BOOKSUM: A Collection of Datasets for Long-form Narrative Summarization

Wojciech Kryściński† Nazneen Rajani§ Divyansh Agarwal†
Caiming Xiong† Dragomir Radev‡
†Salesforce Research §Huggingface ‡Yale University
{kryscinski, divyansh.agarwal, cxiong}@salesforce.com
nazneen@huggingface.co
dragomir.radev@yale.edu

Abstract

The majority of existing text summarization datasets include short-form source documents that lack long-range causal and temporal dependencies, and often contain strong layout and stylistic biases. While relevant, such datasets will offer limited challenges for future text summarization systems. We address these issues by introducing BOOKSUM, a collection of datasets for long-form narrative summarization. Our dataset covers documents from the literature domain, such as novels, plays and stories, and includes highly abstractive, human written summaries on three levels of granularity of increasing difficulty: paragraph-, chapter-, and book-level. The domain and structure of our dataset poses a unique set of challenges for summarization systems, which include: processing very long documents, non-trivial causal and temporal dependencies, and rich discourse structures.

To facilitate future work, we trained and evaluated multiple extractive and abstractive summarization models as baselines for our dataset.

1 Introduction

Text summarization aims at condensing long documents into a short, human-readable form which contains only the salient parts of the source. Leveraging the cutting-edge findings in natural language processing, such as multi-task learning methods (Raffel et al., 2019), pre-training strategies (Zhang et al., 2019a), and memory-efficient architectures (Zaheer et al., 2020), text summarization has seen substantial progress.

The majority of papers published in the field focus on summarizing newswire documents from popular datasets, such as CNN/DailyMail (Nallapati et al., 2016). Other domains gaining interest of the research community are scientific and legal documents, with notable datasets being Arxiv/PubMed (Cohan et al., 2018) and BigPatent (Sharma et al., 2019). While the performance of state-of-the-art methods on those datasets is impressive, the mentioned domains have inherent shortcomings and thus pose limited challenges for future generations of text summarization systems. First, the length of summarized documents is limited, ranging from a few hundred words in case of news articles, to a few pages for scientific documents and patent applications. In most cases, such short-form documents can be quickly read by humans, thus limiting the practical value of automatic summarization systems. Second, the domains under consideration impose strict requirements regarding the document’s layout and stylistic features. Statements follow a logical order and all facts are offered explicitly, leaving limited space for interpretation and reasoning. Additionally, such constraints, can introduce layout biases into the datasets which later dominate the training signal of the summarization systems. The lead-bias present in news articles being one example of such effects (Kedzie et al., 2018; Kryściński et al., 2019). Third, documents in the mentioned domains lack long-range causal and temporal dependencies, and rich discourse structures. Due to the limited length and fact-centric style of writing, most causal dependencies span only a few paragraphs, temporal dependencies are organized in a monotonic fashion where newly introduced facts refer only to previously stated information, and document lacks features such as parallel plot lines.

In this work we address the shortcomings of existing datasets and introduce BOOKSUM, a collection of data resources for long-form narrative summarization Ladhak et al. (2020). The data covers documents from the literature domain, including stories, plays, and novels (Fig. 1), each provided with a highly abstractive, human-written summary. Leveraging the characteristics of fiction writing, BOOKSUM introduces a set of new challenges for summarization systems: processing long-form texts ranging up to hundreds of pages, under-
standing non-trivial causal and temporal dependencies spread out through the entirety of the source, handling documents with rich discourse structure which include parallel plots or changes between narration and dialogue, and generating highly abstractive and compressive summaries. Solving such challenges will require progress in both automatic document understanding and processing of long inputs. To support incremental progress, the BOOKSUM collection includes examples on three levels of granularity with increasing difficulty: 1) paragraph-level, with inputs consisting of hundreds of words and short, single-sentence summaries, 2) chapter-level, with inputs covering several pages and multi-sentence summaries, 3) book-level, with inputs spanning up to hundreds of pages and multi-paragraph summaries. The hierarchical structure of the dataset, with aligned paragraph, chapter, and book-level data, makes it a viable target for single- and multi-document summarization approaches.

