. SEAHORSE: A Multilingual, Multifaceted Dataset for Summarization Evaluation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9397–9413, Singapore. Association for Computational Linguistics.

Reuben Cohn-Gordon and Noah Goodman. 2019. Lost in Machine Translation: A Method to Reduce Meaning Loss. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 437–441, Minneapolis, Minnesota. Association for Computational Linguistics.

Reuben Cohn-Gordon, Noah Goodman, and Christopher Potts. 2018. Pragmatically Informative Image Captioning with Character-Level Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 439–443, New Orleans, Louisiana. Association for Computational Linguistics.

Judith Degen. 2023. The Rational Speech Act Framework. *Annual Review of Linguistics*, 9(1):519–540. _eprint: https://doi.org/10.1146/annurev-linguistics-031220-010811.

G. Erkan and D. R. Radev. 2004. LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. *Journal of Artificial Intelligence Research*, 22:457–479.

Michael C. Frank and Noah D. Goodman. 2012. Predicting Pragmatic Reasoning in Language Games. *Science*, 336(6084):998–998. Publisher: American Association for the Advancement of Science.

Suyu Ge, Jiaxin Huang, Yu Meng, and Jiawei Han. 2023. FineSum: Target-Oriented, Fine-Grained Opinion Summarization. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, WSDM '23, pages 1093–1101, New York, NY, USA. Association for Computing Machinery.

Nianlong Gu, Elliott Ash, and Richard Hahnloser. 2022. MemSum: Extractive Summarization of Long Documents Using Multi-Step Episodic Markov Decision Processes. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6507–6522, Dublin, Ireland. Association for Computational Linguistics.

Hayate Iso, Xiaolan Wang, Stefanos Angelidis, and Yoshihiko Suhara. 2022. Comparative Opinion Summarization via Collaborative Decoding. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3307–3324, Dublin, Ireland. Association for Computational Linguistics.

Hayate Iso, Xiaolan Wang, Yoshihiko Suhara, Stefanos Angelidis, and Wang-Chiew Tan. 2021. Convex Aggregation for Opinion Summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3885–3903, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wenjun Ke, Jinhua Gao, Huawei Shen, and Xueqi Cheng. 2022. ConsistSum: Unsupervised Opinion Summarization with the Consistency of Aspect, Sentiment and Semantic. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, WSDM '22, pages 467–475, New York, NY, USA. Association for Computing Machinery.

Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. 2020. Will I Sound Like Me? Improving Persona Consistency in Dialogues through Pragmatic Self-Consciousness. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 904–916, Online. Association for Computational Linguistics.

Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. 2021. Perspective-taking and Pragmatics for Generating Empathetic Responses Focused on Emotion Causes. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2227–2240, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To Ship or Not to Ship: An Extensive Evaluation of Automatic Metrics for Machine Translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the Factual Consistency of Abstractive Text Summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Haoyuan Li, Somnath Basu Roy Chowdhury, and Snigdha Chaturvedi. 2023a. Aspect-aware Unsupervised Extractive Opinion Summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12662–12678, Toronto, Canada. Association for Computational Linguistics.

Miao Li, Eduard Hovy, and Jey Han Lau. 2023b. Towards Summarizing Multiple Documents with Hierarchical Relationships. ArXiv:2305.01498 [cs].

Xinnian Liang, Shuangzhi Wu, Mu Li, and Zhoujun Li. 2021. Improving Unsupervised Extractive Summarization with Facet-Aware Modeling. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1685–1697, Online. Association for Computational Linguistics.

Inderjeet Mani and Eric Bloedorn. 1999. Summarizing Similarities and Differences Among Related Documents. *Information Retrieval*, 1(1):35–67.

Jiefu Ou, Benno Krojer, and Daniel Fried. 2023. Pragmatic Inference with a CLIP Listener for Contrastive Captioning. ArXiv:2306.08818 [cs].

