April 29, 2024

A Pragmatic Approach to Non-Generic Summarization

Cesare Spinoso-Di Piano
McGill University
`cesare.spinoso-dipiano@mail.mcgill.ca`

Abstract

There has been growing interest in modeling the task of summarization through a pragmatic point of view by imagining what a reader might understand from a generated summary. This approach to summarization has been taken using the Rational Speech Acts (RSA) framework in which the goodness of a generated summary is determined by how well a reader might reconstruct the initial source document from the generated summary. However, in this paper, we argue that this *source reconstruction* objective is unrealistic and is a symptom of the underspecified nature of *generic* summarization. As a result, we move this pragmatic modeling focus away from generic summarization and towards *non-generic summarization*, the task of generating a summary which meets some additional information request (e.g., providing the answer to a question). To use RSA in this setting, we introduce the *latent reconstruction* objective which rescores candidate summaries based on a reader's ability to reconstruct the value of a latent variable related to the summary's information request. With this more realistic implementation of the intended meaning of a non-generic summary, we are able to achieve competitive ROUGE scores offering a path towards pragmatically modeling non-generic summarization.

Unrealistic is a strong word!

What does "realistic" mean? To make this claim seriously, we might need to cite or do some kind of study involving stakeholders in a deployment context.

1 Introduction

Maybe saying "reflective of imagined downstream use cases" might be more scientifically accurate, though a bit weak.

Automatic summarization is the computational task consisting of distilling the contents of a source document down to its most important parts. This task has seen growing interest in recent years with numerous publications contributing novel, often neural-based, approaches to both extractive and abstractive summarization (Rush et al., 2015; Nallapati et al., 2016; See et al., 2017). In particular, one recent line of work has focused on developing summarization systems which improve the informativeness of generated summaries by imagining what a reader might take away after reading a system-generated summary (Shen et al., 2019). Intuitively, this focus on the reader's interpretation of the generated summary should force the summarization system to act more pragmatically e.g., by providing additional details for surprising or ambiguous information and by omitting details which should be understood from the context.

force -> encourage

More formally, this pragmatic way of viewing summary generation is inspired by the Rational Speech (RSA) framework (Degen, 2023). In RSA, a speaker attempts to convey a piece of information to a listener by generating an utterance which the speaker believes will be easily and unambiguously understood by the listener. To do so, the speaker, often referred to as the pragmatic speaker, will generate an utterance u where the intended meaning of the utterance m is likely to be recovered by the listener, often referred to as the literal listener, when observing u. Translating this to the context of summarization, a pragmatic summarizer will, given a source text x, produce a summary y for which a literal reader is likely to recover the summary's intended meaning m.

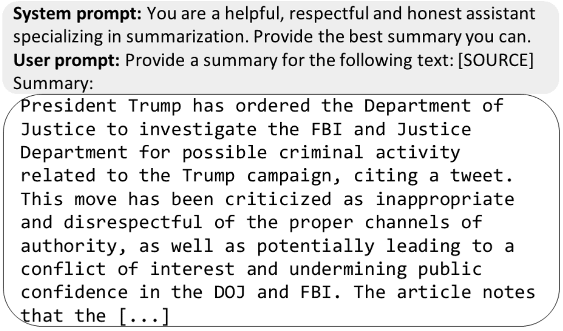
Be clear Degen, 2023 is the overview, not the original paper proposing RSA. RSA is still quite new and not well known in mainstream NLP.

Although appealing in principle, existing approaches which embed the task of automatic summarization within RSA suffer from one common difficulty: implementing the intended meaning m of a summary y. One existing approach approximates the recovery of the intended meaning as a source reconstruction objective (Shen et al., 2019). That is, under the source-reconstruction-based meaning implementation, a good generated summary y should allow a literal listener to easily reconstruct the source document x. Convenient as it may be, we do not believe source reconstruction to be a faithful approximation of recovering the intended meaning of a summary. For instance, a summarization system may correctly choose to drop unimportant information from the source text in the summary thereby making it virtually impos-

Good insight. May be useful to contrast this against the scenarios in which RSA was originally designed, maybe in Related Work section.

May need to tone down the intro overall a little, at least for paper submission purposes.

Including the prompt in these examples makes it unclear what point is intended with these Figures.



System prompt: You are a helpful, respectful and honest assistant specializing in summarization. Provide the best summary you can.
User prompt: Provide a summary for the following text: [SOURCE] Summary:

```
President Trump has ordered the Department of
Justice to investigate the FBI and Justice
Department for possible criminal activity
related to the Trump campaign, citing a tweet.
This move has been criticized as inappropriate
and disrespectful of the proper channels of
authority, as well as potentially leading to a
conflict of interest and undermining public
confidence in the DOJ and FBI. The article notes
that the [...]
```

Figure 1: An example of a *generic summary* (truncated) for this article from the MultiOpEd dataset (Section 4.2) generated by Llama2.

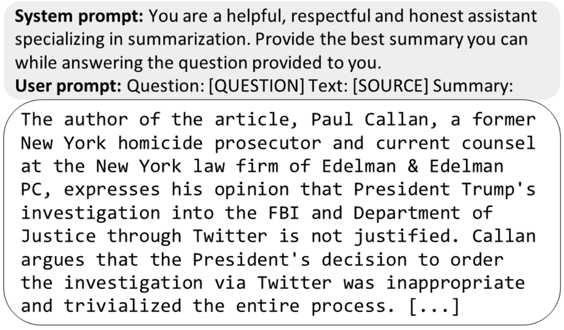
sible to perfectly recover the source text from the generated summary.

