Two-step Text Summarization
for Long-form Biographical Narrative Genre

Avi Bleiweiss
BShalem Research
Sunnyvale, CA, USA
avibleiweiss@bshalem.onmicrosoft.com

Abstract

Transforming narrative structure to implicit discourse relations in long-form text has recently seen a mindset shift toward assessing generation consistency. To this extent, summarization of lengthy biographical discourse is of practical benefit to readers, as it helps them decide whether immersing for days or weeks in a bulky book turns a rewarding experience. Machine-generated summaries can reduce the cognitive load and the time spent by authors to write the summary. Nevertheless, summarization faces significant challenges of factual inconsistencies with respect to the inputs. In this paper, we explored a two-step summary generation aimed to retain source-summary faithfulness. Our method uses a graph representation to rank sentence saliency in each of the novel chapters, leading to distributing summary segments in distinct regions of the chapter. Basing on the previously extracted sentences we produced an abstractive summary in a manner more computationally tractable for detecting inconsistent information. We conducted a series of quantitative analyses on a test set of four long biographical novels and showed to improve summarization quality in automatic evaluation over both single-tier settings and external baselines.

1 Introduction

Text summarization is a principal tool for reasoning about narrative structure and foretell the content of a literary novel in a succinct form. Dated four decades back, the earlier seminal work by Lehnert (1981) pursued analytical summarization of narratives, and offered a graphical representation of human-generated plot units. In this graph, plot units are defined as conceptual elements referring to propositions or states that are linked by character relations. To produce a distilled version of the original discourse, a vast amount of information are selectively ignored by the reader. Similarly, traversing the graph identifies complex elements that are central to the story, and thus points of high relevance for summaries, and ones considered peripheral details.

e-summary: (1) It was committed in the presence of slaves, and they of course could neither institute a suit, nor testify against him; and thus the guilty perpetrator of one of the bloodiest and most foul murders goes unwhipped of justice, and uncensured by the community in which he lives. (2) He was, of all the overseers, the most dreaded by the slaves. (3) He was just proud enough to demand the most debasing homage of the slave, and quite servile enough to crouch, himself, at the feet of the master. [...]

a-summary: The guilty perpetrator of one of the bloodiest and most foul murders goes unwhipped of justice, and uncensured by the community in which he lives . He was cruel enough to inflict the severest punishment, artful enough to descend to the lowest trickery.

Table 1: An example of text generation in our two-stage summarization. In the first step, we extract top-ranked sentences from an extended source chapter of a biographical novel with an average length of over 15K tokens. Then, we produce from the extracted summary (e-summary) an order-of-magnitude compressed abstractive summary (a-summary) that faithfully rephrases its predecessor. Shown are the leading three out of ten top ranked relevant sentences for the e-summary.

Recently, the domain of narrative understanding has gained interest of the research community (Piper et al., 2021). A wide array of computational models developed by language technology professionals provided for expressive generative textual-summaries. Presently, the prevailing approach to natural language generation (NLG) tasks, including summarization, is data driven and uses a sequence-to-sequence neural model pretrained on large text

Proceedings of the 4th Workshop on Computational Approaches to Discourse (CODI 2023), pages 145–155
July 13-14, 2023 ©2023 Association for Computational Linguistics

corpora. Our work centers on evaluating the quality of producing summaries from chapters of long-form biographical novels. Unlike fictional narratives that require concatenating chapter summaries due to an inherit progressive plot nature, biographical chapters are relatively context independent and thus more readily manageable individually. Automatic generation of fluent summaries in the literary domain can be useful to complement the short description of a book provided by the author and to a certain extent assist in constructing expert critiques. The work by Berov (2019) demonstrated that a functional unit approach to summarizing computational storytelling can perform at around human level and contribute to better framing. We note that the narrative summarization task— while a rich source of innovation— is by and large untapped.

Pretrained language models based on the Transformer network (Vaswani et al., 2017) have achieved state-of-the-art performance generating fluent summaries from short input text. However, for long documents, model efficiency and summary quality characterized by remaining faithful to the respectful source present a challenge to natural language generation practitioners (Huang et al., 2021; Zhang et al., 2022). To mitigate the severity, NLG research applied both topical and generic approaches to the task of summary generation, distinguishing extractive summarization that produces high lexical overlap between a summary and the source document, and hence tends to be factually consistent. While abstractive summaries are prone to unaligned content that is not obviously inferable from the original text.