To demonstrate the new set of challenges for text summarization models introduced by BOOKSUM and lay the groundwork for future research, we evaluated several state-of-the-art extractive and abstractive summarization architectures on the newly introduced task. We share the data preparation scripts here: https://github.com/salesforce/booksum.

2 Related Work

The availability of digital documentation has translated into a number of novel, large-scale datasets for text summarization that span a variety of domains. In the news domain, Sandhaus (2008) introduced a corpus of news articles from the New York Times magazine with summaries written by library scientists. Nallapati et al. (2016) collected articles from the CNN and DailyMail portals with multi-sentence article highlights repurposed as reference summaries. Narayan et al. (2018) aggregated articles from the BBC website with highly abstractive, single sentence summaries. Grusky et al. (2018) introduced a dataset spanning 38 news portals, with summaries extracted from the websites metadata. In the academic article domain, Cohan et al. (2018) collected scientific articles from the Arxiv and Pub-Meb article repositories and used paper abstracts as reference summaries. Wang et al. (2020) aggregated a set of articles in the medical domain related to the Covid-19 pandemic, also using paper abstracts as reference summaries. Hayashi et al. (2020) introduced a multi-domain collection of scientific articles each with two associated summaries, one covering the article’s contributions, the other explaining the context of the work. Related to dialogue summarization, Pan et al. (2018) repurposed image captioning and visual dialogue datasets to create a summarization dataset containing conversations describing an image, with image captions considered the summaries. Gliwa et al. (2019) introduced a corpus of conversations between hired annotators designed to mimic interactions on a messaging application with human written summaries. In the legal domain, Sharma et al. (2019) has collected a large collection of patent filings with associated, author-written invention descriptions.

Despite the increased interest in the broader field of text summarization, little work has been done in summarizing stories and novels. In Kazantseva (2006), the authors focused on generating extractive overviews of short works of fiction. The work proposed two modeling approaches, one utilizing decision trees the other based on a manually designed system of rules with experiments conducted on a set of 23 short stories. Mihalcea and Ceylan (2007) introduced the task of book summarization along with a set of resources and baselines. The authors collected and curated a set of 50 books from the Gutenberg Project with two human-written summaries associated with each book collected from online study guides. More recently, Zhang et al. (2019b) tackled the problem of generating character descriptions based on short fiction stories. The authors collected a dataset of stories with associated, author-written summaries from online story-sharing platforms and proposed two baseline methods for solving the task. Ladhak et al. (2020) ex-
explored the problem of content selection in novel chapter summarization. The authors studied different approaches to aligning paragraphs from book chapters with sentences from associated summaries and created a silver-standard dataset for extractive summarization. The work also studied the performance of extractive models on the task.

Our work extends the efforts made by Ladhak et al. (2020). The BOOKSUM corpus prioritizes abstractive summarization and offers aligned data on three levels of granularity (paragraph, chapter, full-book), substantially increasing the number of available examples. We also benchmark the performance of state-of-the-art extractive and abstractive methods on all introduced data subsets.

3 Dataset
In this section we describe the data sources and pre-processing steps taken to create the BOOKSUM data collection and conduct an in-depth analysis of the collected resources.

3.1 Data Collection
Data Sources Despite the popularity of books in electronic format, aggregating and sharing literature pieces is a non-trivial task due to the copyright law protecting such documents. The source documents available in BOOKSUM were collected from the Project Gutenberg public-domain book repository and include plays, short stories, and novels of which copyrights have expired. Associated summaries were collected using content provided by the Web Archive. The summary data includes both book- and chapter-level summaries.