Ciyang Qing and Michael Franke. 2015. Variations on a Bayesian Theme: Comparing Bayesian Models of Referential Reasoning. In Henk Zeevat and Hans-Christian Schmitz, editors, *Bayesian Natural Language Semantics and Pragmatics*, pages 201–220. Springer International Publishing, Cham.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Sheng Shen, Daniel Fried, Jacob Andreas, and Dan Klein. 2019. Pragmatically Informative Text Generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4060–4067, Minneapolis, Minnesota. Association for Computational Linguistics.

J. Steinberger and Karel Jezek. 2004. Using Latent Semantic Analysis in Text Summarization and Summary Evaluation.

Yoshihiko Suhara, Xiaolan Wang, Stefanos Angelidis, and Wang-Chiew Tan. 2020. OpinionDigest: A Simple Framework for Opinion Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5789–5798, Online. Association for Computational Linguistics.

Xiaojun Wan, Houping Jia, Shanshan Huang, and Jianguo Xiao. 2011. Summarizing the differences in multilingual news. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, SIGIR '11, pages 735–744, New York, NY, USA. Association for Computing Machinery.

Ke Wang and Xiaojun Wan. 2021. TransSum: Translating Aspect and Sentiment Embeddings for Self-Supervised Opinion Summarization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 729–742, Online. Association for Computational Linguistics.

Rose E. Wang, Julia White, Jesse Mu, and Noah D. Goodman. 2021. Calibrate your listeners! Robust communication-based training for pragmatic speakers. ArXiv:2110.05422 [cs].

Yumo Xu and Mirella Lapata. 2020. Coarse-to-Fine Query Focused Multi-Document Summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3632–3645, Online. Association for Computational Linguistics.

Noga Zaslavsky, Jennifer Hu, and Roger P. Levy. 2021. A Rate–Distortion view of human pragmatic reasoning? In *Proceedings of the Society for Computation in Linguistics 2021*, pages 347–348, Online. Association for Computational Linguistics.

Qi Zeng, Mankeerat Sidhu, Hou Pong Chan, Lu Wang, and Heng Ji. 2023. Scientific Opinion Summarization: Meta-review Generation with Checklist-guided Iterative Introspection. ArXiv:2305.14647 [cs].

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *ICML'20*, pages 11328–11339. JMLR.org.

Chao Zhao and Snigdha Chaturvedi. 2020. Weakly-Supervised Opinion Summarization by Leveraging External Information. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9644–9651. Number: 05.

A Human evaluation

We recruited expert human annotators in academia to evaluate the informativeness of the produced summaries (already enrolled PhD and Master students in computer science and machine learning). They were compensated based on their skills and location (Montréal, Quebec) at a rate of 30 CAD per hour.

A.1 Task description

The annotators were presented a generated summary and they had to guess which review was input to the summarizer to produce said summary. They were given 4 choices: one per review and the option to say that it's hard to guess. In Figure 4 we present the instructions given to the annotators and an example of the actual task in Figure 5.

A.2 Human evaluation statistics

We got 8 annotators to perform a total of 89 evaluations total. They spent on average less than an hour on the task. With some dedicated annotators spending almost two hours and performing up to 20 evaluations (Figure 6).

B Impact of the generation method

Abstractive summarization. GLIMPSE can rely on different generation methods (either extractive or different generative summarizers). In Figure 7 we compare the informativeness induced by the different generative methods. It seems that BART produces the most promising pool of candidates.