One of the more constraining facet of current neural models tasked with producing abstractive summarizations is that the generated text can contain factually incorrect information with respect to the grounding text they are conditioned on. Summary inconsistencies are diverse and may include inversion, also known as negation, incorrect use of an entity that transpires as object swapping, or the introduction of an entity not in the original document, recognized as hallucination. Maynez et al. (2020) conducted a large-scale study and concluded that hallucination is the most critical to the coherence of abstractive summaries, while Cao et al. (2022) developed a detection approach that separates factual from non-factual hallucinations.

The complexity of the summarization task made automatic evaluation particularly challenging. In their recent line of work, Deng et al. (2021) proposed the intuition of information alignment between input and output text, and developed unified and interpretable metrics across a multitude of diverse NLG tasks. Distinctly for generative summaries, they offered effective definitions of relevance and consistency, widely identified as key aspects to characterize generation quality. Supported by robust theoretical grounds, their prevailing definitions strongly correlate with human judgment on how to concisely describe the most salient content in the input document. We adopted their interpretations in our empirical analysis and extended the consistency measure to a chapter-level rather than book-level over our test set of literary novels.

In Table 1, we present an overview of our two-step framework for summary generation. Distinguishing our work from prior research on extract-then-abstract methods, the approach we propose uses Transformer language models end-to-end, and experiments we conducted were run on exceptionally long-form chapters drawn from biographical literary novels. Our main contribution is twofold: (1) a high-quality and sustainable biographical literary dataset with each chapter consisting of its source text paired with both the extractive and abstractive summary constructs, and (2) through extensive experiments on a diverse biographical literary dataset, we demonstrate the effectiveness of our proposed approach and show similarity and consistency results that are exceeding or comparable to external baseline performance. Our biographical dataset is publicly accessible online. [1]

2 Related Work

We briefly survey existing methods that propose multi-stage text summarization systems evaluated on datasets from a broad range of domains.

Ling and Rush (2017) introduced a coarse-to-fine attention model that reads a document hierarchically, using coarse attention to select top-level blocks of text and fine attention to read the tokens of the chosen blocks. Their proposed summarizer scales linearly with the number of top-level chunks and effectively handles long sequences. However, their model performance lagged behind the standard instantiation baseline of the attention function on ROUGE similarity metrics.

Xu and Lapata (2020) proposed a coarse-to-fine modeling framework for extractive summarization

[1] https://github.com/bshalem/bns

applied to query focused multi-document. Their system incorporates a relevance estimator for retrieving textual segments– such as sentences or longer passages associated with a query—an evidence estimator which further isolates segments likely to contain answers to the query, and a centrality estimator which finally selects which segments to include in the summary. Our extractive summary component is resemblant in spirit to their centrality estimator, however we use a Sentence Transformers (SBERT; Reimers and Gurevych, 2019) model to generate contextual sentence embeddings that follows producing a sentence similarity matrix for computing graph centrality based ranking.

Pilault et al. (2020) explored Transformer language models and proposed an extract-then-summarize computational pipeline for long documents. Their model consists of an extractive element comprising a hierarchical neural encoder that outputs sentence representations of either a pointer to input sentences or to the result of sentence classification; and a Transformer language model conditioned on the extracted sentences as well as on either a part of or the entire input document to generate the summary. Their system was shown to outperform several baselines on similarity metrics, however, a discussion on factual correctness and consistency analyses of experimental results appears relatively sparse.

Gidiotis and Tsoumakas (2020) proposed a divide-and-conquer method by splitting the input into multiple segments, summarizing them separately, and combining the summary pieces. Basing on smaller source and target summary pairs that are focused on a specific aspect of the text, results in better alignment and considerable reduction of computation complexity. They used a basic sequence-to-sequence model and incorporated a rotational unit of memory (Dangovski et al., 2019) in its decoder that led to a more stable training and slightly improved F1 similarity scores. Content quality of their generated summaries relies entirely on ROUGE similarity metrics and could benefit from a broader evaluation framework such as offered by Deng et al. (2021).

More recently, Zhang et al. (2022) proposed a multi-stage split-then-summarize framework to generate summaries from long-form documents. Each source text divides into segments, matching each with a subset of target text. A coarse summary is generated for each segment and further concatenated as input to the next stage. After multiple stages of compression and summarization, a final stage produces a fine-grained summary. Their improved performance across baselines renders relatively low bi-gram scores, most likely owing to over-compression of source text.

An effective abstractive text summarization approach that first compresses long input text into a relatively short input sequence, and follows with efficient long-form document finetuning demonstrated comparable performance at a significantly lower computational cost (Choi et al., 2019; Su et al., 2020). Keen on a specific application, Pu et al. (2022) generate movie plots given movie scripts, by applying heuristic evaluation to extract actions and essential dialogues, a representation that reduces the average length of input movie scripts by 66%. Their system outperforms baselines on various automatic metrics.