Data Acquisition Source texts were downloaded in plain text format in accordance with Project Gutenberg’s guidelines. The data collection contains texts exclusively from the US edition of Project Gutenberg. Summaries were collected using content provided by the Web Archive and processed using the BeautifulSoup library. Collecting summaries from several independent sources with small content overlaps between them resulted in certain texts having multiple associated summaries. Upon manual inspection, substantial stylistic differences were found between the related summaries, thus such coverage overlap was considered advantageous for the dataset.

Data Cleaning & Splitting To ensure high quality of the data, both the source texts and summaries were cleaned after collection. Metadata containing author, title, and publisher information was removed from source files. The documents were manually split into individual chapters to accommodate chapter-level summarization. Due to the unstructured nature of plain text files, heuristic approaches were used to extract chapter content. Initial, automatic chapterization was done using the regex-based Chapterize tool. However, an inspection of outputs revealed many partially processed and unprocessed files, such instances were chapterized manually by the authors of this work. Paragraph-level data was obtained by further splitting the extracted chapter into individual paragraphs based on a white-character pattern. Short paragraphs and dialogue utterances were aggregated to form longer paragraphs. Collected summaries were also inspected for scraping artifacts and superfluous information. Regular expressions were used to remove leftover HTML tags, author’s notes, and analysis parts that were not directly related to the content of the summary.

Data Pairing Source texts and associated summaries were collected independently of each other and required alignment. The pairing procedure was conducted in phases, starting with coarse-grained full-text alignments and ending with fine-grained paragraph alignments, with each phase involving automatic alignments followed by manual inspection and fixes. Full texts were paired with summaries based on title matches and later verified by matching author names. To accommodate automatic alignment, titles were normalized into a common format with lower-case letters and all punctuation characters removed. Chapter alignments were based on chapter metadata, extracted during source text chapterization, and chapter titles collected from online study guides. Similar to full-text titles, chapter names were transformed to a common format with chapter names lower-case and cleaned from punctuation characters, and chapter numbers translated to roman numerals. Paragraph-level alignments were computed between paragraphs extracted from chapters and

---

2US edition: https://www.gutenberg.org/
3https://web.archive.org/
4https://www.gutenberg.org/policy/robot_access.html
5https://crummy.com/software/BeautifulSoup/
6https://github.com/JonathanReeve/chapterize
individual sentences of chapter-level summaries. Following a two step process introduced by Ladhak et al. (2020), the alignment process was preceded by a human-based study aimed at finding an optimal alignment strategy, with its details presented in Appendix B. With the insights from the study, paragraph-sentence similarities were computed using a SentenceTransformer (Reimers and Gurevych, 2019), and leveraged a stable matching algorithm (Gale and Shapley, 1962) to obtain the final alignments. All examples on the chapter- and book-level, and a random subset of examples on the paragraph-level were manually inspected to ensure high quality of data. Quantitative verification of alignment quality is discussed in Appendix C.

**Data Splits** The data was split into training, validation, and test subsets in a 80/10/10% proportion. To prevent data leakage between data subsets, the splits were assigned per book title, meaning that all paragraph, chapter, and full-book examples belonging to the same book title were assigned to the same data split. For consistency with the dataset introduced by Ladhak et al. (2020), all titles overlapping between the two datasets were assigned to the same splits. Remaining titles were assigned to splits at random following the predefined size proportions. The data collection and pre-processing pipeline is visualized in Figure 3 in the Appendix D.