In fact, the difficulty of providing a suitable proxy for the intended meaning of a summary may be due to the inherently ill-defined nature of *generic* summarization, the task of generating a summary from a source text without any additional specifications on the output summary (Figure 1). Indeed, as Kryściński et al. (2019) argues, the task of generic summarization is underspecified since what information is included in a summary and, by extension, its intended meaning, depend on unspecified user preferences and information needs. Consequently, expecting a literal reader to recover the intended meaning of a summary which was, at the outset, unobserved by the summarizer may be an unattainable objective. Thus, due to its underspecification, generic summarization may not be an appropriate task for the RSA modeling framework and even, more generally, for a pragmatics-based modeling approach.

This is a bit strong too.

Also, it'd be good to be able to unify this perspective with the perspective of Maxime's summarization evaluation paper.

As a result, in this paper, we re-focus the use of the RSA modeling framework to the task of *non-generic* summarization, the task of generating a summary from a source text while meeting some additional information request. The advantage of this setting over generic summarization is that it specifies an additional input which could be helpful in formulating a more realistic proxy to a summary's intended meaning. For example, a non-generic summary of a news editorial might involve summarizing the article while answering a question related to it as illustrated in Figure 2. In this case, we believe a realistic approximation of the intended meaning of a summary is providing an answer to the initial question about the news editorial. This example motivates the introduction of the *latent reconstruction* objective whereby a good summary should allow the literal reader to easily reconstruct the value of some latent variable Z related to the information request of the non-generic summary.



System prompt: You are a helpful, respectful and honest assistant specializing in summarization. Provide the best summary you can while answering the question provided to you.
User prompt: Question: [QUESTION] Text: [SOURCE] Summary:

```
The author of the article, Paul Callan, a former
New York homicide prosecutor and current counsel
at the New York law firm of Edelman & Edelman
PC, expresses his opinion that President Trump's
investigation into the FBI and Department of
Justice through Twitter is not justified. Callan
argues that the President's decision to order
the investigation via Twitter was inappropriate
and trivialized the entire process. [...]
```

Figure 2: An example of a *non-generic summary* (truncated) generated by Llama2 for the same article as in Figure 1. The question which was asked about this article ([QUESTION]) was "Is Trump's investigation into the FBI justified?"

We evaluate several summarization systems on multiple non-generic summarization datasets and find that the latent reconstruction is a competitive alternative to modeling non-generic summarization. In addition, we find that combining the latent and source reconstruction objectives together leads to performance gains over both reconstruction objectives used individually. This offers a new promising future direction in pragmatically improving non-generic summarization systems by balancing different meaning recovery objectives in the context of RSA.

To summarize, the research question we address in our work is "Is a pragmatics-based approach to non-generic summarization a viable modeling approach?". Or, more specifically, "Can RSA, through the latent reconstruction objective, be used for non-generic summarization?". To this end, the contributions of this paper are as follows:

- We provide a pragmatics-based perspective to modeling the task of non-generic summarization by using the RSA framework and approximating the intended meaning of the summary as the latent reconstruction objective.

- We compare against existing summarization methods and show that using RSA with the latent reconstruction objective offers a promising path towards pragmatically modeling non-generic summarization.

2 Related Work

2.1 Non-generic summarization

Several types of non-generic summarization tasks have already been studied. The most prominent one is the query-focused summarization (QFS) task in which a summary is generated from a single or multiple documents such that it can provide information related to a given query or queries (Dang, 2006; Daumé and Marcu, 2006; Su et al., 2020; Xu and Lapata, 2020). The query can be a collection of keywords, a question or a longer passage describing some information request. In addition to QFS, other forms of non-generic summarization tasks have attempted to tailor the generated summary to some specified user preference. For instance, previous studies have explored personalizing a summary based on a user's preferred summary length, style and interests (Fan et al., 2018), their preferences between lay and technical summaries (Tauchmann et al., 2018; Shaib et al., 2023) as well as their news interests in the context of personalized news headlines generation (Ao et al., 2021).

There have been several proposed modeling approaches for non-generic summarization. In QFS, the two most common approaches have been a two-step retrieval-abstraction method and an end-to-end method (Vig et al., 2022). In other non-generic summarization tasks, studies have attempted to model user preferences either by creating representations for the preferences and injecting them into an encoder-decoder summarization model (Fan et al., 2018; Ao et al., 2021) or, more recently, via prompt engineering and in-context learning (Shaib et al., 2023; Yang et al., 2023a). To the best of our knowledge, there has not been a pragmatics-based approach to non-generic summarization.

2.2 RSA for Language Generation

The RSA framework has been used to model several language generation tasks in the hopes of introducing some of the pragmatics-based notions which make human communication so effective (Grice, 1975). In Andreas and Klein (2016), the authors model the task of informative image captioning by using, for the first time, *neural* listeners and speakers. Cohn-Gordon et al. (2018) offers to solve some of the inefficiencies of using RSA for image captioning by generating captions at the character level. Fried et al. (2018) extend the use of RSA to instruction following and generation and See et al. (2017) investigate its use for generic summarization and for generation from structured data. More recently, Nie et al. (2020) develop an RSA model for issue-sensitive image captioning and Nguyen (2023) try to characterize the behaviour of large language models (LLMs) by framing them in the context of the speaker-listener RSA model of communication.

3 RSA for Non-Generic Summarization

We present our approach to modeling non-generic summarization via RSA by first discussing the general underlying principles of RSA and, subsequently, by discussing the extensions we propose to allow this framework to support the modeling of non-generic summarization.