3 Chapter Summarization

Our summarization task commences with producing an extractive summary from the source text of a book chapter, and follows with generating an abstractive summary from the salient extractive content (Table 1).

3.1 Importance Extraction

Extractive summarization generates text by selecting a subset of sentences in the original document. To this task we applied LexRank (Erkan and Radev, 2004) that computes sentence importance based on eigenvector centrality in a graph representation of sentences. The graph uses a cosine similarity matrix where each entry in the matrix is the similarity between the corresponding sentence pair. Formally, given n sentences in a novel chapter, we use a colon notation $s_{1:n} = (s_1, \dots, s_n)$ to denote the collection of sentences. We used bag-of-words to represent each sentence as a $|V|$-dimensional vector p, where V is the chapter vocabulary. Hence, the similarity matrix $M \in \mathbb{R}^{n \times n}$ contains elements $m_{ij} = \text{sim}(p_i, p_j)$, where $1 \leq i, j \leq n$ and *sim* a similarity function. LexRank hypothesizes that sentences more similar to many other sentences in the book chapter are more central, or salient to the topic. The algorithm further emits the degree centrality of a node in the similarity graph— the count of similar sentences for each sentence.

Our extractive summarization task uses SBERT

(Reimers and Gurevych, 2019). [2] SBERT derives semantically meaningful sentence embeddings that can be compared using cosine-similarity. We chose the distilled RoBERTa (Liu et al., 2019) variant of the BERT (Devlin et al., 2019) model, a pretrained Transformer network (Vaswani et al., 2017) on a paraphrase dataset. This model generates a dense embedding vector for each input sentence, of which we construct a similarity adjacency matrix M that stores a weighted graph of all sentence-pairs. Matrix M is further provided to LexRank for sentence importance ranking. The chapter extractive summary produced thus comprises a collection of top-ranked sentences with a sentence count that is proportional to the chapter text length, and commonly defaults to a defined maximal saliency.

3.2 Factual Abstraction

Extractive summary generation contrasts with abstractive summarization, where the information in the text is rephrased. Consistent with the Transformer architecture, BART (Lewis et al., 2020), considered a state-of-the-art model for the task of abstractive summarization, introduced denoising autoencoding objectives to pretrain sequence-to-sequence models. As a result, input texts are corrupted in two ways: (1) Text Infilling, where sampled token spans are replaced with a sequence of mask tokens [MASK], and (2) Sentence Permutation that splits a document into declarative sentences thereafter shuffled in random order.

Abstractive summary generation can be cast as a typical sequence-to-sequence learning problem. The pretraining objective of the core transformer model is to minimize the negative log-likelihood of the original document over corrupted text

$$\mathcal{L}_G(\theta) = -\frac{1}{|Y|}\log p(Y|X;\theta),$$

where X is our extractive generated summary rendered as a set of sentences, $|Y|$ is the number of tokens in summary Y, and θ denotes the model parameters. In our experiments, we used the distilled version of BART, [3] from which we drew sentence level representation for our automatic evaluation.

4 Information Alignment

The goal of a summarization task is to concisely describe the most salient information of the input text. Thus, the summary generated should be consistent and only contain content from the input, and the included content must be relevant. Using the intuition of information alignment, defined as the extent to which the information in one generative component is grounded in another, we can evaluate summary consistency and relevance (Deng et al., 2021).

More formally, let $x_{1:n}$ and $y_{1:m}$ be our respective extractive and abstractive summary text-sequences for each book chapter. Summary tokens are each represented with contextual embeddings we extracted from pretrained BERT (Devlin et al., 2019). Using embedding matching, the alignment vector $align(y \rightarrow x)$ consists of scores $\in [0,1]$ for each token in y, and amount to the maximum cosine similarity with the tokens in x

$$(i,j) = \underset{i\in 1:n,j}{\operatorname{argmax}}\, \operatorname{cossim}(x_i, y_j),$$

where (i,j) is a pair of token indices pointing each to a distinct summary text sequence, and $1 \leq j \leq m$. The consistency metric that measures faithfulness thus follows naturally as the average of the alignment vector scores: $\operatorname{mean}(align(y \rightarrow x))$. On the other hand, relevance is implicit in our two-step model that commences with ranking source sentences by their importance.

Individual	Chapters	Tokens	FRE
Frederick Douglass	11	154,293	77.5
Mark Twain	60	620,312	75.1
Ulysses Grant	70	1,269,660	65.3
Napoleon Bonaparte	115	2,238,248	65.5

Table 2: Metadata for our test set of biographical novels.