**3.2 Data Analysis**

**Data Statistics** The data collection and matching process described in Section 3.1 yielded 217 unique book titles with a total of 6,327 book chapters. After the pre-processing and alignment steps, the BOOKSUM collection contains 146,532 paragraph-level, 12,630 chapter-level, and 405 book-level examples. Figure 1 shows the distribution of literary genres in our corpus. Following Grusky et al. (2018), we computed statistics of the BOOKSUM collection and compared them with other popular summarization datasets in Table 1. Coverage and density, which measure the extractive span similarity between source and summary, indicate that while the extractiveness of summaries increases from 0.5 and 0.92 for paragraphs to 0.89 and 1.83 for full-books, the summaries are still highly abstractive when compared to other datasets, such as CNN/DM or Newsroom. Relatively low coverage and density scores for paragraph-level alignments might partially be an artifact of the heuristic approach to aligning the data. The lengths of source and summary texts substantially increases across data granularity. Paragraph-level data includes short documents with an average of 159 words which fit within the limitations of existing models, chapter-level examples contain texts with average of over 5000 words, which are longer than in most of existing datasets and go beyond limitations of many state-of-the-art methods (Liu et al., 2019), while book-level examples contain inputs with over 110,000 words on average, which are orders of magnitude longer than any document previously used in NLP tasks. While long source documents create computational challenges for encoding components of models, the associated summaries on chapter- and book-level are also much longer than in any other dataset, thus creating challenges for the generative component of summarization methods.

**Salient Content Distribution** To assess the difficulty of content selection in our datasets we measure the distribution of salient unigrams in the source texts (Sharma et al., 2019). The distribution is computed as the percentage of salient unigrams in four equally sized segments of the source
Figure 2: The datasets statistics of BOOKSUM and previously introduced datasets. Figure (a) shows the salient unigram distribution over 4 equally sized segments of the source documents. Figure (b) shows the percentage of novel n-grams in the reference summaries when compared with the source documents.

Summary Abstractiveness To quantify the abstractiveness of summaries in BOOKSUM we measured the percentage of n-grams from summaries not appearing in the associated source document (See et al., 2017). Results presented in Figure 2 (b) show that BOOKSUM contains highly abstractive summaries across all measured n-gram sizes. The highest ratio of novel n-grams in BOOKSUM was found for the paragraph-level alignments, followed by chapter-level data and full-books. Results also indicate that our dataset is substantially more abstractive than most previous datasets, with the exception of XSum. High scores for trigrams also indicate that summaries included in BOOKSUM do not contain long extractive spans, which aligns with the Density statistics shown in Table 1.

Qualitative Study For a deeper understanding of the data beyond quantitative evaluation, we manually analyzed subsets of BOOKSUM. First we compared summaries on different levels of granularity assigned to the same title. Summaries on the chapter- and book-level partially overlap in the summarized content, however substantially differ in the level of detail with which they cover the content. This relation could be leveraged for training models in a hierarchical fashion, from shorter to longer source texts (Li et al., 2015). Next, we compared summaries coming from different sources which were aligned with the same book or chapter. We noticed that the summaries had high semantic and low lexical overlap, meaning that they covered the same content of the summarized documents, but were written in a unique way. Such examples contain useful training signal for abstractive summarization models. Table 7 shows examples of chapter summaries of "Sense and Sensibility".

4 Experiments

To motivate the challenges posed by the BOOKSUM corpus, we study the performance of multiple baseline models, both extractive and abstractive, on the different levels of alignment: paragraph, chapter and books. We refer to these levels of alignment as BOOKSUM-Paragraph, BOOKSUM-Chapter, and BOOKSUM-Book accordingly.

4.1 Baseline Models

Lead-3 (See et al., 2017) is an extractive heuristic where the first three sentences from the source document are treated as the summary. Despite its simplicity, Lead-3 is a strong baseline for domains which show layout biases, such as newswire.
which capture long-range dependencies using a 
with automatic metrics: ROUGE-n (R-n), BERTScore (BS), and SummaQA (SQA).

The model can 
large-scale seq-to-seq Transformer architecture in 
training the baselines on long input sequences. To 
Computational constraints and input 
length limits of pre-trained models prevent us from 
Training the baselines on long input sequences. To 
Circumvent those issues we follow a 
and rank approach for BOOKSUM-Chapter and 
We use baseline models fine-tuned on 
BOOKSUM-Paragraph, to 
generate individual summaries for all paragraphs in BOOKSUM-Chapter and 
Next, we rank the generated 
summaries based on the model’s confidence. In 
In case of ablative models we look at the perplexity-
level, for ablative models we take the model assigned scores. As the final chapter- or book-level 
summaries we combine the top-k ranked paragraph-
summaries, where k is chosen based on summary length statistics in the training set.

