Factorizing Content and Budget Decisions in Abstractive Summarization of Long Documents

Marcio Fonseca   **Yftah Ziser**   **Shay B. Cohen**

Institute for Language, Cognition and Computation
School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh, EH8 9AB
`m.fonseca@ed.ac.uk, yftah.ziser@ed.ac.uk, scohen@inf.ed.ac.uk`

Abstract

We argue that disentangling content selection from the budget used to cover salient content improves the performance and applicability of abstractive summarizers. Our method, FAC-TORSUM[1], does this disentanglement by factorizing summarization into two steps through an energy function: (1) generation of *abstractive summary views* covering salient information in subsets of the input document (*document views*) ; (2) combination of these views into a final summary, following a budget and content guidance. This guidance may come from different sources, including from an *advisor* model such as BART or BigBird, or in oracle mode – from the reference. This factorization achieves significantly higher ROUGE scores on multiple benchmarks for long document summarization, namely PubMed, arXiv, and GovReport. Notably, our model is effective for domain adaptation. When trained only on PubMed, it achieves a 46.29 ROUGE-1 score on arXiv, outperforming PEGASUS trained in domain by a large margin. Our experimental results indicate that the performance gains are due to more flexible budget adaptation and processing of shorter contexts provided by partial document views.

1 Introduction

Casting summarization as a language transduction problem is convenient given the existence of powerful neural sequence-to-sequence models that produce high-quality textual outputs (Zhang et al., 2020; Lewis et al., 2020). However, this framework conflates multiple steps of the summarization decision-making process into a single feedforward step without taking into account the contextual factors involved (Jones et al., 1999).

One such decision depending on context factors is the quantity of information to be included in a summary, reflecting on the generated outputs'

length. This factor is particularly relevant for long documents that cover many different aspects of interest for which different summaries may be suitable. For instance, in samples from summarization datasets such as PubMed and arXiv (Cohan et al., 2018), there are many abstracts including terse passages about the background or methods of the research. In contrast, others will add more details about those aspects. Often, those choices are due to the author's preferences and do not necessarily represent an ideal summary for the document.

Furthermore, current evaluation protocols based on n-gram overlap are sensitive to summary lengths (Sun et al., 2019). Generating summaries that match ground-truth lengths increases performance (see Section 4.1). Thus, recent progress in summarization may be the effect of better length prediction and not the actual summarization desideratum: a reductive transformation of the source text that keeps the important information (Jones et al., 1999).

To address this issue, we propose to avoid budget information as a confounding factor as much as possible in sequence-to-sequence training. Instead, we treat budget decisions as *extrinsic guidance* during summary generation, that is, an objective that is unrelated to the content of the documents. In this setting, the neural abstractive model is responsible for the generation of short passages (*summary views*) capturing relevant topics of the input document (*intrinsic importance* objective), while the *extrinsic importance* objective will encourage the adherence of generated summaries to context factors such as budgets or aspect coverage.

Specifically, we formulate FACTORSUM, a factorized energy-based model (LeCun et al., 2006) aiming to find a summary that maximizes the total importance given a source document, a reference dataset, and contextual factors such as budget and content guidance. A key piece of our model is the sampling of *random document views* (and corresponding reference *summary views*) which

---

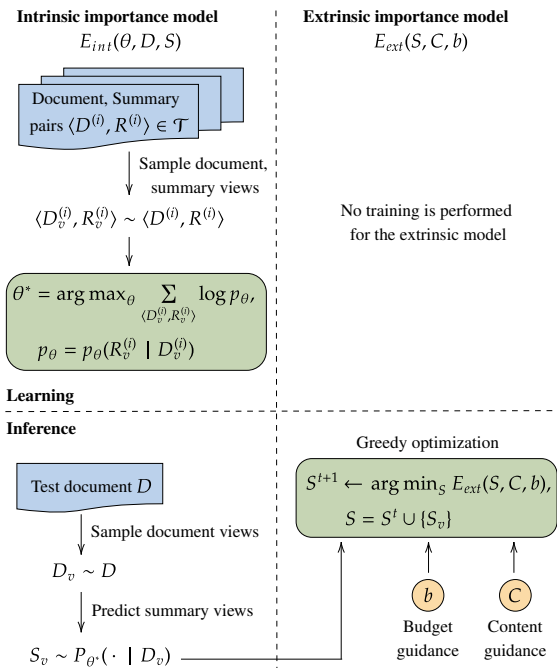[1] Code is available at `https://github.com/thefonseca/factorsum`.

Figure 1: An overview of the summarization model. Each quadrant represents the learning/inference step of either the intrinsic or extrinsic importance model. Intrinsic learning is implemented as the usual training of a sequence-to-sequence model p_θ, but using shorter sampled document and summary views. Optimization procedures are represented by rounded rectangles.

allows the abstractive model to focus on shorter summarization tasks with less influence of varying summary lengths. Also, this approach allows the processing of long documents without truncation, which is a recurring problem in summarization (Beltagy et al., 2020; Zaheer et al., 2020). The sampling procedure is detailed in Section 2.1.

Our model comprises two optimization procedures: learning and inference. In the learning phase, the model parameters are optimized so that summary views with important content (as informed by reference summaries) will have lower energies. In practice, this is implemented by training a neural sequence-to-sequence model to predict summary views. During inference, a greedy optimization algorithm is used to find the combination of summary views that maximize the compatibility with the target budget and other types of guidance. This process is illustrated in Figure 1.

Our experimental results on the PubMed, arXiv, and GovReport summarization benchmarks (Cohan et al., 2018; Huang et al., 2021) show that our approach using budget guidance alone is competitive with resource-intensive baselines such as PEGA-

SUS (Zhang et al., 2020). The results confirm that matching reference summary lengths significantly impacts ROUGE scores, often more than different modeling approaches. We also investigate the use of existing baselines as additional guidance during summary generation. In contrast to teacher models in knowledge distillation literature (Hinton et al., 2015), we leverage existing model prediction during inference only, and thus, we adopt the term *advisor* model to refer to our summarization guidance approach. When guided by BigBird or BART, our model obtains state-of-the-art results on PubMed, arXiv, and GovReport.

Finally, we perform domain adaptation experiments in which models trained on PubMed, arXiv, and GovReport have no access to samples from the evaluation dataset during training. Our results indicate that FACTORSUM can adapt better to out-of-domain data, outperforming strong baselines trained in domain on both PubMed and arXiv. This finding suggests a good generalization capacity and is evidence that we achieved our objective to disentangle content selection from budget decisions.

2 Intrinsic and Extrinsic Importance

Since our objective is to explicitly model budget decisions, we need a definition of importance that accounts for context factors. Inspired by the information-theoretic notion of importance developed by Peyrard (2019), we introduce the notion of importance with respect to *intrinsic* and *extrinsic* semantic units.

Intrinsic semantic units are those specific to the document (e.g., salient topics) whereas extrinsic units are related to a priori external preferences or require domain knowledge and grounding that is hard to capture from the textual corpora alone. We argue that the usual setting of end-to-end supervised summarization evaluated by ROUGE (Lin, 2004) optimizes for intrinsic importance. In this work, budget and content guidance provided by advisor model summaries (Section 2.1.2) play the role of extrinsic information.

Formally, we define the best summary S^* for a document D as the summary that minimizes the following factorized energy (LeCun et al., 2006):

$$E(\theta, D, S, C, b) = \qquad (1)$$
$$E_{int}(\theta, D, S) + E_{ext}(S, C, b),$$

where we call E_{int} and E_{ext} intrinsic and extrinsic energies respectively, while b denote the summary

budget guidance and C is a guidance content provided by an advisor model as explained in Section 2.1.2. This energy point of view allows us to unify the notion of extrinsic and intrinsic semantic units, and present the duality of the energy functions with respect to learning and inference.

Furthermore, the factorization of the total energy function makes the problem more tractable and leads to the following advantages:

- Model components can be changed or replaced more cost-effectively. For example, adding more components to the extrinsic objective would not require retraining the intrinsic importance model.
- Issues with differentiability of the extrinsic guideline loss with respect to the summary views generator parameters are avoided.
- More complex inference procedures that just feedforward computation are possible.

An overview of the model components and the summary inference procedure is represented in Figure 1. Given a document D, n_d document views are generated, each covering a random subset of sentences from the original document (Section 2.1). Then, the intrinsic importance model generates *summary views* S_v, each partially covering salient content from the original document D (Section 2.1.1). Finally, the extrinsic importance model will optimize the final summary S so that it maximizes the alignment of the content to a target budget and content guidance. In the following sections, we detail the model components as well as the training and inference procedures.

2.1 Sampling Document Views

Our model requires multiple summary proposals (or views) to allow the optimization of the extrinsic energy (Eq. 1). One further motivation for using document views is that it allows the intrinsic encoder-decoder model to focus on shorter sequences, which makes the less affected by truncation issues. To generate multiple *views* for the same document, we implement the following steps:

- From a document D, we generate a random sample of sentences, which we call a *document view* D_v. The number of sentences in D_v is controlled by the sampling factor parameter $s_f \in [0, 1]$, so that $\text{n\_sents}(D_v) \approx s_f \cdot \text{n\_sents}(D)$.
- Also from D, we extract oracle sentences o_i corresponding to each sentence r_i in the reference summary R. We choose

as oracle the sentences that maximize the sum of ROUGE-1 and ROUGE-2 F1 scores: $o_i = \text{argmax}_{s \in D} \text{ROUGE\_1}(s, r_i) + \text{ROUGE\_2}(s, r_i)$.