Individual	Sentences	Min	Max	Mean
Frederick Douglass	1,812	69	703	164.7
Mark Twain	6,614	10	711	110.2
Ulysses Grant	12,139	67	301	173.4
Napoleon Bonaparte	20,514	19	861	178.4

Table 3: Chapter sentence distribution across our test set of biographical novels.

5 Evaluation

Our proposed two-step summarization method is evaluated on our curated biographical literary test-set. Automatic evaluation results are reported using

[2] https://huggingface.co/sentence-transformers

[3] https://huggingface.co/sshleifer/distilbart-cnn-12-6

	e-summary					a-summary				
Individual	Tokens	Min	Max	Mean	STD	Tokens	Min	Max	Mean	STD
Frederick Douglass	14,989	197	510	314.2	86.9	1,835	38	45	41.1	2.2
Mark Twain	98,059	63	744	369.9	144.7	9,458	21	46	38.1	4.3
Ulysses Grant	121,865	269	731	387.1	80.4	12,206	29	46	41.0	4.0
Napoleon Bonaparte	248,878	35	835	477.5	104.8	20,209	23	46	39.0	3.7

Table 4: Token-length distribution of e-summary and a-summary across our biographical narrative test set.

the canonical ROUGE measure (Lin, 2004), and we have also experimented with the recently developed BARTScore metric (Yuan et al., 2021), more suitable to NLG tasks. We compared our performance with a handful of external baselines set to reach similar objectives like ours, and analyzed the newly proposed information alignment concept and consistency metric (Deng et al., 2021). Unless otherwise noted, we report novel-level summary quality using the average of chapter scores.

Novel Test Set We obtained unicode encoding of the literature text from Project Gutenberg, and carried our work on four biographies including `Narrative of the Life of Frederick Douglass, An American Slave` by Frederic Douglass (2006), [4] `Life on the Mississippi` by Mark Twain (2004), [5] `Personal Memoirs of U. S. Grant` by Ulysses S. Grant (2004), [6] and `Memoirs of Napoleon Bonaparte` by Louis Antoine Fauvelet de Bourrienne (2006). [7] These texts total 256 chapters and over four million words (Table 2). We also post for the literary set the Flesch Reading Ease (FRE) score that identifies a difficulty level range from standard to fairly easy.

In Table 3, we present chapter sentence distribution across our narrative literary set. Per book chapter there are on average 15,491 tokens (Table 2), and about 150 sentences with a little over 100 tokens per sentence. Chapter text is notably long in form and present a challenge to generate fluent and faithful summaries in a single computational pass.

Generated Summaries Our model provides two user-settable parameters to control summary generation: (1) the number of top-ranked sentences in a chapter ordered by their relevance to the input source text and concatenated to construct the e-summary. This number is set to ten by default; (2) the maximum token-length of the predicted abstractive summary set by the user to either fifty or one hundred words. We conducted ablation experiments and analyzed the impact of the bound token-length parameter on the a-summary generation quality. In Table 4, we provide token-length distribution of both e-summary and a-summary across our literary test set. On average, e-summaries consist of 387 tokens, while a-summaries, set to a maximum length of 50 tokens, have a mean of close to 40 words. Thus, the first stage of our summarization system presents a compression ratio of roughly 40 between source chapter text and e-summaries. In the second step, generated a-summaries are more concise than their respective e-summaries by an almost order of magnitude.

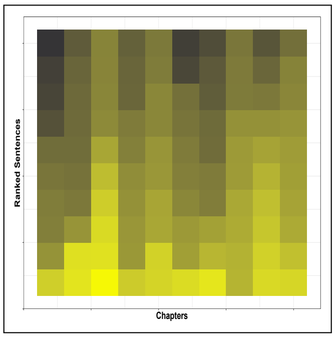

Figure 1: Chapter sentence ranking for the biography of Napoleon Bonaparte. Showing ten randomly sampled chapters and for each we highlight its respective ten top-ranked sentences in descending order. The brighter the tile, the higher the rank.