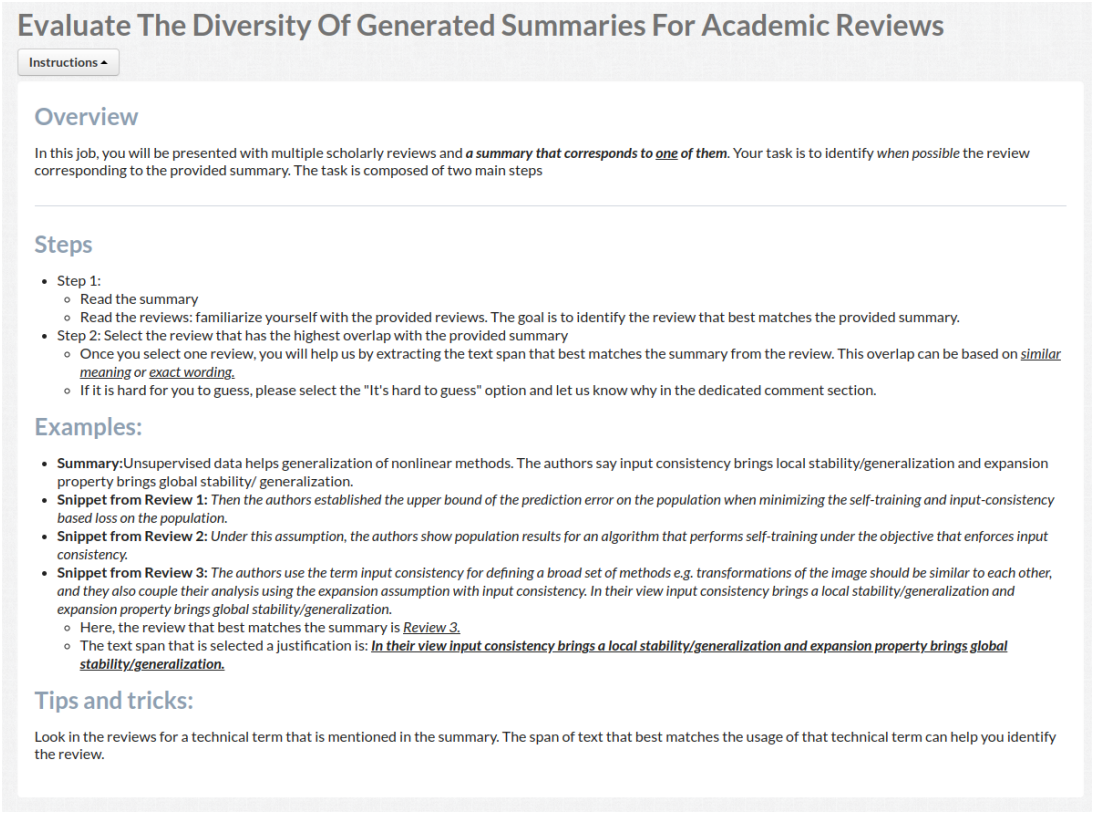


Figure 4: Instructions given to the annotators to perform the discriminative summarization task.