- For each oracle sentence $o_i \in D_v$ we collect the corresponding sentence r_i from the reference summary R. These sentences r_i form the reference summary R_v for the document view D_v. If there is no oracle sentence in the document view, the reference summary view R_v is empty[2].

For each document D from a training dataset $\mathcal{T}$, we repeat the sampling procedure described above n_d times, yielding a new dataset $\mathcal{T}' = \{(D_v^{(i)}, R_v^{(i)}) : i = 1, \ldots, |\mathcal{T}| \times n_d\}$ with n_d times more samples than the original data. The number of samples n_d and the sample fraction s_f are hyperparameters that to be tuned for each dataset. For PubMed, by sampling $n_d = 20$ views per document, each with 20% of the document sentences, we obtain document views with 17.2 sentences on average, while covering 99.1% of the original oracle sentences. In Appendix A, we provide statistics for different n_d and s_f and the heuristics we use for choosing appropriate values.

The intuition behind this sampling method is that if a sentence is relevant for the entire document, it should also be relevant in different contexts. Besides allowing the decoupled energy minimization objective, this approach also makes the input documents and corresponding summaries much shorter than the original data. Thus, our method scales to long documents without requiring specialized architectures for modeling long sequences (Beltagy et al., 2020; Zaheer et al., 2020). Also, in contrast to previous work, this summary sampling is domain-agnostic as it does not make any assumption about the discourse structure of the document (Dong et al., 2021; Gidiotis and Tsoumakas, 2020).

2.1.1 Intrinsic Importance Model

Powerful sequence-to-sequence models such as PEGASUS (Zhang et al., 2020) and BART (Lewis et al., 2020) are trained to estimate the probability of a sequence of tokens given a document by minimizing cross-entropy with respect to the data distribution. We hypothesize that these models are good candidates to fulfill the intrinsic importance objective, as described below.

---

[2]Except for training, when we enforce that each document view has at least one oracle sentence.

6343

Learning Given the training dataset $\mathcal{T}'$ consisting of document views $D_v^{(i)}$ and reference summary views $R_v^{(i)}$ (Section 2.1), we define the intrinsic loss as a negative log-likelihood functional:

$$L(E_{int}, \mathcal{T}') = \frac{1}{|\mathcal{T}'|} \sum_{i=1}^{|\mathcal{T}'|} \underbrace{L(R_v^{(i)}, E_{int}(\theta, D_v^{(i)}, \mathcal{S}))}_{-\log p_\theta(R_v^{(i)}|D_v^{(i)})}$$

where $p_\theta(R_v^{(i)}|D_v^{(i)})$ is a distribution over the possible summaries $\mathcal{S}$, specifically a sequence-to-sequence neural network model (Lewis et al., 2020). During learning, we find the parameters θ^* that minimize the loss above.

Inference The summary generation is performed as usual in sequence-to-sequence models via beam search decoding (Sutskever et al., 2014). We sample summary views by generating a summary conditioned to the document views:

$$S_v^{(i)} \sim p_{\theta^*}(\,\cdot\,|D_v^{(i)}). \tag{2}$$

We assume these summary views are samples from low-energy regions of $E_{int} = -\log p_{\theta^*}(S_v^{(i)}|D_v^{(i)})$, thus contributing to the minimization of the factorized energy (Eq. 1).

2.1.2 Extrinsic Importance Model

The extrinsic importance energy function E_{ext} measures the compatibility between the summary, the guidance budget b, and the guidance content C. Thus, the optimal summary S^* is defined as:

$$S^* = \arg \min_S E_{ext}(S, C, b). \tag{3}$$

The extrinsic energy is defined in terms of the squared deviation with respect to the guidance budget and ROUGE-1 score between the generated summary and the guidance content:

$$E_{ext}(S, C, b) \tag{4}$$
$$= \alpha \left(|S|/b - 1\right)^2 - \beta \,\text{ROUGE\_1}(S, C),$$

where $|S|$ denotes the length of the summary in words, the content C is a summary provided by an advisor model. The hyperparameters α and β weight the contribution of each guidance signal. In our experiments, we use $\alpha = \beta = 1.0$ and, as advisor models, PEGASUS (Zhang et al., 2020) and BigBird (Zaheer et al., 2020).

In our implementation, there is no learning step for the extrinsic importance model. For inference,

we design a greedy algorithm to minimize the energy as detailed in Algorithm 1. Let V_D be the set of n_d summary views for the document D. Starting from the initial condition $S = \emptyset$, the procedure selects the summary view $S_v \in V_D$ that minimizes the energy $E_{ext}(S \cup \{S_v\}, C, b)$. The view S_v is added to the summary if it satisfies the following additional conditions:

- **Non-redundancy**: the view S_v cannot be redundant with respect to the current summary S. We consider as redundant a summary view that has a (word-level) normalized Levenshtein distance[3] (Levenshtein et al., 1966) to any sentence in S lower than a threshold $t = 0.4$[4] (is_redundant function in Algorithm 1).
- **Energy reduction**: $S \cup \{S_v\}$ must have a lower energy than the current best summary S^*. After p iterations without improvement, the algorithm returns the current best summary S^*. Unless otherwise stated, this patience parameter is set to $p = n_d$, which means the algorithm iterates over all available views.

When $\beta = 0$ in Eq. 4 (no content guidance), each step of the greedy algorithm adds the longer summary view that satisfies the non-redundancy condition above, except for the last step, when a shorter view may better match the budget guidance. We provide further details on pre- and postprocessing summary views in Appendix D.

3 Experimental Setup

We turn to discuss the experimental setup to assess the effectiveness of the factorized model proposed in Section 2. We specify details about datasets, baselines, and the evaluation protocol. (See Appendices C and D for further implementation details).

Datasets Our experiments are performed on the PubMed and arXiv datasets, consisting of documents extracted from the homonymous scientific repositories (Cohan et al., 2018). To further test the generalization capacity of the model, we also perform the experiments on GovReport, a dataset containing long reports published by U.S. Government Accountability Office (GAO; Huang et al. 2021). The only preprocessing applied is to filter out documents with empty articles or summaries,

---

[3]We use the textdistance library:
https://github.com/life4/textdistance
[4]The redundancy threshold was manually tuned by inspecting sample outputs from the validation set.

Algorithm 1: Greedy summary generation. Input parameters are the set of summary views V_D for document D, content guidance C, budget guidance b, redundancy threshold t, and patience p. See the "Non-redundancy" paragraph in Section 2.1.2 for a discussion about the `is_redundant` function.

> **Input:** V_D, C, b, t, p
> **Output:** S^*
> $S \leftarrow \emptyset, S^* \leftarrow \emptyset$;
> $i \leftarrow 0$;
> **while** $V_D \neq \emptyset$ and $i \leq p$ **do**
> $S_v^* \leftarrow \arg\min_{S_v \in V_D} E_{ext}(S \cup \{S_v\}, C, b)$
>   **if** $E_{ext}(S \cup \{S_v^*\}, C, b) > E_{ext}(S^*, C, b)$ **then**
> $i \leftarrow i + 1$;
> $S \leftarrow S \cup \{S_v^*\}$;
>   **else if not** $is\_redundant(S, S_v^*, t)$ **then**
> $i \leftarrow 0$;
> $S \leftarrow S \cup \{S_v^*\}$;
> $S^* \leftarrow S$;
> $V_D \leftarrow V_D \setminus \{S_v^*\}$;
> **end**

Dataset	Samples			Summaries	
	Train	Val	Test	Sents	Words
PubMed	119,920	6,631	6,658	6.8	204.8
arXiv	202,917	6,436	6,440	12.6	292.6
GovReport	17,517	973	973	17.6	546.0

Table 1: Key statistics for the summarization datasets. "Sents" and "Words" denote the average number of words and sentences in the summaries (training split).

which lead to the training, validation, and test splits shown in Table 1. We do not truncate the articles or their abstracts.

Evaluation We evaluate our models using the ROUGE F-measure metric (Lin, 2004), with the implementation used by Zhang et al. (2020)[5].

Baseline models We use the following summarization baselines, for which implementations and pre-trained models are publicly available:

- **PEGASUS** (Zhang et al., 2020), an encoder-decoder transformer-based model that uses a specialized pre-training task of predicting entire masked sentences and achieves strong performance across several datasets.
- **BigBird** (Zaheer et al., 2020), a model based on a sparse attention mechanism that allows

transformer-based models to process up to 8 times longer sequences efficiently.
- **BART** (Lewis et al., 2020), a transformer-based denoising autoencoder that has strong performance on text generation tasks. We train our own version of BART-large on GovReport with a longer maximum target length of 768 tokens.

Also, we add results for the following abstractive systems: DANCER (Gidiotis and Tsoumakas, 2020), HEPOS (Huang et al., 2021), DYLE (Mao et al., 2021), and SUMM^N (Zhang et al., 2021).

4 Results

In this section, we analyze the contribution of different guidance factors on summarization performance by controlling summary budget and content guidance. We continue by conducting an ablation study, examining the summary views and the greedy summary generation contributions to the overall performance. Finally, we discuss domain adaptation results. Sample summaries are provided in Appendix F.

4.1 Effects of Budget Guidance

To test the impact of budgets on the summarization performance we provide three types of guidance:

- **Fixed**: the budget guidance is set at a fixed value for all summaries. The budget is 205, 165, and 648 words, which are the average summary lengths in the validation sets of PubMed, arXiv, and GovReport respectively.
- **Oracle**: the model uses the reference summary length as the budget guidance.
- **Model-based**: the model uses the length of the summary produced by an advisor model (BART for GovReport and BigBird for PubMed/arXiv) as budget guidance.