In Figure 1, we provide visualization of ten top-ranked sentences extracted from ten randomly sampled chapters in the Napoleon Bonaparte novel. We formulate extractive summaries as a matrix $\in \mathbb{R}^{m \times n}$, where m is the number of chapters in a book and n the number of top-ranked sentences that are concatenated to found an extractive summary. In our setup, LexRank is set to return a fixed number of n most relevant sentences, noting that the extracted list may contain ties. Over our experiments, we observed on average a fairly low— a slight over six percentage points— duplicated sentence salience across our test set. Most ties were an occurrence

[4] https://www.gutenberg.org/files/23/23-0.txt
[5] https://www.gutenberg.org/files/245/245-0.txt
[6] https://www.gutenberg.org/cache/epub/4367/pg4367.txt
[7] https://www.gutenberg.org/cache/epub/3567/pg3567.txt

	maxlen=50						maxlen=100					
Individual	ROUGE-2			ROUGE-L			ROUGE-2			ROUGE-L		
	r	p	f	r	p	f	r	p	f	r	p	f
Frederick Douglass	0.13	0.95	0.23	0.18	0.97	0.30	0.18	0.94	0.29	0.24	0.97	**0.38**
Mark Twain	0.11	0.91	0.19	0.16	0.96	0.26	0.17	0.91	0.27	0.22	0.96	0.35
Ulysses Grant	0.10	0.92	0.18	0.15	0.97	0.27	0.15	0.92	0.25	0.21	0.97	0.34
Napoleon Bonaparte	0.08	0.91	0.15	0.13	0.96	0.22	0.13	0.91	0.22	0.18	0.96	0.30

Table 5: ROUGE scores of a-summary generation (r - recall, p - precision, and f - F measure).

of two and a handful were of three sentences. Operating as a modular component, we applied the distilled RoBERTa-based pretrained SBERT model to generate contextual sentence embeddings. This model renders about 82 million trained parameters.

In our automatic evaluation we used the distilled checkpoint of BART, DistilBART-CNN-12-6, pretrained and finetuned on the CNN/Daily Mail news corpus (Nallapati et al., 2016) that comprises multi-sentence summaries, and on the extreme summarization dataset (XSUM; Narayan et al., 2018), both sustain a strong abstractive property. To generate a-summaries, we used the BART checkpoint model with a neural network of over 305 million parameters and ran inference on our biographical narrative test set.

ROUGE Scores We compute an a-summary from a reference e-summary. Rather than sentence-level that could potentially result in overlapping content and thus redundant summaries, we report summary-level ROUGE scores (Lin, 2004). [8] Following standard practice, we chose F1 ROUGE as our evaluation metric to estimate the generation quality of summaries. Concretely, we used bi-gram ROUGE (ROUGE-2) that is a proxy for assessing informativeness and the longest common subsequence (ROUGE-L) to represent fluency. In Table 5, we show recall, precision, and F1 scores of produced a-summaries bound to a maximum token-length (maxlen) of 50 and 100 over our biographical literary set. Consistently ROUGE-L scores are higher than the respective bi-gram performance by about twenty five percentage points, on average. As expected, summary quality reduces proportionally to the e-summary token count (Table 4). Although limited to only two settings, our results support the conjecture that the longer the summary text sequence produced the higher the performance by up to 36%.

Individual	ROUGE-2	ROUGE-L
Frederick Douglass	0.06	0.11
Mark Twain	0.07	**0.12**
Ulysses Grant	0.04	0.08
Napoleon Bonaparte	0.03	0.07

Table 6: ROUGE F1 scores for a single-tier setting. Summary maximum token-length is set to 500.

In Table 6, we report F1 ROUGE scores for a single-tier setting. This method collapses our summarizer stages and generates a-summary directly from the grounded source text of a chapter in a single computation pass. The summary maximum token length is implicitly set to 500 to account for the excessively long chapter document. Compared to our two-step summarization method, single-tier ROUGE-2 and ROUGE-L scores are shown to decline quadruply and triply, respectively.

We compared our summary generative performance with the quality of a half dozen of external baselines, presenting top F1 scores for both ROUGE-2 and ROUGE-L metrics in Table 7. At 0.29 F1, our ROUGE-2 measure exceeded state-of-the-art Gidiotis and Tsoumakas (2020) by 0.11 F1, while for ROUGE-L we came closely second with 0.38 F1 behind their best score of 0.41 F1. At an average of 15,491 words per novel chapter our dataset exceeded the token complexity of the baselines by at least 1.7X.

BARTScore We leveraged BARTScore (Yuan et al., 2021), [9] a recently introduced evaluation metric for generated text that is unsupervised and does not require human judgments to train. Owing to its ability to utilize the entirety of the BART pre-trained parameters, BARTScore can better support evaluation from a factual perspective. BARTScore relies on contextual word embeddings extracted

[8] https://pypi.org/project/rouge/

[9] We obtained code to calculate BARTScore from https://github.com/neulab/BARTScore