In the RSA theory for modeling communication, a pragmatic speaker, S_1, attempts to convey a message with some meaning, m, by generating an utterance, u. To do so, the pragmatic speaker uses a *literal speaker*, S_0, to first generate candidate utterances $\mathcal{U}$ where the utterances $u \in \mathcal{U}$ are generated from a uniform distribution, $P_{S_0}(U|M = m)$, over utterances with literal meaning matching m. The pragmatic speaker then rescores each candidate utterance, $u \in \mathcal{U}$, by using the probability that a literal listener, L_0, recovers the intended meaning of their message by observing u, $P_{L_0}(M = m|U = u)$. Finally, the pragmatic speaker combines these two probabilities with a rationality parameter λ to provide a final score for each utterance $u \in \mathcal{U}$ represented by[1]

$$S_1(u|m) = P_{S_0}(u|m)^{\lambda} \cdot P_{L_0}(m|u)^{1-\lambda} \qquad (1)$$

To adapt the RSA model to the task of non-generic summarization, we provide new definitions for the speakers, S_0, S_1, and listener, L_0, which we refer to as summarizers, S_0, S_1, and readers, R_0, respectively. We define the literal summarizer as a base summarization system, $P_{S_0}(Y|X = x, R = r)$, which takes as input some source text x and information request r. To define the literal reader, we approximate the recovery of the intended meaning of a summary y through the *latent reconstruction* objective. In this reconstruction objective, we use a language model conditioned on y with reconstruction target being the value z of a latent variable Z related to the initial information request

---

[1] We drop the P_{S_1} notation for the pragmatic speaker as, except for $\lambda = 0, 1$, this product does not necessarily induce a valid probability distribution.

r, $P_{R_0}(z|y)$. This latent reconstruction objective is then used by the pragmatic summarizer, S_1, to provide a final score to a summary y

$$S_1(y|x) = P_{S_0}(y|x,r)^{\lambda} \cdot P_{R_0}(z|y)^{1-\lambda} \quad (2)$$

The latent reconstruction objective is, to the best of our knowledge, the first implementation of the intended meaning of a non-generic summary in the context of RSA. Previous approaches in generic summarization have included the source reconstruction objective where the reconstruction target is the source text x as well as the distractor-based objective where, given a generic summary, y, R_0 should distinguish between the true source text x and a distractor source text x' (Andreas and Klein, 2016; Cohn-Gordon et al., 2018; See et al., 2017). We believe that, in the context of non-generic summarization, the latent reconstruction objective is more in line with the true intended meaning of a summary than previous implementations of meaning in generic summarization.

4 Experiments

In this section, we describe the experiments, datasets and evaluations we use to verify our hypothesis that a viable approach to non-generic summarization via RSA involves the latent reconstruction objective presented in Section 3.

4.1 Experimental Setup

We describe the pieces of our RSA-based non-generic summarization pipeline which starts with candidate summary generations by the literal summarizer, then goes through summary rescoring by the literal reader and final summary selection by the pragmatic summarizer.

4.1.1 Literal Summarizers

We use different pre-trained summarization systems as literal summarizers in our RSA setup of non-generic summarization. For each source text-information request pair, we use each summarization system to produce 5 candidate summaries using diverse beam search (Vijayakumar et al., 2016) (DBS). DBS is a variant of beam search which optimizes its decoding step for a diversity-augmented objective. We briefly describe the different summarization systems below. All model checkpoints are taken from HuggingFace [2].

BART BART (Lewis et al., 2019) is an encoder-decoder Transformer model trained with a denoising objective function and commonly used for language generation tasks. We use the `facebook/bart-large-cnn` checkpoint of BART which is fine-tuned on the CNN/Daily Mail summarization dataset. To account for the additional information request, we follow previous query-focused summarization work (Vig et al., 2022) and preprocess the input to BART by concatenating the information request and the source text with the model's delimiter token. We truncate the input from the right so that it fits within BART's 1024 maximum input token limit.

LED Similar to BART, the Longformer Encoder-Decoder (Beltagy et al., 2020) (LED) is an auto-denoising encoder-decoder which uses a Longformer as its encoder. The Longformer encoder uses a memory-efficient version of attention to allow for a maximum input token limit of 16384 tokens which may be beneficial in cases where the source text exceeds BART's 1024 input token limit. We use the `allenai/led-large-16384-arxiv` checkpoint of LED which is fine-tuned on the arXiv scientific article summarization dataset. We use the same input preprocessing decision as for BART to account for the additional information request of the non-generic summaries.

LLama2 Llama2 (Touvron et al., 2023) is a collection of LLMs developed by Meta pre-trained and fine-tuned on a "mix of publicly available online data." This class of models has been shown to be competitive on some summarization benchmarks (Yang et al., 2023a; Zhang et al., 2024). We use the `meta-llama/Llama-2-7b-chat-hf` checkpoint which is the 7 billion parameter version of Llama2 fine-tuned with both instruction tuning and reinforcement learning from human feedback (RLHF). We use the prompt shown in Figure 2 for non-generic summarization where the [QUESTION] placeholder is replaced by the information request.

4.1.2 Literal Reader

To implement the literal reader, we use the `google/flan-t5-large` checkpoint of the FLAN-T5 (Chung et al., 2022) series of general purpose LLMs. The FLAN-T5 LLM series is a collection of large language models with the same architecture as the original T5 series but trained on additional tasks and fine-tuned with instruction tuning. We

---

[2] `https://huggingface.co/`

As we discussed in a meeting, this step of constraining to 5 candidate summaries is probably the main bottleneck in the system right now. This with an interpolation between latent and source reconstruction would be fruitful to explore.

The paper is well-structured and easy to understand.

use the large (700 million parameters) version as a compromise between inference speed and capacity.

4.1.3 Pragmatic Summarizer

The pragmatic summarizer combines the likelihood scores produced by the summarization system and the LLM to select the final summary from the pool of 5 candidate summaries i.e., the final selected summary is

$$\arg\max_{i=1...5} P_{S_0}(\hat{y}_i|x,r)^{\lambda} \cdot P_{R_0}(z|\hat{y}_i)^{1-\lambda} \tag{3}$$

for some $\lambda \in [0,1]$.