To fairly compare the different models, we use an additive budget correction so that the average number of tokens in system summaries is close to the average length of reference summaries from the validation set (Table 9). The average summary lengths for each model is presented in Tables 2 and 3. Also, we provide ROUGE scores for varying budget guidance values in Appendix B.

Fixed budget The simpler version of FACTOR-SUM is only guided by a fixed budget and its performance is competitive with most baselines, including PEGASUS and BigBird (Table 2, "no content

Model	PubMed				arXiv				GovReport			
	R-1	**R-2**	**R-L**	**Len**	**R-1**	**R-2**	**R-L**	**Len**	**R-1**	**R-2**	**R-L**	**Len**
Previous work												
DANCER†	46.34	19.97	42.42	-	45.01	17.60	40.56	-	-	-	-	-
HEPOS†	**48.12**	**21.06**	42.72	-	48.24	20.26	41.78	-	56.86	22.62	53.82	-
DYLE†	-	-	-	-	46.41	17.95	41.54	-	**61.01**	**28.83**	**57.82**	-
SUMM^N†	-	-	-	-	-	-	-	-	56.77	23.25	53.90	-
PEGASUS	43.83	18.72	40.29	180	43.06	16.39	38.65	168	-	-	-	-
BigBird	45.48	19.92	41.81	185	46.15	18.60	41.46	164	-	-	-	-
BART-large	-	-	-	-	-	-	-	-	52.82	19.12	49.99	596

Guidance												
Budget	**Content**											
FACTORSUM - no content guidance												
Oracle	-	47.37	19.10	43.27	208	48.87	18.83	43.96	167	59.80	24.13	56.12	651
Fixed	-	45.41	18.66	41.63	206	47.22	18.60	42.61	165	58.77	23.99	55.19	650
Model	-	44.64	17.98	40.76	185	46.40	18.21	41.85	164	57.18	23.34	53.66	638
FACTORSUM - content guidance												
Oracle	Lead	48.31	19.99	44.35	208	49.69	19.32	44.85	166	60.73	25.24	57.20	650
Fixed	Lead	46.27	19.29	42.57	205	48.05	19.05	43.49	165	59.67	25.02	56.22	649
Fixed	Model	47.50	20.33	**43.76**	205	**49.32**	**20.27**	**44.76**	165	60.10	25.28	56.65	648
Model	Model	47.34	20.31	_43.52_	185	48.74	_20.12_	44.19	164	58.78	24.87	55.37	638

Table 2: ROUGE F1 scores and average words per summary on the test sets for different types of guidance during inference. *Lead* guidance is the first k sentences from the source document (Section 4.2). Model guidance is provided by BART-large for GovReport and BigBird for PubMed and arXiv. The choice of budget guidance values is described in Appendix D and validation scores are provided in Appendix E. Results for models marked with † are taken from the original publications. Underlined results are statistically equivalent to the best methods ($p < 0.05$).

guidance"). It is important to note that the intrinsic importance model is based on a `bart-base` model with 139M parameters, which is 4 times smaller than PEGASUS and BigBird.

Oracle budgets The second section of Table 2 shows that for FACTORSUM with no content guidance, having access to the oracle lengths improves the scores by about 2, 1.6, and 1 ROUGE-1 on PubMed, arXiv, and GovReport respectively. We observe a similar effect with content guidance (third section of Table 2). These results agree with our hypothesis that the impact of budgets on ROUGE is significant and often larger than the differences between different modeling approaches.

Model-based budgets One may argue that inferring summary lengths is part of the task. Thus, we also test how summary lengths provided by BART and BigBird affect the summarization performance. For all datasets, we observe that using model budget guidance is detrimental to the scores compared

to fixed budget guidance (second and third section of Table 2). These results suggest that summary lengths are hard to predict from the source documents, and highlights the potential benefits of divorcing content selection and budget optimization.

4.2 Effects of Content Guidance

We also examine the impact of different types of content guidance. The first sort of guidance, *Lead*, takes the first k sentences from the source article. We choose the lowest k so that the guidance text has at least the number of words as the fixed target budget for each dataset (see Section 4.1). This content guidance improves the scores by ~0.8 (PubMed and arXiv) and ~0.9 (GovReport) ROUGE-1 points over the model without guidance. Notably, FACTORSUM with *Lead* guidance achieves 48.05 ROUGE-1 on arXiv and 59.67 ROUGE-1 on GovReport *without relying on predictions from strong baselines*. We can further improve performance by providing content guidance

| Evaluation | PubMed | | | | arXiv | | | | GovReport | | | |
Training	R-1	R-2	R-L	Len	R-1	R-2	R-L	Len	R-1	R-2	R-L	Len
End-to-end baseline (BigBird for PubMed and arXiv; BART-large for GovReport)												
PubMed	45.48	19.92	41.81	185	42.33	15.16	38.09	161	19.35	3.57	18.10	222
arXiv	39.47	14.95	35.77	177	46.15	18.60	41.46	164	16.61	2.25	15.09	352
GovReport	37.18	11.10	33.96	203	35.11	8.94	31.67	203	52.82	19.12	49.99	596
FACTORSUM - fixed budget, no content guidance												
PubMed	45.41	18.66	41.63	206	44.61	15.88	40.16	165	42.49	15.07	39.92	350
arXiv	44.40	16.87	40.51	209	47.22	18.60	42.61	165	**48.75**	**18.07**	**45.94**	414
GovReport	39.67	12.63	35.37	213	38.34	10.74	33.72	167	58.77	23.99	55.19	650
FACTORSUM - fixed budget and content guidance (BigBird guidance for PubMed and arXiv; BART-large guidance for GovReport)												
PubMed	47.50	20.33	43.76	205	**46.29**	**17.13**	**41.86**	166	42.24	15.03	39.68	344
arXiv	**45.87**	**18.10**	**42.02**	210	49.32	20.27	44.76	165	48.65	18.03	45.85	410
GovReport	41.27	14.01	37.10	211	40.00	11.85	35.47	176	60.10	25.28	56.65	648

Table 3: ROUGE F1 scores and average words per summary for the domain adaptation experiments. Models trained on PubMed, arXiv, and GovReport samples (rows) are used to summarize articles from the other dataset test splits (columns). The choice of budget guidance values is described in Appendix D. Shaded scores are in-domain results from Table 2. Underlined results are statistically equivalent to the best cross-domain scores ($p < 0.05$).

from BigBird and BART, leading to strong performance on all datasets (Table 2). From these empirical results, we conclude that content guidance is a simple and effective method to turn strong sequence-to-sequence baselines into more flexible summarization systems. It should be possible to add more types of guidance to adapt the summaries to specific needs such as topic coverage.

4.3 Ablation Study

Our method comprises two main components, a document and summary views and a ranker extracting the most salient information from those summaries using budget guidance. To shed some light on the contribution of each component to the overall performance, we conduct two ablation analyses. First, to better understand the inherent potential of the summary views, we feed them to the FACTORSUM ranker with reference (oracle) content guidance. Having reference as content guidance serves as an upper bound for what we can achieve using the summary views. We do the same to an ensemble of PEGASUS and BigBird, which is a concatenation of their summaries, for comparison.

Second, to understand the importance of the FACTORSUM ranker (see Algorithm 1), we replace it with TextRank (Mihalcea and Tarau, 2004), a prominent algorithm for extractive summarization.

Having two input variants (summary views and PEGASUS and BigBird ensemble) and three ranker variants (TextRank, FACTORSUM - no content guidance, and FACTORSUM - reference content guidance) results in six models. We use fixed-length guidance for all of the models.

Table 4 shows our results on the PubMed and arXiv datasets. We observe significantly higher results for FACTORSUM - reference content guidance) when applied on summary views, compared to PEGASUS and BigBird Ensemble Summaries. This shows that when both inputs reach their full potential under ideal guidance, summary views are superior to an ensemble of the two strong baselines, thus containing more salient information. We observe a similar pattern when using FACTORSUM - no content guidance as a ranker, showing that summary views serve as a better input to a more realistic ranker. In addition, we notice that for both types of input, FACTORSUM - no content guidance significantly outperforms TextRank, thus contributing to FACTORSUM overall performance.

4.4 Domain Adaptation

Our last experiment, unusual in the summarization literature, aims to test if a model trained on PubMed/arXiv/GovReport performs well when applied to out-of-domain (OOD) samples. Our in-

Ranker	PubMed			arXiv		
	R-1	R-2	R-L	R-1	R-2	R-L
PEGASUS + BigBird Ensemble Summaries						
TextRank	43.93	18.33	38.40	44.15	16.89	37.32
FS	<u>45.38</u>	**19.43**	<u>41.49</u>	45.30	17.60	40.38
FS-Oracle	48.90	21.81	44.76	49.34	20.13	43.99
Summary Views						
TextRank	42.10	16.71	37.54	42.66	16.41	37.70
FS	<u>**45.41**</u>	18.66	<u>**41.63**</u>	<u>**47.22**</u>	<u>**18.60**</u>	<u>**42.61**</u>
FS-Oracle	51.75	23.31	47.53	53.51	22.94	48.29

Table 4: ROUGE F1 scores on the test sets for the ensemble experiments. We compare summary predictions given by the concatenation of PEGASUS and BigBird summaries against summaries derived from FACTORSUM summary views. We use two sentence rankers: an unsupervised TextRank baseline and FACTORSUM extrinsic importance ranker. FS and FS-oracle use FACTORSUM without content guidance and with reference summary guidance, respectively. All models use *fixed budget* as described in Section 4.1. Best non-oracle results are **bold-faced**. <u>Underlined results</u> are statistically equivalent to the best scores ($p < 0.05$).

tuition is that FACTORSUM should adapt well to OOD budget distributions, whereas the intrinsic model captures domain-specific patterns with less influence from length and content position noise.