System	Domain	Tokens	Model	ROUGE-2	ROUGE-L
Ling and Rush (2017)	News	804	Finetuned	0.15	0.29
Xu and Lapata (2020)	QA	400	Finetuned	0.12	0.17
Pilault et al. (2020)	News	3,615	Pretrained	0.12	0.34
Gidiotis and Tsoumakas (2020)	Sci,Med	5,069	Finetuned	0.18	**0.41**
Zhong et al. (2021)	Meetings	9,070	Pretrained	0.11	0.31
Zhang et al. (2022)	TV,Reports	8,883	Pretrained	0.09	0.29
Ours	News	**15,491**	Pretrained	**0.29**	0.38

Table 7: Token complexity and ROUGE F1 scores comparison with external baselines. Neural models are at least pretrained on a large text corpus and optionally finetuned on the target dataset.

from pretrained sequence-to-sequence models and explores weighted conditional log-probabilities of a summary sequence given source tokens. In Table 8, we report BARTScore figures in average log-likelihood of probabilities $\in [0, 1]$. The calculated scores are less than zero, thus the higher the log-likelihood, the higher the probability. BARTScore appears far less affected by varying the maximum token-length of the produced a-summary, suggesting BARTScore captures aspects complementary to ROUGE. Consistent with ROUGE, BARTScore performance decreases with a higher e-summary word count.

Individual	BARTScore	
	maxlen=50	maxlen=100
Frederick Douglass	**-10.89**	-11.01
Mark Twain	-11.07	-11.04
Ulysses Grant	-11.12	-11.10
Napoleon Bonaparte	-11.17	-11.19

Table 8: BARTScore metric in log-likelihood for our biographical test set. The higher the measure the better the performance.

Individual	Pearson	Kendall	Spearman
Frederick Douglass	0.13	0.09	0.13
Mark Twain	0.24	0.19	**0.27**
Ulysses Grant	0.23	0.17	0.24
Napoleon Bonaparte	0.14	0.11	0.15

Table 9: BARTScore correlation between a-summary generation of 50 and 100 limited token-length.

We also measured the BARTScore correlation between a-summaries confined to 50 and 100 token-length, respectively. The strength of association between the two measures and the direction of the relationship are outlined in Table 9. We present Pearson, Kendall, and Spearman correlation types, all indicating a stronger positive relation for the books on Mark Twain and Ulysses Grant that share a similar token complexity per chapter.

6 Discussion

Individual	Min	Max	Mean	SD
Frederick Douglass	0.41	0.75	0.61	0.11
	0.62	0.92	**0.81**	0.11
Mark Twain	0.27	0.85	0.59	0.12
	0.43	0.94	0.76	0.11
Ulysses Grant	0.42	0.89	0.63	0.10
	0.49	0.94	0.77	0.12
Napoleon Bonaparte	0.48	0.89	0.67	0.07
	0.43	0.96	0.77	0.10

Table 10: Factual consistency distribution across our test set of biographical novels. The figures for each title show consistency measures for generated a-summaries, contrasting their alignment with the source text (grayed) and to their respective e-summary.

Factual Consistency In this section, we offer qualitative analysis of factual consistency as it relates to biographical literary using embedding-matching alignment estimation. To extract contextual embeddings we used a pretrained BERT model that has nearly 109 million parameters. Our extractive summarization step warrants textually grounded generation of a summary, thus the following discussion pertains exclusively to the abstractive-summary computational stage. In Table 10, we show the distribution of factual correctness for aligning both (a-summary → e-summary) and (a-summary → source) across our biographical literary test set. The Frederick Douglass narrative scored the highest consistency of 0.81, along with the rest of the novels slightly behind, however, we contend that the three novels uphold a more faithful score of 0.77 owing to a larger sample of chapters. Using a comparable metric for compression tasks, Deng et al. (2021) report consistency performance at 0.33 on the CNN/Daily Mail news corpus.

The impact on consistency performance gained by contrasting alignment of a-summaries with the source text and aligning a-summaries with e-summaries is a considerable 25% on average (Table 10). Evidently accurate automatic evaluation of generated summaries from long-form literary narratives is a multi-dimensional problem and pose a key challenge for optimization.

We note that extending the maximal generative token-length is not indefinite or else the summarizer aim to effectively balance both fluency and succinctness will be adversely affected.

Individual	Finetuned Train	 Test	 F1	Pretrained F1
Frederick Douglass	9	2	0.17	**0.30**
Mark Twain	48	12	0.14	0.26
Ulysses Grant	56	14	0.17	0.27
Napoleon Bonaparte	92	23	0.15	0.22
Unified	205	51	0.15	0.26

Table 11: Contrasting ROUGE-L F1 scores for finetuned and pretrained BART models across our biographical novels. Finetuned narrative chapter allocations are shown for train and test subsets in individual and consolidated datasets.