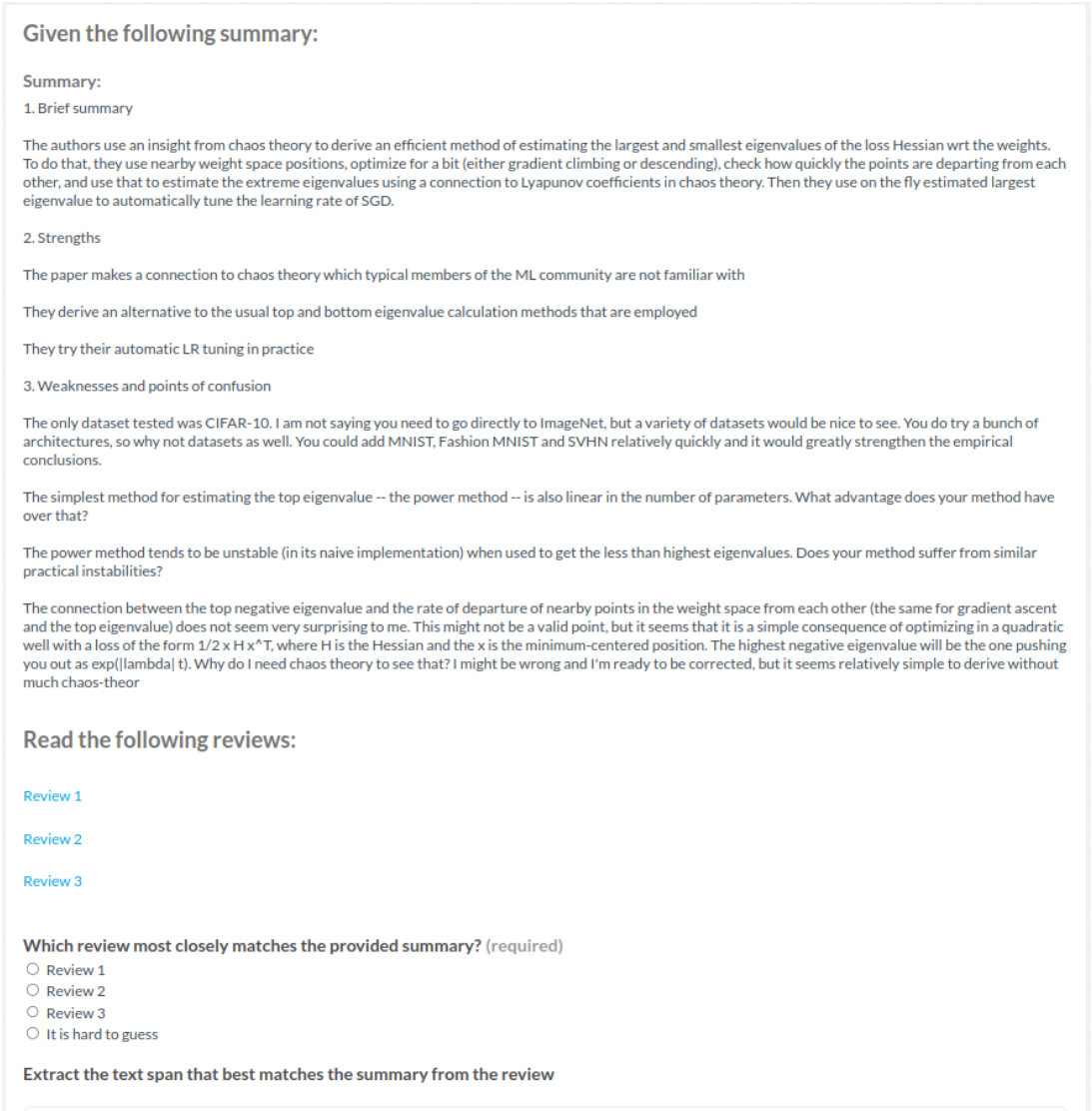
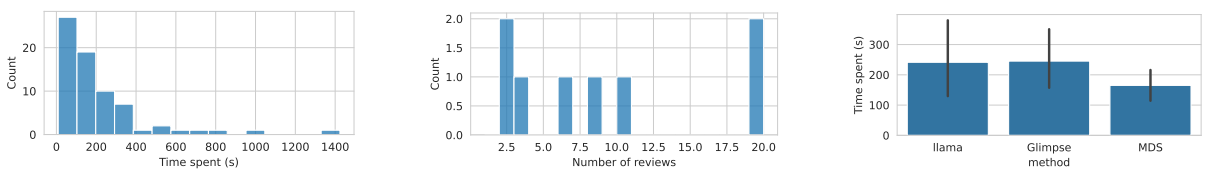


Figure 5: Example of task



(a) Distribution of the time spent by the annotators.

(b) Number of samples examined by each worker.

(c) Time spent by the reviewer per method evaluated.

Figure 6: Statistics of the annotators involved.

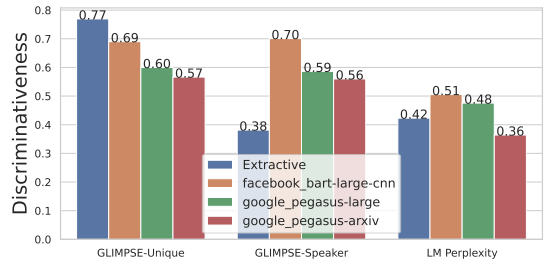


Figure 7: Informativeness of the summaries selected using different scoring methods (GLIMPSE-Unique, GLIMPSE-Speaker) and the baseline LM Perplexity from different set of candidates. Candidates generated by different generative models or simply sentences extracted from the document.