The adaptation performance for similar domains (PubMed and arXiv) is much higher than summarizing GovReport documents when trained on scientific articles. FACTORSUM outperforms end-to-end baselines in all cases, especially when there is a large gap in average summary lengths between the domains. However, it can still achieve significant improvements on arXiv, for which all models output summaries with similar average lengths.

Our most important finding is that the ROUGE scores are not as severely affected as expected in this cross-domain setting. Notably, FACTORSUM trained on PubMed without content guidance achieves 44.61 ROUGE-1 on arXiv, outperforming PEGASUS trained in-domain. When guided by BigBird summaries (also OOD), FACTORSUM scores 46.29 ROUGE-1 on arXiv, also outperforming BigBird trained in-domain. Similar results are observed for models trained on arXiv and evaluated on PubMed. On GovReport, FACTORSUM can produce much longer summaries that cover more

relevant content, which explains the substantial improvements in ROUGE scores over the end-to-end baselines. However, summaries generated by FACTORSUM trained on arXiv/PubMed cannot match the average length of 650 words produced by the in-domain version. The reason for this gap is that OOD models generate summary views with a lower variety of content, which are eliminated by the redundancy control described in Section 2.1.2.

5 Related Work

Ranking and aggregating candidate summaries Recent work explores the idea of sampling summary candidates that are scored to form a final summary. The SimCLS (Liu and Liu, 2021) model generates several summary candidates using diverse beam search (Vijayakumar et al., 2016), which are ranked according to a learned evaluation function. Also, the Perturb-and-Select summarizer (Oved and Levy, 2021) uses similar ideas in a multi-document opinion summarization task. Instead of a diverse beam search, it performs random perturbations in the model inputs to generate candidates ranked according to a coherence model. Iso et al. (2021) presented COOP, a framework that improves the aggregation methods for multi-document opinion summarization representations by maximizing their input-output word overlap. Our approach differs from these models as our sampled summary views are not full summary candidates but partial views with important semantic content from the original document. Also, the SimCLS ranking uses intrinsic importance only (similarily with respect to the original document).

Divide-and-conquer approaches for long document summarization The DANCER model of Gidiotis and Tsoumakas (2020) breaks a summarization task into multiple smaller sequence-to-sequence sub-tasks, which share similar motivation to our intrinsic importance model. However, they use several heuristics to select specific sections of the papers (introduction, methods, results, and conclusion) while our sampling approach is domain agnostic. Recently, Mao et al. (2021) proposed DYLE, an *extract-then-generate* method that extract text snippets from chunks of the input document, obtaining state-of-the-art results on GovReport. Zhang et al. (2021) presented a multi-stage summarization method that generates coarse summaries for each document segment, which are then used to obtain a fine-grained summary. Cao and

Wang (2022) add a learnable hierarchical bias term to the transformer attention mechanism, which allows the model to capture information about the structure (sections) of long documents.

Controllable summarization Controlling length in summaries has been addressed by leveraging positional encodings (Takase and Okazaki, 2019), a length-aware attention mechanism (Liu et al., 2022), and optimization objectives that include a length constraint (Makino et al., 2019). Kikuchi et al. (2016) explore different length control techniques at the learning and decoding stages. Control of other attributes such as entity coverage and summary style were achieved with control tokens (Fan et al., 2018; He et al., 2020) and constrained Markov Decision Processes (Chan et al., 2021).

Similar to this work, GSum (Dou et al., 2021) uses content guidance to improve summary quality. However, we note that GSum's guidance is used as *input* of its sequence-to-sequence model, and shifts in guidance distribution would require further training. In contrast, FactorSum allows one to change the budget or content guidance without expensive retraining. Regarding evaluation, the best GSum variant achieves 45.09 ROUGE-1 on PubMed, whereas FactorSum achieves 45.41 R-1 without content guidance and 47.5 R-1 with content guidance. Finally, our sequence-to-sequence architecture is based on BART-base, thus requiring significantly fewer training parameters than GSum's dual BART-large encoders.

6 Conclusion and Future Work

In this work, we embrace the idea that general-purpose summary is an elusive goal and that contextual factors such as preferences for summary lengths are essential in the design of summarization systems (Jones et al., 1999). We propose a framework to separate budget decisions from selecting important content in the document using neural sequence-to-sequence models as building blocks. Our results suggest improved performance in both in-domain and cross-domain summarization of long documents. In future work, we plan to investigate the effects of domain-specific extrinsic guidance that would encourage the summaries to cover aspects of interest.

Limitations

While our model requires fewer compute resources for training, the inference step is more expensive.

For each document, $n_d = 20$ document views are sampled and n_d feedforward computations are performed for the intrinsic model (BART-base) before the greedy summary generation algorithm is applied. In our experiments, BART-base averages $0.27 \times n_d$ seconds per sample. The greedy generation adds in the worst case an average of 1.5 seconds per sample using a naive single-threaded implementation, which gives a total of 6.9 seconds per document. Fortunately, these computations are highly parallelizable, and more careful tuning of the number of views per document n_d would make the runtime similar to a single large neural model. For comparison, PEGASUS and BigBird-PEGASUS take on average 3.13 and 3.85 seconds per sample on a single GPU (batch size = 4).

Ethical Considerations

In Section 6, we discussed how our model requires more compute power during inference, despite being relatively cheaper to train. Thus, if similar methods are used as part of a service in a production environment that serves a large number of requests, the inference procedure may result in unnecessary use of compute resources. In this case, we recommend careful tuning of the document sampling parameters n_d and s_f (Section 2.1) to keep the number of summary views adequate for the appliceeds. In Appendix A, we provide some heuristics to optimize those sampling parameters.

Acknowledgements

References

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

Shuyang Cao and Lu Wang. 2022. HIBRIDS: Attention with Hierarchical Biases for Structure-aware Long Document Summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 786–807, Dublin, Ireland. Association for Computational Linguistics.

Hou Pong Chan, Lu Wang, and Irwin King. 2021. Controllable summarization with constrained Markov decision process. *Transactions of the Association for Computational Linguistics*, 9:1213–1232.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. 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 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Yue Dong, Andrei Mircea, and Jackie Chi Kit Cheung. 2021. Discourse-aware unsupervised summarization for long scientific documents. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1089–1102, Online. Association for Computational Linguistics.

Zi-Yi Dou, Pengfei Liu, Hiroaki Hayashi, Zhengbao Jiang, and Graham Neubig. 2021. GSum: A general framework for guided neural abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4830–4842, Online. Association for Computational Linguistics.

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.

Alexios Gidiotis and Grigorios Tsoumakas. 2020. A divide-and-conquer approach to the summarization of long documents. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:3029–3040.

Junxian He, Wojciech Kryściński, Bryan McCann, Nazneen Rajani, and Caiming Xiong. 2020. CTRLsum: Towards Generic Controllable Text Summarization.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *ArXiv preprint*, abs/1503.02531.

Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. 2021. Efficient attentions for long document summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1419–1436, Online. 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.

K Sparck Jones et al. 1999. Automatic summarizing: factors and directions. *Advances in automatic text summarization*, pages 1–12.

Yuta Kikuchi, Graham Neubig, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. 2016. Controlling output length in neural encoder-decoders. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1328–1338, Austin, Texas. Association for Computational Linguistics.

Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and F Huang. 2006. A tutorial on energy-based learning. *Predicting structured data*, 1(0).

Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.

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.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yixin Liu and Pengfei Liu. 2021. SimCLS: A simple framework for contrastive learning of abstractive summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1065–1072, Online. Association for Computational Linguistics.

Yizhu Liu, Qi Jia, and Kenny Zhu. 2022. Length Control in Abstractive Summarization by Pretraining Information Selection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6885–6895, Dublin, Ireland. Association for Computational Linguistics.

Takuya Makino, Tomoya Iwakura, Hiroya Takamura, and Manabu Okumura. 2019. Global optimization under length constraint for neural text summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1039–1048, Florence, Italy. Association for Computational Linguistics.

Ziming Mao, Chen Henry Wu, Ansong Ni, Yusen Zhang, Rui Zhang, Tao Yu, Budhaditya Deb, Chenguang Zhu, Ahmed H. Awadallah, and Dragomir Radev. 2021. DYLE: Dynamic Latent Extraction for Abstractive Long-Input Summarization.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

Nadav Oved and Ran Levy. 2021. PASS: Perturb-and-select summarizer for product reviews. 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 351–365, Online. Association for Computational Linguistics.

Maxime Peyrard. 2019. A simple theoretical model of importance for summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1059–1073, Florence, Italy. Association for Computational Linguistics.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2019. Zero: Memory optimizations toward training trillion parameter models. ArXiv.

Simeng Sun, Ori Shapira, Ido Dagan, and Ani Nenkova. 2019. How to compare summarizers without target length? pitfalls, solutions and re-examination of the neural summarization literature. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 21–29, Minneapolis, Minnesota. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

Sho Takase and Naoaki Okazaki. 2019. Positional encoding to control output sequence length. 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 3999–4004, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontañón, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

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, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.

Yusen Zhang, Ansong Ni, Ziming Mao, Chen Henry Wu, Chenguang Zhu, Budhaditya Deb, Ahmed H. Awadallah, Dragomir Radev, and Rui Zhang. 2021. Summ^N: A Multi-Stage Summarization Framework for Long Input Dialogues and Documents.