Finetuning We explored finetuning the BART checkpoint on our biographical literary set and looked at the model ability to generalize across datasets. To this end, we built a distinct model for each and all novels unified, and applied an 80/20 percent chapter split for training and testing, respectively. We trained the BART model for three epochs using a cross-entropy loss, the Adam optimizer, a batch size of 32, and a learning rate of 1e-3. In Table 11, we present finetuned ROUGE-L F1 scores using a generation not to exceed a length of 50 tokens, and contrast them with the pretrained model (Table 5). Both finetuned and pretrained results follow a similar performance decline with a growing chapter token complexity. Finetuned scores are lower than the pretrained measures by about 1.75X on average, because the BART model weights are fitting to a much smaller dataset that is genre-different from the pretrained domain. Results of finetuning on the unified dataset appear commensurate with the rates obtained on individual novel data.

Human Evaluation Perceived as the best practice to evaluate auto-generated summaries, human judgment of long-form content similar to our scale remains challenging, time consuming, and often delivers only moderately reliable results. In a more recent study, Krishna et al. (2023) conducted a survey to understand best practices for applying human evaluation to summarization of large-scale documents. Their findings concluded that summaries derived from greater length articles are rarely evaluated by humans and the results obtained are often irreproducible.

Individual	ROUGE-2	ROUGE-L
Frederick Douglass	0.56	0.61
Mark Twain	0.63	0.67
Ulysses Grant	0.66	**0.72**
Napoleon Bonaparte	0.63	0.69

Table 12: ROUGE F1 scores for a human evaluation. Summary maximum token-length is set to 100.

To ameliorate these shortfalls, our text generation process for human evaluation of summaries offers a span-based approach that resembles evidence annotation in question answering systems. A summary is thus a set of non-overlapping spans of contiguous text snippets from the chapter source. The total number of tokens across the spans is bound to the summary maximal token-length parameter. We considered twenty five readers from a book club as expert annotators, each assigned between ten to eleven distinct chapters for span labeling. We were less concerned about bias and avoided allocating more than one reader to a chapter.

In Table 12, we outline ROUGE F1 scores for human evaluation of span-based summaries. Top human scoring is at 0.72 ROUGE-L exceeding machine generation performance (Table 5) by up to about 2X. Given the current pace for developing state-of-the-art NLG systems, this apparent performance gap is expected to diminish rather precipitously, as research continues to reason the trade-off between cost and reward for conducting human annotation.

Method Generalization To evaluate the generalizability of our proposed two-step summarization method to other text genres or domains, we explored NarrativeQA (Kočiský et al., 2018). Destined for the reading comprehension (RC) problem space, NarrativeQA is a large-scale question answering dataset constructed from a collection of large documents in the form of full-length books and movie scripts. Learning to understand books through effective summarization modeling become key to a successful RC system.

NarrativeQA comprises full-length books with an average of slightly over 60K tokens per story. While its human-curated abstractive summaries has a token complexity of about 650 on average. This suggests an end-to-end compression ratio of roughly 100 from source to summary. In contrast to our automatic method that yields a data compaction rate of close to 400 across the two computational steps on our biographical test set. We note that a NarrativeQA book is represented as a cohesive long sequence of text, rather than a collection of chapter entities like ours, the result of performing a data preprocessing step on each of our novels to improve model scalability.

The authors of NarrativeQA performed question answering quality experiments comparing the use of a book in its entirety to its labor-intensive human-created summary for retrieving an answer. Using the ROUGE-L metric they achieved 0.37 for summaries and 0.14 on full length stories. Although for a different goal, these results highly resemble our automatic evaluation scores of 0.38 and 0.12 for two-step and single-tier configurations, respectively.

7 Conclusion

In this paper, we presented a summarization approach that ensures hallucination-free text generation in its first step, and follows by a more regulated and manageable production of a final abstractive summary. On a biographical literary dataset with doubled to quadrupled chapter token complexity, our method achieved superior or similar performance compared to six baseline models. Empirical results show that our fact-unaware summarization can produce abstractive summaries with compelling factual consistency. Noting that author-created book descriptions are often of less than adequate quality, we encourage not only span-based but also free-form reader-written chapter summaries that are factually faithful and benefit a plausible load sharing for curating annotations.

Limitations

Our proposed summarization model is pretrained exclusively on news datasets, however, our experiments and analysis were conducted on biographical narratives. We only studied English summarization and our processes and in particular relevance findings are likely not entirely applicable to long multi-lingual documents. Moreover, single-domain trained models may propagate inductive biases rooted in the data they were pretrained on. This was evidenced in finetuning on our target dataset as the model demonstrated a moderate degree of transferability in adapting the newswire domain to our biographical discourse genre.