4.2 Datasets

In this section, we describe the datasets that we use to run our experiments on the different implementations of S_0, S_1, R_0 in the context of non-generic summarization. We also describe our choice of information request and latent variable for each dataset. In addition, we provide examples of the different types of source texts, information requests and latent variables in Table 1 and summary statistics for each dataset in Table 2.

CovidET The CovidET dataset (Zhan et al., 2022) is a collection of Reddit[3] posts about different personal events experienced by people during the Covid-19 pandemic. Each post has multiple human-written reference summaries which each summarize one of the different emotions found in the Reddit post as well as the emotion's associated cause(s) or trigger(s).

In the case of CovidET, no information request is explicitly provided, thus we take the emotion e being summarized and convert it to the instruction "Describe the e of this post." to make an information request. We use the emotion e as the latent variable.

Debatepedia The Debatepedia dataset (Nema et al., 2017) is a collection of argumentative texts with a for or against stance related to a certain topic (e.g., politics, crime, environment, etc.). Each text has a certain topic and question associated with it, with certain texts having multiple associated topic-question pairs. Each text and topic-question pair has a human-written reference summary which summarizes the text by providing an answer to its related question.

For Debatepedia, we use the question associated with the source text as the summary's information request and the topic of the question as the latent variable.

DUC 2007 The DUC 2007 dataset (Over et al., 2007) is a multi-document query-focused summarization dataset with source texts consisting of news articles from the Associated Press, the New York Times (1998-2000) and the Xinhua News Agency (1996-2000). The news articles are grouped into collections of 25 news articles, all covering the same topic. A question is then associated with each group of 25 news articles and 4 human-written reference summaries are produced summarizing the 25 news articles while answering the initial question. Because handling multiple documents is beyond the scope of this work, we associate each topic, question and reference summaries with each of their corresponding 25 news articles to create a single-document non-generic summarization dataset.

Similar to Debatepedia, we use the question as the information request and the topic of the question as the latent variable.

MultiOpEd The MultiOpEd dataset (Liu et al., 2021) is a collection of news editorials which provide a perspective on a certain question about a controversial topic. The news editorial is accompanied by an abstract summarizing its stance and arguments and a perspective which provides a concise answer to the initial question.

For MultiOpEd, we use the question being asked as the information request and the perspective as the latent variable.

QMSum The QMSum dataset (Zhong et al., 2021) is a query-focused meeting summarization dataset where the source texts are excerpts from different meeting transcripts. Each excerpt has an associated question and reference summary summarizing the excerpt while providing an answer to the question.

In the case of QMSum, we use the question being asked about the meeting transcript excerpt as the information request. Since no answers to questions are explicitly provided in the dataset, we assume that the latent variable is the same as the summary's information request.

4.3 Evaluation

We evaluate our non-generic summarization pipeline by computing the ROUGE-1, ROUGE-2 and ROUGE-L scores between the summary se-

---

[3] `https://www.reddit.com/`

value

Dataset Name	Source text	Information request	Latent variable
CovidET	I dont even know how to speak of this grief. I have read of many stories of people losing their loved ones, but it didnt happen to [...]	Describe the disgust of this post	Disgust
Debatepedia	[...] Inviting foreigners to come to america as guest workers is equivalent to sending the message : you people are only fit to do menial jobs that americans think they are too good to do [...]	Would a guest worker program be fair to guest workers?	Guest workers
DUC 2007	In February 1999 some of OJ Simpson's possessions including his Heisman Trophy were sold at an auction which raised about $430,000. The proceeds were to go to the Brown [...]	Give an account of the developments in the life of OJ Simpson.	OJ Simpson developments
MultiOpEd	[...] Early Sunday afternoon President Donald Trump elevated Twitter to a virtual Cabinet position in his administration. He used it, rather than the attorney general, to order a possible criminal [...]	Is trump's investigation into the FBI justified?	Trump's trying to smear the Russia probe
QMSum	PhD F : We um {disfmarker} So we just put in an order for about twelve new machines, uh, to use as sort of a compute farm. And um, uh, we ordered uh, SUN - Blade - one - hundreds, [...]	What did PhD F think about computational resources?	-

Table 1: Examples of source texts (truncated) along with their corresponding information request and latent variable for each dataset we use. We omit the latent variable for QMSum since it is the same as its information request.

Dataset Name	Test Size	Average length			
		Source	Information Request	Latent Variable	Reference Summary
CovidET	1043	171	6	1	22
Debatepedia	893	70	9	1	10
DUC 2007	1124	466	22	4	243
MultiOpEd	531	925	7	7	102
QMSum	244	1020	13	13	58

Table 2: Summary statistics for each of the datasets. Test size refers to the number of unique source text-information request pairs in the test set of each dataset and the average length refers to the average space-separated string lengths of different dataset fields.

lected by the pragmatic summarizer and the reference summary. In the case that several reference summaries have been written for the same source text-information request pair (e.g., DUC 2007), we take the maximum ROUGE scores.

5 Results and Analysis

In this section, we present the results and analyses of our experiments which test the viability of using different variants of pragmatic summarizers in the context of modeling non-generic summarization via RSA.

5.1 Reconstruction-Only S_1

We first present the results for what we call the reconstruction-only pragmatic summarizer in which we set the rationality parameter value to $\lambda = 0$. In the case of $\lambda = 0$, the pragmatic summarizer selects the final summary uniquely based on $P_{R_0}(z|\hat{y})$ (Equation 2). To understand the usefulness of the latent reconstruction objective, we implement an additional pragmatic speaker which uses the source reconstruction objective in its scoring function with $\lambda = 0$. Additionally, we compare the reconstruction-only S_1's with the setting where $\lambda = 1$ i.e., where the summary selected by S_1 is the one with the highest likelihood score according to the literal summarizer, S_0. Finally, we also implement a Random baseline where the final summary is selected at random and an Oracle upper-bound which selects the candidate summary which maximizes the ROUGE score with the reference summary(ies). The results for the MultiOpEd dataset are presented in Table 3 and the results for the other datasets can be found in Appendix A.1.