A Document Sampling Experiments

The sampling factor (s_f) and number of samples per document (n_d) are important hyperparameters that affect the summarization performance and computation costs of FACTORSUM. By increasing the number of samples per document, the coverage of oracle sentences is also increased. Also, a smaller sampling factor will make each document view shorter and less prone to input truncation, at the cost of oracle sentence coverage.

For our experimental setup in Sections 3 and 4, we pick the sampling factor so that the number of sentences/tokens fit BART input limit (1024 tokens) with minimal truncation. The number of samples per document is chosen so that the coverage of oracle sentences in the original article is close to 100%, while keeping the resulting dataset size and training costs manageable. According to Table 5, a sampling factor $s_f = 0.2$ and number of samples $n_d = 20$ fulfil the requirements above for all datasets.

Sampling factor (s_f)	Oracle coverage (%)			Average sentences
	$n_d = 5$	$n_d = 10$	$n_d = 20$	
PubMed				
0.1	38.9	65.1	87.6	8.4
0.2	66.7	90.2	99.1	17.2
0.25	77.0	95.6	100	21.6
0.33	87.9	99.7	100	29.0
0.5	98.3	100	100	43.8
arXiv				
0.1	37.9	63.8	88.1	27.6
0.2	64.6	88.4	99.0	55.7
0.25	75.6	93.3	99.7	69.8
0.33	86.8	97.7	100	93.2
0.5	96.5	99.9	100	140.0
GovReport				
0.1	39.6	65.2	88.5	29.8
0.2	65.4	88.7	99.0	60.2
0.25	76.0	93.4	99.7	75.4
0.33	86.4	97.5	100	100.7
0.5	96.6	100	100	151.3

Table 5: Oracle sentence coverage and average number of sentences in sampled documents (validation sets) for different configurations of sampling factor s_f and samples per document n_d.
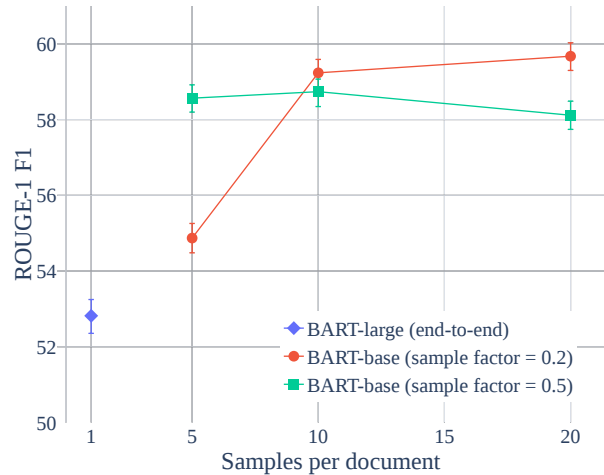
To further investigate the effects of different sam-

Figure 2: ROUGE-1 (F1) scores for different values of sampling factor (s_f) and number of samples per document (n_d), evaluated on the GovReport test set. BART-large is an end-to-end baseline, which is equivalent to $n_d = 1$ and $s_f = 1$.

pling strategies, we train FACTORSUM versions with sampling factor $s_f \in \{0.5, 0.2\}$ and samples per document $n_d \in \{5, 10, 20\}$. In Figure 2, we report evaluation results on the GovReport test set, which confirm the importance of oracle sentence coverage for the final summarization performance. Specifically, for a fixed $n_d = 5$, the model with $s_f = 0.2$ (65.4% oracle coverage) achieves 54.87 ROUGE-1 versus 58.57 ROUGE-1 for the version $s_f = 0.5$ (96.6% oracle coverage). A model with the same sampling factor $s_f = 0.2$ but using more samples per document achieves up to 59.67 ROUGE-1 ($n_d = 5$, 99.1% oracle coverage).

Finally, we observe that lower sampling factors achieve higher ROUGE scores, specially compared to the BART-large end-to-end baseline ($s_f = 1$, $n_d = 1$). These results suggest that working with shorter inputs is beneficial for summarization. We believe that this difference in performance is due to less truncation of the inputs.

B Evaluation Results for Varying Budgets

In Figure 3, we provide results for varying budget guidance values on PubMed, arXiv, and GovReport test sets. For all datasets, there is a consistent improvement for content-guided summaries versus FACTORSUM without content guidance. We also note that BigBird content guidance leads to significant improvements on PubMed and arXiv but BART-large is statistically equivalent to Lead guidance on GovReport, which means there is still room for improvement in our end-to-end BART-

Dataset	Docs	Summary Views	Steps	Hours
PubMed	115k	2.3M	100k	91
arXiv	200k	4M	200k	180
GovReport	17.5k	340k	50k	51

Table 6: Statistics for number of documents, summary views, training steps, and total training time for each dataset (intrinsic model based on BART-base).

large baseline.

C Training Details

The intrinsic importance model $p_\theta(R_v^{(i)}|D_v^{(i)})$ described in Section 2.1.1 is implemented using the BART sequence-to-sequence model (Lewis et al., 2020). We fine-tune the `bart-base` checkpoint from huggingface[6] on the datasets of document and summary views $\mathcal{T}'$ presented in Section 2.1. Unless otherwise stated, we use $n_d = 20$ samples per document and a sampling factor $s_f = 0.2$, as explained in Appendix A. To ensure replicability, we use a random seed for document views sampling.

For the training process, we use 4 GeForce GTX 1080 Ti GPUs each with 12GB of memory. Table 6 details the training set size (number of documents and summary views), number of training steps, and time to train the intrinsic importance models for each dataset. The main training hyperparameters are presented in Table 7.

End-to-end summarization We train our own end-to-end BART-large baseline on the GovReport dataset. Since the target summaries are long, we the maximum summary generation length to 768 tokens, which makes the memory requirements to exceed most single-GPU capacities (even with batch size equal to one). To address this problem, we resort to model parallelism techniques provided by the DeepSpeed library (Rajbhandari et al., 2019), allowing the efficient distribution of the model across 4 GeForce GTX 1080 Ti GPUs, each with 12GB of memory. We use gradient accumulation to achieve a effective batch size of 32. Table 7 details the training hyperparameters.

D Inference Details

For reproducibility purposes, we provide the generation details for the end-to-end baseline models

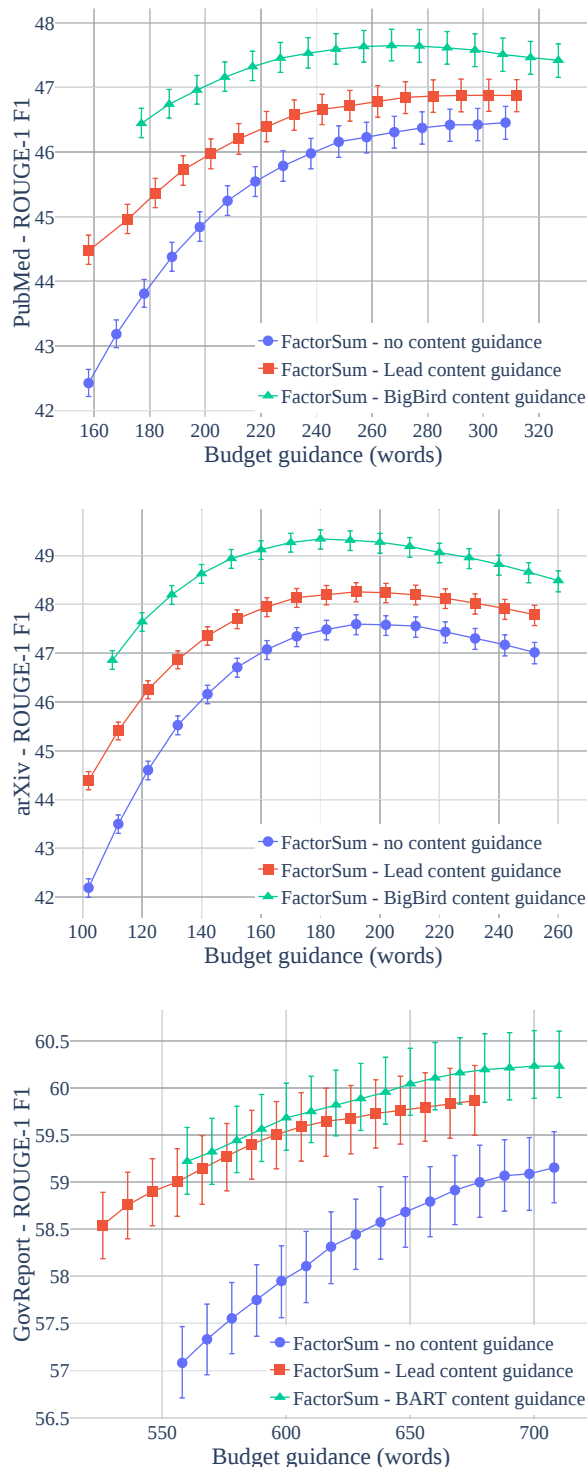[6] https://huggingface.co/facebook/bart-base

Figure 3: ROUGE-1 (F1) scores for different summary budget and content guidance computed on PubMed (top), arXiv (middle), and GovReport (bottom) test sets. Error bars indicate 95% confidence interval.