Extractive Oracle 
We follow the steps described 
by Zhong et al. (2020) to generate oracle candidates for the 
BOOKSUM-Paragraph data. First, we 
compute a mean ROUGE-\{1,2,L\} score between 
each sentence in a paragraph and the associated summary. Next, we select the 5 highest scoring sentences and generate all combinations of 1, 2, and 3 sentences to serve as candidate oracles. The final oracle chosen from the set of candidates is the one which maximizes the mean ROUGE-\{1,2,L\} score with the paragraph summary.

**Table 2**: Performance of baseline models on the Paragraph, Chapter, and Full-Book subsets of BOOKSUM evaluated with automatic metrics: ROUGE-n (R-n), BERTScore (BS), and SummaQA (SQA).

**Random Sentences** follows the Lead-3 heuristic and extracts 3 sentences sampled at random from the source document. It represents the performance of an untrained extractive baseline.

**CNN-LSTM Extractor** (Chen and Bansal, 2018) builds hierarchical sentence representations which capture long-range dependencies using a CNN and bi-directional LSTM-RNN layers. A separate LSTM-based pointer network is applied to the representations to extract summary sentences.

**BertExt** (Liu and Lapata, 2019) extends the BERT (Devlin et al., 2019) model with the ability to generate distinct representations for multiple text spans. Based on those representations the model selects sentences into the extractive summary.

**MatchSum** (Zhong et al., 2020) formulates extractive summarization as a semantic text matching problem. Multiple candidate summaries are extracted and embedded as dense vectors using a Siamese-BERT model and matched with the reference text in the semantic space.

**BART** (Lewis et al., 2019) uses a denoising autoencoder pre-training strategy designed specifically for NLG tasks. It has achieved state-of-the-art results on many generative tasks, including ablative text summarization.

**T5** (Raffel et al., 2019) approaches transfer learning by unifying multiple NLP tasks into a common text-to-text format. All tasks are modeled with a large-scale seq-to-seq Transformer architecture in the order of billions of parameters. The model can be used to generate ablative summaries using a summarize: prefix added to the text.

**PEGASUS** (Zhang et al., 2019a) uses a pre-training objective designed for ablative text summarization which includes masked language modeling and gap sentence generation. The model achieved state-of-the-art performance on multiple summarization datasets.

### 4.2 Setup

**Modeling** Computational constraints and input length limits of pre-trained models prevent us from training the baselines on long input sequences. To circumvent those issues we follow a **generate & rank** approach for BOOKSUM-Chapter and BOOKSUM-Book. We use baseline models fine-tuned on BOOKSUM-Paragraph, to **generate** individual summaries for all paragraphs in BOOKSUM-Chapter and BOOKSUM-Book. Next, we **rank** the generated summaries based on the model’s confidence. In case of ablative models we look at the perplexity-level, for ablative models we take the model assigned scores. As the final chapter- or book-level summary we combine the top-\(k\) ranked paragraph-summaries, where \(k\) is chosen based on summary length statistics in the training set.

---

**Table 2**: Performance of baseline models on the Paragraph, Chapter, and Full-Book subsets of BOOKSUM evaluated with automatic metrics: ROUGE-n (R-n), BERTScore (BS), and SummaQA (SQA).