We observe that, across all datasets and all literal summarizers, a reconstruction-only pragmatic summarizer, via either the latent or the source reconstruction objective, is likely to lead to higher ROUGE scores than relying on the initial likelihood scores of the literal summarizer. This finding is illustrated through Table 5 which aggregates the ROUGE performance scores from Table 3 and Appendix A.1 by counting, for each dataset and each literal summarizer, which final summary selection method achieves the highest ROUGE score.

In terms of determining which of the reconstruction objectives is more suitable for non-generic summarization, the frequencies provided by Table 5 do not provide decisive evidence. As a result, to further compare these two reconstruction objectives, we compute the relative ROUGE score performance difference between using the latent reconstruction objective and the source reconstruction objective in the pragmatic summarizer's scoring function. These relative score differences can be found in Table 4. Although it appears that the latent reconstruction objective is always detrimental

Might need an additional baseline: lexical similarity to information request, such as by word overlap.
See this paper for an old implementation of the idea that worked very well:

https://aclanthology.org/P06-2020.pdf

	BART			LED			Llama2		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
Random	0.257	0.049	0.153	0.251	0.040	0.146	**0.322**	**0.070**	**0.176**
$P_{S_0}(\hat{y}\|x)$	0.260	0.047	0.152	0.219	0.038	0.127	0.321	0.069	0.175
$P_{R_0}(x\|\hat{y})$	0.257	0.048	0.150	**0.288**	**0.047**	**0.154**	0.318	0.069	0.172
$P_{R_0}(z\|\hat{y})$	**0.267**	**0.053**	**0.156**	0.242	0.043	0.138	0.320	0.069	0.174
Oracle	0.297	0.071	0.178	0.320	0.066	0.180	0.355	0.089	0.197
SOTA	0.315/0.138/0.298 (Xu et al., 2023)								

Table 3: ROUGE scores for the reconstruction-only S_1 setting described in Section 5.1 on the MultiOpEd dataset. We also include the ROUGE scores of the existing state-of-the-art (SOTA) non-generic summarization model on this dataset.

	BART			LED			Llama2		
Dataset	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
CovidET	+2.24	+3.66	+4.08	−16.42	−18.90	−13.25	+0.55	+0.52	+1.93
Debatepedia	−0.80	+0.77	−0.23	−6.63	−8.13	−5.08	−5.60	−9.85	−5.34
DUC 2007	+0.35	+4.24	+2.14	−8.45	−5.52	−5.02	+0.74	+0.90	+0.70
MultiOpEd	+3.74	+11.11	+4.23	−15.94	−7.64	−10.56	+0.70	+0.99	+1.03
QMSum	+4.71	+3.53	+4.87	−5.04	−5.90	−1.06	−0.09	−0.65	−1.01

Table 4: Relative average ROUGE score performance differences between a pragmatic summarizer which uses the latent reconstruction objective and one that uses the source reconstruction objective. A value of $+x$ implies that the average ROUGE score achieved with the latent reconstruction objective is $x\%$ higher than the one achieved with the source reconstruction objective.

	R-1		R-2		R-L	
	Freq.	%	Freq.	%	Freq.	%
Random	2	13.3	2	13.3	2	13.3
$P_{S_0}(\hat{y}\|x)$	2	13.3	4	26.7	2	13.3
$P_{R_0}(x\|\hat{y})$	6	40.0	6	40.0	6	40.0
$P_{R_0}(z\|\hat{y})$	5	33.3	3	20.0	5	33.3

Table 5: Frequencies for when the maximal ROUGE score is achieved by the different models described in Section 5.1. The frequencies are computed across all datasets and for every literal summarizer implementation.

for the LED literal summarizer, it appears to be beneficial in most cases when the literal summarizer is either BART of LLama2. These results offer a promising case for using the latent reconstruction objective to implement meaning in the RSA framework when modeling non-generic summarization.

Despite promising results, the method of selecting final summaries using the latent reconstruction objective and a rationality parameter value of $\lambda = 0$ remains sub-optimal when comparing its ROUGE scores to the ones achieved by the Oracle upperbound. These relative performance differences are presented in Table 12 of Appendix A.1 and motivate the exploration of pragmatic speaker variants where $\lambda \in (0, 1)$.

5.2 Rationality Parameter Interpolation

To further investigate the viability of the latent reconstruction objective for modeling non-generic summarization via RSA, we evaluate the performance of pragmatic summarizers when the rationality parameter λ from Equation 2 varies from 0 to 1. In particular, we compute the ROUGE scores for the final summaries selected by S_1 based on values of $\lambda \in [0, 0.1, \dots, 0.9, 1.0]$. We show a plot of ROUGE-1 scores for the pragmatic summarizer using the latent-reconstruction-based literal reader as well as the source-reconstruction-based literal reader in Figure 3.

By analyzing the plots for every dataset and for every literal summarizer, we conclude that a rationality parameter which interpolates between the scores of the literal summarizer and the literal reader is more beneficial under the latent reconstruction objective than the source reconstruction objective. As Table 6 demonstrates, the highest attained ROUGE scores when using the latent reconstruction objective often occur for intermediate values of $\lambda \in (0, 1)$. For the pragmatic summarizer which uses the source reconstruction objective, the highest ROUGE scores occur more com-

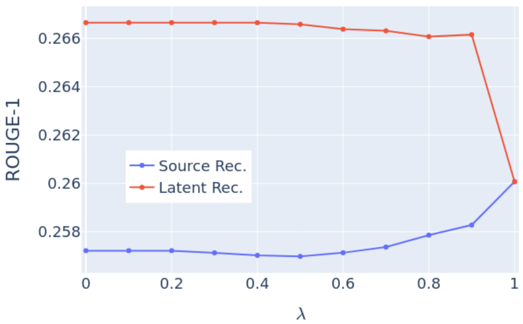


Figure 3: Plot of the average ROUGE-1 scores for the summaries selected by a pragmatic summarizer using a latent reconstruction objective as well as a source reconstruction objective. In this case, the dataset is MultiOpEd and the literal summarizer used to generated the summaries is BART.