Parameter	BART base	BART large
Checkpoint	base	large
Number of parameters	139.4M	406.3M
Training time	Table 6	224 hours
Batch size per GPU	4	1
Gradient accumulation	4	32
Effective batch size	64	32
Learning rate	5×10^{-5}	2×10^{-5}
Learning rate scheduler	linear	
Optimizer	AdamW	
Adam β_1	0.9	
Adam β_2	0.999	
Adam ϵ	1×10^{-8}	
Metric best model	ROUGE-1 F1	
Floating point precision	FP16	
Max source length	1024	
Max target length	128	768
Generation beams	4	8
Length penalty	1.0	0.6
DeepSpeed ZeRO[7]	-	Stage 1

Table 7: Training details and hyperparameters for the intrinsic model (BART-base) and the end-to-end baseline for GovReport (BART-large).

in Table 8. The intrinsic importance model (BART-base) generation uses the same maximum source length, maximum target length, beam size, and length penalty defined for training in Table 7.

Extrinsic importance model For summary length control, we adjust the budget guidance so that the average summary lengths is close to the average length of the first 1,000 summaries from the validation set (or the entire validation set for GovReport). The budget guidance for each model/guidance type is shown in Table 9. For the domain adaptation results in Table 3, the budget guidance corresponds to the target dataset, i.e., it is the in-domain budget listed in Table 9.

In addition to the inference procedure described in Algorithm 1, we apply a preprocessing step that divides each summary view $S_v \in V_D$ into sentences. The resulting sentence-tokenized set of summary views V_D is the union of all sentences. We use the sentence tokenizer provided by NLTK[8].

For FACTORSUM versions using content guidance, the best summary S^* returned by Algorithm 1 is reordered according to the following procedure:

[8] nltk.org

PEGASUS	
Checkpoint (arXiv)	pegasus-arxiv
Checkpoint (PubMed)	pegasus-pubmed
Number of parameters	570.8M
Max source length	1024
Generation beams	8
Length penalty	0.8

BigBird	
Checkpoint (arXiv)	bigbird-pegasus-large-arxiv
Checkpoint (PubMed)	bigbird-pegasus-large-pubmed
Number of parameters	576.9M
Max source length	3072
Generation beams	5
Length penalty	0.8

BART-large	
Checkpoint (GovReport)	See Appendix C
Number of parameters	406.3M
Max source length	1024
Generation beams	8
Length penalty	1.0

All models	
Max target length (arXiv, PubMed)	256
Max target length (GovReport)	768

Table 8: Summary generation details and parameters for the end-to-end baselines.

Guidance Budget Content		Pubmed	arXiv	GovReport
FACTORSUM - no content guidance				
Oracle	-	216	169	656
Fixed	-	213	167	656
Model	-	217	170	698
FACTORSUM - content guidance				
Oracle	Lead	221	169	632
Fixed	Lead	217	167	624
Fixed	Model	232	175	658
Model	Model	227	177	658

Table 9: Budget guidance used for FACTORSUM models in Table 2. Model guidance is provided by BART-large for GovReport and BigBird for PubMed and arXiv.

Ranker	PubMed			arXiv		
	R-1	R-2	R-L	R-1	R-2	R-L
PEGASUS + BigBird Ensemble Summaries						
TextRank	44.02	18.40	38.73	44.20	16.97	37.66
FS	44.83	**19.12**	40.90	44.77	17.36	39.84
FS-Oracle	48.77	21.72	44.40	49.18	20.07	43.72
Summary Views						
TextRank	42.17	16.82	37.62	42.49	16.39	37.58
FS	**45.33**	18.69	**41.62**	**47.16**	**18.57**	**42.57**
FS-Oracle	51.64	23.27	47.48	53.27	22.75	48.08

Table 10: ROUGE F1 scores on the **validation sets** for the ensemble experiments. We compare summary predictions given by the concatenation of PEGASUS and BigBird summaries against summaries derived from FACTORSUM summary views. We use two sentence rankers: an unsupervised TextRank baseline and FACTORSUM extrinsic importance ranker. FS and FS-oracle use FACTORSUM without content guidance and with reference summary guidance, respectively. All models use *fixed budget* as described in Section 4.1. Best non-oracle results are **bold-faced**. Underlined results are statistically equivalent to the best scores ($p < 0.05$).

(1) for each summary view in S^*, we collect the index of the oracle sentence in the content guidance text[9]; (2) The summary views are sorted according to the list of corresponding oracle indexes, using the Python `sorted` function[10].

E   Validation results

In Table 11, we provide validation scores corresponding to the in-domain summarization test results in Table 2. Additionally, Table 10 shows the validation scores for the ensemble experiments corresponding to the test results in Table 4.

F   Sample Summaries

In Tables 12 to 19, we provide samples of summaries from PubMed, arXiv, and GovReport test sets. We compare BigBird and BART summaries to FACTORSUM with lead and model-provided content guidance. Budget guidance is fixed to the same values as described in Section 4.

---

[9]Oracle sentences determined as described in Section 2.1.
[10]https://docs.python.org/3/library/functions.html#sorted

Model	PubMed				arXiv				GovReport			
	R-1	**R-2**	**R-L**	**Len**	**R-1**	**R-2**	**R-L**	**Len**	**R-1**	**R-2**	**R-L**	**Len**
Previous work												
PEGASUS	43.73	18.77	40.15	181	43.07	16.39	38.66	170	-	-	-	-
BigBird	45.28	19.77	41.60	186	46.02	18.54	41.35	164	-	-	-	-
BART-large	-	-	-	-	-	-	-	-	53.06	19.11	50.12	597
Guidance												
Budget **Content**	**FACTORSUM - no content guidance**											
Oracle -	47.37	19.16	43.33	210	48.85	18.85	43.92	165	59.54	23.98	55.82	642
Fixed -	45.33	18.69	41.62	205	47.16	18.57	42.57	165	58.41	23.90	54.83	650
Model -	44.71	18.07	40.84	185	46.35	18.22	41.79	165	57.55	23.68	53.92	639
FACTORSUM - content guidance												
Oracle Lead	48.19	19.99	44.25	205	49.61	19.27	44.75	164	60.47	24.93	56.89	649
Fixed Lead	46.10	19.21	42.43	202	47.98	18.99	43.41	165	<u>59.19</u>	<u>24.72</u>	<u>55.78</u>	649
Fixed Model	**<u>47.20</u>**	**<u>20.17</u>**	**<u>43.48</u>**	197	**<u>49.16</u>**	**<u>20.17</u>**	**<u>44.59</u>**	164	**60.00**	**25.33**	**56.49**	648
Model Model	<u>47.00</u>	<u>20.15</u>	<u>43.19</u>	179	48.17	<u>20.06</u>	44.07	164	<u>59.35</u>	25.25	55.84	641

Table 11: ROUGE F1 scores and average words per summary on the **validation sets** for different types of guidance during inference. *Lead* guidance is the first k sentences from the source document (Section 4.2). Model guidance is provided by BART-large for GovReport and BigBird for PubMed and arXiv. The choice of budget guidance values is described in Appendix D. Results for models marked with † are taken from the original publications. Underlined results are statistically equivalent to the best methods ($p < 0.05$).

Reference abstract (Tokens: 156)
we report a 2.5-year - old girl who presented with hoarseness of voice since 3 months of age and failure to thrive . chest x - ray showed cardiomegaly with a deviation of the trachea and mediastinum to the right side . two - dimensional echocardiography showed decreased flow across the right pulmonary artery , a small atrial septal defect (asd) with a right - to - left shunt , and a dilated right atrium and right ventricle with severe tricuspid regurgitation suggestive of severe pulmonary hypertension . a silent large patent ductus arteriosus was also seen . multiple detector computerized tomography aortogram confirmed the findings of absent right pulmonary artery and hypoplastic right lung with small cystic lesions suggestive of congenital cystic adenomatoid malformation in the right lower lobe .

BigBird summary Tokens: 48; ROUGE-1: 45.16; ROUGE-2: 31.52; ROUGE-L: 39.78
we report a 2.5-year - old girl with congenital absence of the right pulmonary artery with associated congenital cystic adenomatoid malformation of the right lower lobe, patent ductus arteriosus, and atrial septal defect, who presented with ortner's syndrome due to severe pulmonary hypertension. .

FACTORSUM - Fixed budget and Lead content guidance Tokens: 199; ROUGE-1: 47.17; ROUGE-2: 26.58; ROUGE-L: 44.65
unilateral absence of pulmonary artery (uapa) is a rare congenital abnormality , with an estimated prevalence of 1 in 200,000. while some patients with uapa are totally asymptomatic , others may have severe pulmonary hypertension. we report a 2.5-year - old girl with congenital absence of the right pulmonary artery with associated congenital cystic adenomatoid malformation (ccam) , patent ductus arteriosus (pda) , and atrial septal defect , who presented with ortner 's syndrome due to severe pulmonary hypertension. a 2.5-year - old girl presented with hoarseness of voice noticed since 3 months of age , breathlessness for the past 15 days and failure to thrive. pulmonary agenesis is usually unilateral , right sided absence of pulmonary artery being more common. the syndrome has since been described in adults with various cardiovascular disorders , but reports in children are less common. congenital cystic adenomatoid malformation (ccam) is a rare cause of congenital cyanotic heart disease. the child was intubated and ventilated and started on pressors , but sustained a cardiac arrest on the 4th hospital day from which she could not be resuscitated.

FACTORSUM - Fixed budget and BigBird content guidance Tokens: 142; ROUGE-1: 42.42; ROUGE-2: 23.66; ROUGE-L: 38.64
congenital cystic adenomatoid malformation (ccam) is a rare cause of congenital cyanotic heart disease. pulmonary agenesis is usually unilateral , right sided absence of pulmonary artery being more common. the syndrome has since been described in adults with various cardiovascular disorders , but reports in children are less common. unilateral absence of pulmonary artery (uapa) is a rare congenital abnormality , with an estimated prevalence of 1 in 200,000. we report a 2.5-year - old girl with congenital absence of the right pulmonary artery with associated congenital cystic adenomatoid malformation (ccam) , patent ductus arteriosus (pda) , and atrial septal defect, who presented with ortner 's syndrome due to severe pulmonary hypertension. while some patients with uapa are totally asymptomatic , others may have severe pulmonary hypertension.