| Models                  | BOOKSUM-Paragraph | BOOKSUM-Chapter | BOOKSUM-Book |
|-------------------------|-------------------|-----------------|--------------|
|                         | R-1\(_f_1\) | R-2\(_f_1\) | R-L\(_f_1\) | BS\(_f_1\) | SQA\(_f_1\) | R-1\(_f_1\) | R-2\(_f_1\) | R-L\(_f_1\) | BS\(_f_1\) | SQA\(_f_1\) | R-1\(_f_1\) | R-2\(_f_1\) | R-L\(_f_1\) | BS\(_f_1\) | SQA\(_f_1\) |
| Lead-3                  | 17.99  | 3.25  | 12.80  | 0.085  | 22.18  | 14.32  | 2.23  | 8.59  | -0.008 | 19.28  | 6.50  | 0.89  | -4.34  | -0.060 | 24.18  |
| Random Sentences        | 17.56  | 3.03  | 12.32  | 0.080  | 19.99  | 12.54  | 1.32  | 7.43  | -0.029 | 7.86  | 5.26  | 0.48  | 3.38  | -0.119 | 6.28   |
| Extractive Oracle       | 27.90  | 7.22  | 20.96  | 0.172  | 17.35  | 42.25  | 9.83  | 20.91 | 0.153  | 17.16  | 46.30  | 9.25  | 17.77 | 0.076  | 17.90  |

| Extractive Models       |                  |                  |              |
|-------------------------|-------------------|-----------------|--------------|
| CNN-LSTM                | 16.71  | 2.93  | 12.84  | 0.096  | 7.17   | 32.50  | 5.51  | 13.91 | 0.081  | 3.81   | 38.79  | 6.90  | 14.28 | 0.033  | 3.09   |
| BertExt                 | 14.55  | 2.29  | 10.52  | 0.053  | 8.12   | 32.06  | 5.37  | 13.68 | 0.076  | 8.61   | 36.32  | 6.04  | 13.29 | 0.020  | 9.64   |
| MatchSum                | 19.13  | 3.38  | 14.17  | 0.125  | 16.47  | 30.97  | 5.34  | 13.23 | 0.082  | 16.83  | 33.76  | 5.58  | 12.68 | 0.012  | 24.11  |

| Abstractive Models      |                  |                  |              |
|-------------------------|-------------------|-----------------|--------------|
| BART fine-tuned          | 17.66  | 2.29  | 13.15  | 0.132  | 12.03  | 31.54  | 5.25  | 14.20 | 0.088  | 11.80  | 35.35  | 5.36  | 12.93 | 0.021  | 11.90  |
| T5 fine-tuned            | 19.84  | 3.61  | 14.05  | 0.096  | 19.07  | 31.28  | 5.25  | 13.05 | 0.075  | 19.00  | 35.50  | 5.60  | 12.04 | 0.011  | 20.43  |
| PEGASUS-large fine-tuned | 16.35  | 2.61  | 12.01  | 0.093  | 14.34  | 30.61  | 4.51  | 13.07 | 0.077  | 14.47  | 34.16  | 5.05  | 12.40 | 0.021  | 20.29  |
| BART base-trained        | 22.79  | 5.09  | 17.70  | 0.211  | 13.96  | 37.51  | 8.49  | 17.05 | 0.156  | 13.64  | 38.71  | 7.59  | 13.65 | 0.051  | 14.89  |
| T5 base-trained          | 21.07  | 4.87  | 16.96  | 0.212  | 13.93  | 36.60  | 7.93  | 16.79 | 0.152  | 13.65  | 39.87  | 8.01  | 13.99 | 0.074  | 15.13  |
| PEGASUS-large base-trained| 20.32  | 4.56  | 16.22  | 0.194  | 10.23  | 35.86  | 7.47  | 16.03 | 0.132  | 11.54  | 36.03  | 7.23  | 12.88 | 0.050  | 14.85  |
Implementation Models were implemented in Python using the PyTorch (Paszke et al., 2019) and Huggingface (Wolf et al., 2019) libraries. Abstractive models were initialized from pretrained checkpoints shared through the Huggingface Model Hub. Additional details are listed in Appendix A.

Training & Inference All models were trained for 10 epochs and evaluated on the validation split at the end of each epoch. Final model checkpoints were chosen based on the performance of models on the validation data. Model outputs were decoded using beam search with 5 beams and n-gram repetition blocking for n > 3 (Paulus et al., 2018).