Our work studies generated summaries for long narrative text. While our taxonomy appears generalizable to other domains, investigating summarization quality of large-scale datasets, such as scientific articles, patent documents, government reports or meeting discourses was confined to the scope of baseline performance comparison.

Ethics Statement

We assembled our biographical dataset for the grounded source consistent with Project Gutenberg permissions and terms of use. Emanating personal identifiable information of the individual history is unavoidable when obtained from biographical literary. However, improving the faithfulness of automatically generated summaries is essential to ensure reliable and trusted factual accuracy. To the extent of our judgment, produced narrative summaries are free of harmful or offensive content, yet we plan to restrict our dataset for research use only.

Acknowledgements

We would like to thank the anonymous reviewers for their insightful suggestions and feedback.

References

Leonid Berov. 2019. Summaries can frame - but no effect on creativity. In *International Conference on Computational Creativity (ICCC)*, pages 164–171, Charlotte, North Carolina. Association for Computational Creativity (ACC).

Meng Cao, Yue Dong, and Jackie Cheung. 2022. Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3340–3354, Dublin, Ireland. Association for Computational Linguistics.

Hyungtak Choi, Lohith Ravuru, Tomasz Dryjański, Sunghan Rye, Donghyun Lee, Hojung Lee, and Inchul Hwang. 2019. VAE-PGN based abstractive model in multi-stage architecture for text summarization. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 510–515, Tokyo, Japan. Association for Computational Linguistics.

Rumen Dangovski, Li Jing, Preslav Nakov, Mićo Tatalović, and Marin Soljačić. 2019. Rotational unit of memory: A novel representation unit for RNNs with scalable applications. *Transactions of the Association for Computational Linguistics*, 7:121–138.

Mingkai Deng, Bowen Tan, Zhengzhong Liu, Eric Xing, and Zhiting Hu. 2021. Compression, transduction, and creation: A unified framework for evaluating natural language generation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 7580–7605, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Günes Erkan and Dragomir R. Radev. 2004. LexRank: Graph-based lexical centrality as salience in text summarization. *Journal Artificial Intelligence Research (JAIR)*, 22(1):457–479.

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.

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

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Kalpesh Krishna, Erin Bransom, Bailey Kuehl, Mohit Iyyer, Pradeep Dasigi, Arman Cohan, and Kyle Lo. 2023. LongEval: Guidelines for human evaluation of faithfulness in long-form summarization. In *European Chapter of the Association for Computational Linguistics (EACL)*, Dubrovnik,Croatia. Association for Computational Linguistics.

Wendy G. Lehnert. 1981. Plot units and narrative summarization. *Cognitive Science*, 5(4):293–331.

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 *Annual Meeting of the Association for Computational Linguistics (ACL)*, 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.

Jeffrey Ling and Alexander Rush. 2017. Coarse-to-fine attention models for document summarization. In *New Frontiers in Summarization*, pages 33–42, Copenhagen, Denmark. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1906–1919, Online. Association for Computational Linguistics.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Computational Natural Language Learning (SIGNLL)*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Jonathan Pilault, Raymond Li, Sandeep Subramanian, and Chris Pal. 2020. On extractive and abstractive neural document summarization with transformer language models. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 9308–9319, Online. Association for Computational Linguistics.

Andrew Piper, Richard Jean So, and David Bamman. 2021. Narrative theory for computational narrative understanding. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 298–311, Online.

Dongqi Pu, Xudong Hong, Pin-Jie Lin, Ernie Chang, and Vera Demberg. 2022. Two-stage movie script summarization: An efficient method for low-resource long document summarization. In *Proceedings of The Workshop on Automatic Summarization for Creative Writing*, pages 57–66, Gyeongju, Republic of Korea. Association for Computational Linguistics.

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

Ming-Hsiang Su, Chung-Hsien Wu, and Hao-Tse Cheng. 2020. A two-stage transformer-based approach for variable-length abstractive summarization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2061–2072.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (neurIPS)*, volume 30. Curran Associates, Inc.

Yumo Xu and Mirella Lapata. 2020. Coarse-to-fine query focused multi-document summarization. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 3632–3645, Online. Association for Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.

Yusen Zhang, Ansong Ni, Ziming Mao, Chen Henry Wu, Chenguang Zhu, Budhaditya Deb, Ahmed Awadallah, Dragomir Radev, and Rui Zhang. 2022. Summ^n: A multi-stage summarization framework for long input dialogues and documents. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1592–1604, Dublin, Ireland. Association for Computational Linguistics.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021. QMSum: A new benchmark for query-based multi-domain meeting summarization. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 5905–5921, Online. Association for Computational Linguistics.