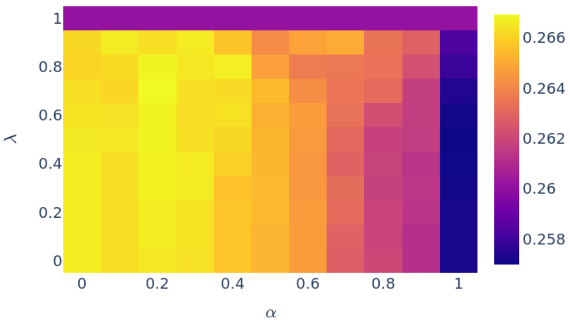


Figure 4: Heatmap of ROUGE-1 scores for different values of λ and α on the MultiOpEd dataset using BART as the literal summarizer.

	Source Reconstruction						Latent Reconstruction					
	R-1		R-2		R-L		R-1		R-2		R-L	
	Freq.	%	Freq.	%	Freq.	%	Freq.	%	Freq.	%	Freq.	%
$\lambda = 0$	6	40.0	7	43.75	7	46.67	7	43.75	5	31.25	6	37.5
$\lambda \in (0, 1)$	2	13.33	6	37.50	1	6.67	5	31.25	7	43.75	8	50
$\lambda = 1$	7	46.67	3	18.75	7	46.67	4	25.0	4	25.0	2	12.5

Table 6: Frequencies for when the maximum ROUGE score is achieved by the pragmatic summarizer using either the source or latent reconstruction objective and either $\lambda = 0$, $\lambda \in (0, 1)$ or $\lambda = 1$.

	R-1		R-2		R-L	
	Freq.	%	Freq.	%	Freq.	%
$\lambda = 1$	1	6.67	1	3.67	0	0.0
$\lambda = 0, \alpha = 1$	5	33.3	2	13.3	4	26.7
$\lambda = 0, \alpha = 0$	0	0.0	0	0.0	1	6.67
$\lambda \in (0, 1), \alpha \in (0, 1)$	9	60.0	12	80.0	10	66.7

Table 7: Frequencies for when the maximum ROUGE score is achieved by the pragmatic summarizer using the latent-source reconstruction objective.

monly when the selection of the final summary is based only on the literal summarizer's score ($\lambda = 1$) or on the literal reader's score ($\lambda = 0$). This finding suggests that the latent reconstruction objective may be more appropriate in non-generic summarization as it forces the pragmatic summarizer to find a middle ground between overfitting to either the literal summarizer's or the literal reader's summary scores.

5.3 Latent-Source Reconstruction Objective

Given the promise of interpolating the rationality parameter, we introduce another pragmatic summarizer which uses a hybrid latent-source reconstruction objective. To do so, we introduce a latent-source parameter $\alpha \in [0, 1]$ which interpolates between the latent and source reconstruction objective. The pragmatic summarizer scoring function thus becomes

$$S_1(y|x) = P_{S_0}(y|x, r)^{\lambda} \cdot \left(P_{R_0}(x|y)^{\alpha} \cdot P_{R_0}(z|y)^{1-\alpha}\right)^{1-\lambda} \quad (4)$$

where we notice that for $\alpha = 0$ and $\alpha = 1$ we recover S_1's scoring function using the latent and source reconstruction objectives respectively. For each dataset and for each literal summarizer, we compute the pragmatic summarizer's average ROUGE scores for every $\lambda \in [0, 0.1, \ldots, 0.9, 1]$ and for every $\alpha \in [0, 0.1, \ldots, 0.9, 1]$. A heatmap for the ROUGE-1 score on the MultiOpEd dataset using the BART literal summarizer is shown in Figure 4.

By computing frequencies across all datasets and all literal summarizers, we observe that in most cases an intermediate value for both $\lambda \in (0, 1)$ and $\alpha \in (0, 1)$ is beneficial in terms of average ROUGE score (Table 7). This finding suggests that the source reconstruction objective may hold additional signal which is orthogonal in terms of performance contribution to the latent reconstruction objective. This result may in part be due to the little information provided by the latent variables of most of the datasets we use. For instance, in the cases of CovidET and Debatepedia the value of the latent variables are typically one word and one noun phrase respectively. In brief, the latent reconstruction objective may be the most beneficial in the context of non-generic summarization when paired with the source reconstruction objective.

And that your current way of instantiating latent reconstruction doesn't capture all aspects of the query-focused summarization task (which is to be expected).

We'll eventually need human evaluations.

Is there a way to use RSA to do prompt tuning or prompt search? i.e., Find the prompt that, over a dataset of samples, best approximates the intended message to convey, using RSA to derive an objective for searching for this prompt.

6 Conclusion

Infeasibility might not be the best word.

In this paper, we have argued that previous attempts to model generic summarization via RSA have been inappropriate due to the infeasibility of the source reconstruction objective. As a result, we have argued for shifting the use of RSA to the context of non-generic summarization where the latent reconstruction objective is a more appropriate implementation of meaning. Through our experiments, we have shown that using the latent reconstruction objective in a pragmatic summarizer's scoring function may lead to performance improvements over existing summarization systems as well as over previous meaning implementations such as the source reconstruction objective. Furthermore, we find that the latent reconstruction objective may be most beneficial when paired with the source reconstruction objective as results suggest that these objectives offer complimentary information to the pragmatic summarizer. Thus, future work may investigate how these meaning implementations interact and whether they can be more seamlessly integrated within a non-generic summarization system which focuses on pragmatic language generation.