Table 12: Sample abstract and generated summaries from the PubMed test set (ID = 5836).

Reference abstract (Tokens: 240)

the navier - stokes - fourier theory of viscous , heat - conducting fluids provides parabolic equations and thus predicts infinite pulse speeds . naturally this feature has disqualified the theory for relativistic thermodynamics which must insist on finite speeds and , moreover , on speeds smaller than c. the attempts at a remedy have proved heuristically important for a new systematic type of thermodynamics : extended thermodynamics . that new theory has symmetric hyperbolic field equations and thus it provides finite pulse speeds.extended thermodynamics is a whole hierarchy of theories with an increasing number of fields when gradients and rates of thermodynamic processes become steeper and faster . the first stage in this hierarchy is the 14-field theory which may already be a useful tool for the relativist in many applications . the 14 fields and further fields are conveniently chosen from the moments of the kinetic theory of gases.the hierarchy is complete only when the number of fields tends to infinity . in that case. the pulse speed of non - relativistic extended thermodynamics tends to infinity while the pulse speed of relativistic extended thermodynamics tends to c , the speed of light. [...]

BigBird summary Tokens: 105; ROUGE-1: 37.27; ROUGE-2: 13.75; ROUGE-L: 32.29

the paradox of pulse speeds in extended thermodynamics has been known for 50 years. it seems to have been caused by eckart s theory of irreversible processes which assumed a constant pulse speed and a fixed temperature. recently it has been shown by boillat & ruggeri that, as the number of moments increases, the pulse speed tends to infinity in the non - relativistic kinetic theory of gases and the relativistic case by which the pulse speed tends to c. these results put an end to the long - standing paradox of pulse speeds. they are reviewed in detail.

FACTORSUM - fixed budget and Lead content guidance Tokens: 216; ROUGE-1: 50.12; ROUGE-2: 14.18; ROUGE-L: 44.77

the pulse speed problem is one of the most important questions in thermodynamics , but it is a question that can be answered , and has to be answered. we derive a set of thermodynamic processes for a , a , b , c , and d. in this paper , we prove that a co - vector exists in which the entropy is a function of the thermal equation of the state. the thermodynamics of viscous , heat - conducting gases is studied by means of the determination of the 14 fields of the field equations. in this paper , we review the recent developments in the field of non - equilibrium thermodynamics. it is possible , and indeed common , to make a specific choice for the fields u and the concavity postulate is contingent upon that choice. the first moments in the kinetic theory of gases are obtained from a homogeneous system where the acceleration waves and their speeds of propagation are to be calculated from the homogeneous systems. the heat fluxes f(x , p , t) of the atoms , viz. the paper deals with the thermodynamics of a non - degenerate gas. in the non - relativistic limit ,

FACTORSUM - fixed budget and BigBird content guidance Tokens: 190; ROUGE-1: 44.67; ROUGE-2: 12.75; ROUGE-L: 42.64

it is then a simple problem of linear algebra to prove that the entropy density h = ha is concave as a function of f. the pulse speed problem is one of the most important questions in thermodynamics , but it is a question that can be answered , and has to be answered. it is possible , and indeed common , to make a specific choice for the fields u and the concavity postulate is contingent upon that choice. the first moments in the kinetic theory of gases are obtained from a homogeneous system where the acceleration waves and their speeds of propagation are to be calculated from the homogeneous systems. we derive a set of thermodynamic processes for a , a , b , c , and d. in this paper , we review the recent developments in non - relativistic kinetic theory. this paper presents the results of an investigation of the non - degenerate gases in the system. we conclude that no paradox of infinite speeds can arise in extended thermodynamics. in the non - relativistic limit ,

Table 13: Sample abstract and generated summaries from the PubMed test set (ID = 1475).

Reference abstract (Tokens: 157)

the metrization of the space of neural responses is an ongoing research program seeking to find natural ways to describe , in geometrical terms , the sets of possible activities in the brain . one component of this program are the _ spike metric , notions of distance between two spike trains recorded from a neuron . alignment spike metrics work by identifying " equivalent " spikes in one train and the other . we present an alignment spike metric having @xmath0 underlying geometrical structure ; the @xmath1 version is euclidean and is suitable for further embedding in euclidean spaces by multidimensional scaling methods or related procedures . we show how to implement a fast algorithm for the computation of this metric based on bipartite graph matching theory . @xmath2center for studies in physics and biology , rockefeller university , new york ny @xmath3harvard faculty of arts and sciences , cambridge ma

BigBird summary   Tokens: 89; ROUGE-1: 28.71; ROUGE-2: 3.00; ROUGE-L: 22.77

we propose a spike metric that is consistent with the time - coding hypothesis of spike generation and has all of the desirable properties of an @xmath0 norm . when @xmath1 , this metric is equal to the victor - purpura metric @xcite . when @xmath2 , this metric is equal to the hungarian algorithm @xcite , @xcite , @xcite . when @xmath3 , this metric is equal to the monge algorithm @xcite , @xcite , @xcite .

FACTORSUM - **Fixed budget and Lead content guidance**   Tokens: 165; ROUGE-1: 42.51; ROUGE-2: 6.32; ROUGE-L: 38.33s

we consider the problem of finding a minimum weight matching on a bipartite graph such that each edge of the graph has a weight or cost. the value of the metric between two spike trains is the @xmath0 norm of the difference between their estimated rate functions . this metric preserves the integrity of individual spikes instead of viewing them as contributions to a rate function . we present a spike metric satisfying two important desiderata : that it be grounded in the time - coding hypothesis of spike generation , and that it is closely related to the euclidean @xmath1 norm . our proof uses the tools of graph theory and combinatorial optimization , and demonstrates that they can be usefully brought to bear on problems in computational biology. the cost of aligning two spikes on different trains is @xmath0. we present a new metric for comparing two spike trains , which we call the hungarian metric.

FACTORSUM - **Fixed budget and BigBird content guidance**   Tokens: 150; ROUGE-1: 41.61; ROUGE-2: 5.88; ROUGE-L: 38.69

we present a spike metric satisfying two important desiderata : that it be grounded in the time - coding hypothesis of spike generation , and that it is closely related to the euclidean @xmath1 norm . the value of the metric between two spike trains is the @xmath0 norm of the difference between their estimated rate functions . the hungarian algorithm is a special case of algorithms to solve general matching problems and assignment problems. the metric is equal to the sum of the costs incurred by the most efficient alignment. we consider the problem of finding a minimum weight matching on a bipartite graph such that each edge of the graph has a weight or cost. we present a new metric for comparing two spike trains , which we call the hungarian metric. our metric is a minimization over all possible matchings .

Table 14: Sample abstract and generated summaries from the arXiv test set (ID = 5946).

Reference abstract (Tokens: 145)
in this paper we discuss the chemical evolution of elliptical galaxies and its consequences on the evolution of the intracluster medium (icm) . we use chemical evolution models taking into account dark matter halos and compare the results with previous models where dark matter was not considered . in particular , we examine the evolution of the abundances of some relevant heavy elements such as oxygen , magnesium and iron and conclude that models including dark matter halos and an initial mass function (imf) containing more massive stars than the salpeter (1955) imf , better reproduce the observed abundances of mg and fe both in the stellar populations and in the icm (asca results) . we also discuss the origin of gas in galaxy clusters and conclude that most of it should have a primordial origin .

BigBird summary Tokens: 109; ROUGE-1: 36.21; ROUGE-2: 9.57; ROUGE-L: 28.45
we discuss a model for the chemical evolution of elliptical galaxies in which supernovae (sne)-driven galactic winds play an important role in the formation of these objects . in this model the star formation is assumed to stop after the occurrence of the galactic wind and the galaxy evolves passively thereafter . the star formation is assumed to stop after the occurrence of the galactic wind and the galaxy evolves passively thereafter . the model includes the most recent ideas on sn progenitors and nucleosynthesis , indicating that sne ia originate from long living stars whereas sne of type ii originate from short living stars .

FACTORSUM - **fixed budget and Lead content guidance** Tokens: 159; ROUGE-1: 43.73; ROUGE-2: 12.27; ROUGE-L: 39.43
this is a very strong conclusion since it implies a very fast process for the formation of big ellipticals at variance with the hierarchical clustering scenario for galaxy formation. we show how abundance ratios in stellar populations and gas in ellipticals can be used to constrain the amount and concentration of dark matter in these objects. in order to reproduce realistic galaxies , namely with the right colors and luminosities. we discuss the chemical evolution of elliptical galaxies in the framework of a simple model based on the idea that the efficiency of star formation should be inversely proportional to the dynamical timescale. we find that the efficiency of star formation in elliptical galaxies is inversely proportional to the dynamical timescale. in particular , we show that it is not possible to explain the increase of the [mg / fe] ratio in the nuclei of ellipticals as a function of galactic luminosity.

FACTORSUM - **fixed budget and BigBird content guidance** Tokens: 147; ROUGE-1: 43.70; ROUGE-2: 14.93; ROUGE-L: 39.26
we discuss the chemical evolution of elliptical galaxies in the framework of a simple model based on the idea that the efficiency of star formation should be inversely proportional to the dynamical timescale. in particular , we show that the efficiency of star formation increases with the total mass of the galaxy and that the more massive galaxies develop a galactic wind before the less massive ones. we show that the presence of dark matter in elliptical galaxies plays a crucial role in determining the onset and the entity of galactic winds. in this paper we discuss the possibility of an inverse wind scenario for the formation of elliptical galaxies. the model includes the most recent ideas on sn progenitors and nucleosynthesis , indicating that sne ia originate from long living stars whereas sne of type ii originate from short living stars.