Limitations

Despite our best efforts, this work suffers from shortcomings which prevent us from making more significant conclusions regarding the use of RSA and the latent reconstruction objective in non-generic summarization. Firstly, we do not explore the use of literal summarizers *fine-tuned* on the datasets with which we experiment in this work. These fine-tuned models are responsible for achieving the existing SOTA ROUGE scores we presented in Table 3 and in Appendix A.1. This investigation is important since it is possible that when literal summarizers are fine-tuned there is no need for additional levels of reasoning as described in the RSA framework. Secondly, our exploration of alternative meaning implementations is limited to the source reconstruction objective and to the rescoring of 5 candidate summaries. A decoding method as described in Cohn-Gordon et al. (2018) which "folds" the reconstruction objective within its decoding step should also be compared against. Finally, in this work, we consider reconstruction objectives purely on the output space. However, a reconstruction objective in the representation space as described by Assran et al. (2023) may be more suitable as it would not suffer from the lack of informativeness found in the latent variables we consider in this work.

References

Jacob Andreas and Dan Klein. 2016. Reasoning about Pragmatics with Neural Listeners and Speakers. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1173–1182, Austin, Texas. Association for Computational Linguistics.

Xiang Ao, Xiting Wang, Ling Luo, Ying Qiao, Qing He, and Xing Xie. 2021. PENS: A Dataset and Generic Framework for Personalized News Headline Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 82–92, Online. Association for Computational Linguistics.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. 2023. Self-supervised learning from images with a joint-embedding predictive architecture.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *CoRR*, abs/2004.05150.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

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.

Hoa Trang Dang. 2006. DUC 2005: evaluation of question-focused summarization systems. In *Proceedings of the Workshop on Task-Focused Summarization and Question Answering - SumQA '06*, page 48, Sydney, Australia. Association for Computational Linguistics.

Hal Daumé and Daniel Marcu. 2006. Bayesian query-focused summarization. *Proceedings of the 21st International Conference on Computational Linguistics* *and the 44th annual meeting of the ACL - ACL '06*, pages 305–312. Conference Name: the 21st International Conference Place: Sydney, Australia Publisher: 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.

Angela Fan, David Grangier, and Michael Auli. 2018. Controllable Abstractive Summarization. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 45–54, Melbourne, Australia. Association for Computational Linguistics.

Daniel Fried, Jacob Andreas, and Dan Klein. 2018. Unified Pragmatic Models for Generating and Following Instructions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1951–1963, New Orleans, Louisiana. Association for Computational Linguistics.

HP Grice. 1975. Logic and conversation. *Syntax and Semantics*, 3:43–58.

Wojciech Kryściński, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. *arXiv preprint arXiv:1908.08960*.

Md Tahmid Rahman Laskar, Elena Khasanova, Xue-Yong Fu, Cheng Chen, and TN ShashiBhushan. 2024. Query-opt: Optimizing inference of large language models via multi-query instructions in meeting summarization. *ArXiv*, abs/2403.00067.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Siyi Liu, Sihao Chen, Xander Uyttendaele, and Dan Roth. 2021. MultiOpEd: A corpus of multi-perspective news editorials. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4345–4361, Online. Association for Computational Linguistics.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Preksha Nema, Mitesh M. Khapra, Anirban Laha, and Balaraman Ravindran. 2017. Diversity driven attention model for query-based abstractive summarization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1063–1072, Vancouver, Canada. Association for Computational Linguistics.

Khanh Nguyen. 2023. Language models are pragmatic speakers. *ArXiv*, abs/2305.17760.

Allen Nie, Reuben Cohn-Gordon, and Christopher Potts. 2020. Pragmatic issue-sensitive image captioning. In *Findings*.

Paul Over, Hoa Dang, and Donna Harman. 2007. Duc in context.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Chantal Shaib, Millicent L. Li, Sebastian Joseph, Iain J. Marshall, Junyi Jessy Li, and Byron C. Wallace. 2023. Summarizing, Simplifying, and Synthesizing Medical Evidence Using GPT-3 (with Varying Success). ArXiv:2305.06299 [cs].

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.

Dan Su, Yan Xu, Tiezheng Yu, Farhad Bin Siddique, Elham Barezi, and Pascale Fung. 2020. CAiRE-COVID: A Question Answering and Query-focused Multi-Document Summarization System for COVID-19 Scholarly Information Management. *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*. Conference Name: Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020 Place: Online Publisher: Association for Computational Linguistics.

Christopher Tauchmann, Thomas Arnold, Andreas Hanselowski, Christian M. Meyer, and Margot Mieskes. 2018. Beyond Generic Summarization: A Multi-faceted Hierarchical Summarization Corpus of Large Heterogeneous Data.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. ArXiv:2307.09288 [cs].

Jesse Vig, Alexander Fabbri, Wojciech Kryscinski, Chien-Sheng Wu, and Wenhao Liu. 2022. Exploring Neural Models for Query-Focused Summarization. *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1455–1468. Conference Name: Findings of the Association for Computational Linguistics: NAACL 2022 Place: Seattle, United States Publisher: Association for Computational Linguistics.

Ashwin K. Vijayakumar, Michael Cogswell, Ramprasaath R. Selvaraju, Qing Sun, Stefan Lee, David J. Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *CoRR*, abs/1610.02424.

Ruochen Xu, Song Wang, Yang Liu, Shuo Wang, Yichong Xu, Dan Iter, Chenguang Zhu, and Michael Zeng. 2023. Lmgqs: A large-scale dataset for query-focused summarization. In *Conference on Empirical Methods in Natural Language Processing*.

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.