Table 15: Sample abstract and generated summaries from the arXiv test set (ID = 6213).

Reference abstract (Tokens: 474)

Congress frequently faces questions about whether and how to commemorate people and events that have influenced the nation's history. Congress often has chosen to do so by establishing national memorials or by conferring a national designation on existing state, local, or private memorials. The National Park Service (NPS) defines national memorials within the National Park System as "primarily commemorative" works that need not be at sites historically associated with their subjects. The Commemorative Works Act (CWA; 40 U.S.C. §§8901-8910) was enacted to govern the establishment process for memorials located in the District of Columbia (Washington, DC) or its environs that are under the jurisdiction of the NPS or the General Services Administration. The CWA includes provisions related to memorial location, design, construction, and perpetual maintenance. Memorials in Washington, DC, include those with the word national in the name and those that are essentially national memorials but do not bear that title. For memorials outside the District of Columbia, no specific law or set of regulations governs their establishment. Congress has established a number of federally administered national memorials throughout the nation, most often as units of the National Park System but also under management of other federal agencies. Various nonfederal entities undertaking commemorative efforts also have petitioned Congress for assistance or statutory recognition, and some individual memorial organizers have titled their works as national memorials without congressional recognition. To clarify options for Congress when considering commemoration of individuals, groups, and events through memorials, this report discusses several types of congressional involvement in memorials outside the District of Columbia. For purposes of the report, these are characterized as high federal involvement (e.g., congressional establishment of a national memorial under federal agency administration); medium federal involvement (e.g., congressional authorization for a memorial to be located on federal property or to receive federal funds); low federal involvement (e.g., statutory recognition without additional federal support); and no federal involvement (e.g., a self-declared national memorial). The report provides examples of memorials of each type and discusses some options for Congress, with regard to both individual memorial designations and consideration of whether to systematize criteria for memorials outside Washington, DC, similar to the CWA's provisions for District of Columbia memorials. Because this report focuses specifically on memorials outside the District of Columbia, please see CRS Report R41658, Commemorative Works in the District of Columbia: Background and Practice, by Jacob R. Straus, for discussion of memorials governed by the CWA in Washington, DC, and its environs.

Table 16: Sample abstract from the GovReport test set (ID = 681). Model predictions are presentend in the next pages.

BART-large summary
Tokens: 390; ROUGE-1: 56.11; ROUGE-2: 19.38; ROUGE-L: 54.23

Virtually all societies attempt to remember and memorialize individuals, groups, and events as part of the preservation of shared rhetoric and history. In the United States, there are hundreds, and possibly thousands, of commemorative works to commemorate such events. Decisions about which places to commemorate are made by many different entities, including Congress, federal agencies, state and local governments, and private citizens, among others. Some memorials receive federal involvement, while others do not. For example, for memorials on federal land in the District of Columbia, the Commemorative Works Act (CWA) requires that Congress authorize the creation of a new memorial. No systematic law or set of regulations governs the establishment of memorials outside Washington, DC. However, Congress also has established or recognized numerous memorials nationwide, and some have been designated by the executive branch. Federal agencies may be classified as "high," "medium," "low," or "none." Memorials with high federal involvement typically are located on U.S. federal land; receive federal funds for design, construction, and maintenance; and are managed by federal agencies. These include memorials established by Congress as units of the National Park System or under the administration of another agency. Other memorials with no federal involvement are those that do not receive any direct federal involvement (i.e., memorials designated by Congress but not administered by a federal agency). Congress, executive branch officials, and other interested parties may place plaques, memorials, and similar works at federal sites in remembrance of a person, group, or event. or on nonfederal land. The National Park Service (NPS) and the General Services Administration (GSA) maintain some of the nation's largest memorials. NPS provides assistance to other federal agencies with assistance in managing memorials located on its lands, including NPS-designated "NPS-affiliated areas." Other agencies, primarily the Army and Air Force, have similar relationships with NPS affiliated areas. Congress has appropriated funds to both NPS and GSA to help fund memorials created on the lands they manage.

Table 17: Summary generated by BART for a document from GovReport test set (ID = 681). Reference summary is presented in Table 16.

FACTORSUM - fixed budget and Lead content guidance
Tokens: 548; ROUGE-1: 63.37; ROUGE-2: 24.77; ROUGE-L: 57.30

Beyond these federally endorsed memorials, a wide variety of other entities have established and maintained memorials throughout the country with no federal connection, including some titled as "national memorials,"In the United States, there are hundreds, and possibly thousands, of memorials to various individuals, groups, and events. Decisions about which people, groups, or events to memorialize are made by many different entities, including Congress, federal agencies, state and local governments, and private citizens, among others. For example, the CWA governs the establishment of memorials on federal lands in the District of Columbia, with provisions for the creation, design, construction, and maintenance of such works. In other areas, various laws, regulations, and policies may provide for different groups and governments to decide what should be commemorated and how. For certain types of commemorations, Congress has taken a more systematized approach. No systematic law or set of regulations governs the establishment of memorials outside Washington, DC. ,Some of these memorials include multiple facilities such as a visitor center or kiosk in addition to the primary commemor For example, the George Washington Masonic National Memorial in Alexandria, VA, and the National Memorial for Peace and Justice in Montgomery, AL, are privately established and maintained. In some cases, memorials located outside of the District of Columbia have been called "national" memorials without being so designated by Congress, such as through the establishment of a program to identify nonfederal memorials deserving of a national designation. A distinction is drawn between memorials located within and outside of Washington, DC, because of the exclusive role the CWA gives Congress to authorize new memorials on federal land in the District of Columbia, and the role of federal agencies—primarily the National Park Service (NPS) and the General Services Administration (GSA)—in maintaining District-based memorials once dedicated. This report considers the extent of federal involvement in memorials located outside the District of Columbia. Congress also could potentially consider a program to provide grants to nonfederal entities for constructing and/or maintaining national memorials outside of Washington, DC. While many such works are established without federal involvement, Congress also has established or recognized numerous memorials nationwide, and some have been designated by the executive branch. For purposes of this report, federal involvement in memorials outside the District of Columbia may be classified as "high," "medium," "low," or "none." For example, P. L. Other variations of federal-nonf For a discussion of the process for creating a new NPS unit and associated issues, see CRS Report RS20158, National Park System: Establishing New Units. Legislation designating these national memorials often includes explicit language stating that the memorial is not an NPS unit and that federal funds shall not be provided for the memorial. In some instances, Congress authorizes a memorial to be created on federal land and administered by a federal agency.

Table 18: Summary generated by FACTORSUM with BART content guidance for a document from GovReport test set (ID = 681). Reference summary is presented in Table 16.

FACTORSUM - fixed budget and BART-large content guidance
Tokens: 548; ROUGE-1: 63.37; ROUGE-2: 24.77; ROUGE-L: 57.30

In other areas, various laws, regulations, and policies may provide for different groups and governments to decide what should be commemorated and how. Beyond these federally endorsed memorials, a wide variety of other entities have established and maintained memorials throughout the country with no federal connection, including some titled as "national memorials,"In the United States, there are hundreds, and possibly thousands, of memorials to various individuals, groups, and events. Decisions about which people, groups, or events to memorialize are made by many different entities, including Congress, federal agencies, state and local governments, and private citizens, among others. This report considers the extent of federal involvement in memorials located outside the District of Columbia. For example, the CWA governs the establishment of memorials on federal lands in the District of Columbia, with provisions for the creation, design, construction, and maintenance of such works. No systematic law or set of regulations governs the establishment of memorials outside Washington, DC. In some cases, memorials located outside of the District of Columbia have been called "national" memorials without being so designated by Congress, such as through the establishment of a program to identify nonfederal memorials deserving of a national designation. For purposes of this report, federal involvement in memorials outside the District of Columbia may be classified as "high," "medium," "low," or "none." Legislation designating these national memorials often includes explicit language stating that the memorial is not an NPS unit and that federal funds shall not be provided for the memorial. For a discussion of the process for creating a new NPS unit and associated issues, see CRS Report RS20158, National Park System: Establishing New Units. While many such works are established without federal involvement, Congress also has established or recognized numerous memorials nationwide, and some have been designated by the executive branch. For certain types of commemorations, Congress has taken a more systematized approach. Congress also could potentially consider a program to provide grants to nonfederal entities for constructing and/or maintaining national memorials outside of Washington, DC. A distinction is drawn between memorials located within and outside of Washington, DC, because of the exclusive role the CWA gives Congress to authorize new memorials on federal land in the District of Columbia, and the role of federal agencies—primarily the National Park Service (NPS) and the General Services Administration (GSA)—in maintaining District-based memorials once dedicated. ,Some of these memorials include multiple facilities such as a visitor center or kiosk in addition to the primary commemor For example, the George Washington Masonic National Memorial in Alexandria, VA, and the National Memorial for Peace and Justice in Montgomery, AL, are privately established and maintained. In some instances, Congress authorizes a memorial to be created on federal land and administered by a federal agency. For example, P. L. Other variations of federal-nonf

Table 19: Summary generated by FACTORSUM with BART content guidance for a document from GovReport test set (ID = 681). Reference summary is presented in Table 16. Note that this summary uses the same set of summary views as FACTORSUM with Lead content guidance in Table 18, just changing their presentation order.