Xianjun Yang, Yan Li, Xinlu Zhang, Haifeng Chen, and Wei Cheng. 2023a. Exploring the Limits of ChatGPT for Query or Aspect-based Text Summarization. Publisher: arXiv Version Number: 1.

Xianjun Yang, Yan Li, Xinlu Zhang, Haifeng Chen, and Wei Cheng. 2023b. Exploring the limits of chatgpt for query or aspect-based text summarization. *ArXiv*, abs/2302.08081.

Hongli Zhan, Tiberiu Sosea, Cornelia Caragea, and Junyi Jessy Li. 2022. Why do you feel this way? summarizing triggers of emotions in social media posts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9436–9453, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2024. Benchmarking Large Language Models for News Summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir R. Radev. 2021. Qmsum: A new benchmark for query-based multi-domain meeting summarization. *CoRR*, abs/2104.05938.

A Additional Results

In this section, we provide additional results for all our experiments.

A.1 Reconstruction-Only S_1

	S_0								
	BART			LED			Llama2		
R_0	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
Random	0.192	0.037	0.132	0.125	0.018	0.095	**0.131**	**0.026**	**0.093**
$P_{S_0}(\hat{y}\|x)$	0.200	0.039	0.134	0.106	0.017	0.082	0.130	**0.026**	0.092
$P_{R_0}(x\|\hat{y})$	0.199	0.039	0.133	**0.150**	**0.025**	**0.109**	0.130	**0.026**	0.091
$P_{R_0}(z\|\hat{y})$	**0.204**	**0.041**	**0.139**	0.125	0.020	0.094	**0.131**	**0.026**	**0.093**
Oracle	0.237	0.058	0.165	0.202	0.043	0.152	0.158	0.040	0.114
SOTA	0.262/0.069/0.179 (Yang et al., 2023b)								

Table 8: ROUGE scores for the reconstruction-only S_1 setting described in Section 5.1 on the CovidET dataset. We also include the ROUGE scores of the existing state-of-the-art (SOTA) non-generic summarization model on this dataset.

	S_0								
	BART			LED			Llama2		
R_0	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
Random	0.144	0.045	0.120	0.079	0.022	0.071	0.080	0.024	0.069
$P_{S_0}(\hat{y}\|x)$	**0.147**	**0.046**	**0.121**	0.075	0.022	0.067	0.079	0.023	0.068
$P_{R_0}(x\|\hat{y})$	0.146	0.045	0.120	**0.089**	**0.027**	**0.079**	**0.083**	**0.025**	**0.071**
$P_{R_0}(z\|\hat{y})$	0.145	0.045	0.120	0.083	0.025	0.075	0.078	0.023	0.067
Oracle	0.175	0.063	0.149	0.125	0.044	0.112	0.106	0.037	0.091
SOTA	0.236/0.076/0.210 (Xu et al., 2023)								

Table 9: ROUGE scores for the reconstruction-only S_1 setting described in Section 5.1 on the Debatepedia dataset. We also include the ROUGE scores of the existing state-of-the-art (SOTA) non-generic summarization model on this dataset.

	S_0								
	BART			LED			Llama2		
R_0	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
Random	0.169	0.047	0.108	0.238	0.056	0.139	0.367	0.096	0.187
$P_{S_0}(\hat{y}\|x)$	**0.176**	**0.050**	**0.111**	0.230	0.054	0.136	0.369	0.097	0.188
$P_{R_0}(x\|\hat{y})$	0.173	0.048	0.109	**0.261**	**0.061**	**0.147**	0.374	0.098	0.188
$P_{R_0}(z\|\hat{y})$	0.173	**0.050**	**0.111**	0.239	0.058	0.140	**0.377**	**0.099**	**0.189**
Oracle	0.193	0.063	0.125	0.295	0.075	0.169	0.402	0.115	0.205
SOTA	-/-/-								

Table 10: ROUGE scores for the reconstruction-only S_1 setting described in Section 5.1 on the DUC 2007 dataset. Because of the manipulation we carry out on this dataset to make it a single-document non-generic summarization dataset, there are no existing SOTA model ROUGE scores.

	S_0								
	BART			LED			Llama2		
Dataset	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
CovidET	14.02	30.12	15.91	25.84	41.76	28.62	17.21	35.16	18.40
Debatepedia	15.86	27.54	18.72	29.27	39.59	29.73	21.40	31.93	22.67
DUC 2007	9.04	19.68	10.80	11.31	18.26	12.78	6.31	14.11	7.61
MultiOpEd	10.27	25.09	12.38	9.95	29.52	14.04	9.41	22.01	10.77
QMSum	16.50	29.83	19.18	18.90	33.25	20.40	14.04	23.49	16.09

Table 12: Relative ROUGE score performance difference between the Oracle final-summary selection method and the “next best” selection method, i.e. the method that received the highest ROUGE score after the Oracle.

	S_0								
	BART			LED			Llama2		
R_0	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
Random	0.281	0.087	0.193	0.170	0.035	0.129	**0.295**	**0.095**	**0.193**
$P_{S_0}(\hat{y}\|x)$	0.284	**0.094**	0.195	0.155	0.032	0.125	0.289	0.092	0.188
$P_{R_0}(x\|\hat{y})$	0.278	0.089	0.190	**0.192**	**0.045**	**0.138**	0.293	**0.095**	0.191
$P_{R_0}(z\|\hat{y})$	**0.291**	0.093	**0.199**	0.182	0.043	0.136	0.293	0.094	0.189
Oracle	0.349	0.133	0.247	0.237	0.068	0.173	0.343	0.125	0.230
SOTA	0.535/0.263/0.329 (Laskar et al., 2024)								

Table 11: ROUGE scores for the reconstruction-only S_1 setting described in Section 5.1 on the QMSum dataset. We also include the ROUGE scores of the existing state-of-the-art (SOTA) non-generic summarization model on this dataset.