Evaluation Metrics Models were evaluated using a suite of automatic evaluation metrics included in the SummEval toolkit (Fabbri et al., 2021). Lexical overlap between n-grams in generated and reference summaries was measured using ROUGE-\{1,2,L\} metrics (Lin, 2004). Semantic overlap between mentioned summaries was evaluated using BERTScore (Zhang et al., 2020), which aligns summaries on a token-level based on cosine similarity scores between token embeddings. We also inspect content overlap between generated summaries and source documents by employing SummaQA (Scialom et al., 2019), which generates questions based on the input document and next applies a QA system to evaluate how many of those questions can be answered using the summary. Due to the input length limits of SummaQA, the metric was applied individually to paragraphs of chapters and books and next aggregated by averaging to obtain chapter and book-level scores.

4.3 Automatic Evaluation We first evaluate the baseline models using automatic metrics, with results shown in Table 2.

A general trend showing across all evaluated models is low BERTScore values which decrease as reference summaries get longer (from paragraphs to full books). The metric operates on a \([-1, 1]\) range, and the highest scores, slightly above 0.19, were achieved by the fine-tuned T5 model on a paragraph level. This suggests that BERTScore might not be a good fit for evaluating highly abstractive, long summaries. We decided to include it in the evaluation process to highlight this issue for future investigation.

Heuristics The performance of the Lead-3 baseline is relatively low, scoring an R-1 of 17.99, 14.32, and 6.50 on the paragraph-, chapter-, and book-level respectively. The random sentence baseline closely trails Lead-3 across all metrics and data splits. Both results suggest that data from the literature domain included in BOOKSUM may be less susceptible to layout biases present in other domains, such as newswire. Extractive oracle scores on paragraph data substantially underperformed those on the chapter and book data. This could be an artifact of the data pairing procedure where the content of a highly abstractive summary sentences is partially covered by the matched paragraph.

Extractive Models The performances of the CNN-LSTM and BertExt models are very similar, with the first model being better on paragraph data, and the second model performing better on chapters and books. The small performance gap between the two mentioned models is surprising considering that the BERT based model was initialized from a pre-trained checkpoint, while the CNN-LSTM model was trained from scratch. The MatchSum baseline which reported state-of-the-art performance on news domain datasets (Zhong et al., 2020) achieved the best performance on a paragraph level, but underperformed the other models on chapter and book summaries.

Abstractive Models We evaluated the performance of abstractive models both in a zero-shot setting and after fine-tuning on the BOOKSUM-Paragraph data. We find that fine-tuning models on the BOOKSUM data leads to consistent improvements across all models and data granularities, with the exception of the BART model on the book-level which performed better in a zero-shot fashion according to the ROUGE metric, and the T5 model on the SQA metrics. Upon manual inspection of model outputs we noticed that zero-shot models included fragments of dialogues in the summaries which are less likely to be found in reference summaries, this in turn could contribute to the lower evaluation scores of zero-shot baselines. The BART model achieved the best performance out of all the baseline models on paragraph- and chapter-level data, while T5 performed best on the book-level. Despite its state-of-the-art performance on most summarization datasets (Zhang et al., 2019a), we found PEGASUS to underperform other baseline models, both in the zero-shot
Table 3: Performance of baseline models on the Paragraph, Chapter, and Full-Book subsets of BOOKSUM evaluated by human annotators. Judges were asked to assess the fluency (Flu.), coherence (Coh.), relevance (Rel.) and factuality (Fact.) of generated summaries. Relevance and factuality were not evaluated on the chapter- and book-level due to the length of the source texts.

| Models                | BOOKSUM-Paragraph | BOOKSUM-Chapter | BOOKSUM-Book |
|-----------------------|-------------------|-----------------|--------------|
|                       | Flu.  | Coh.  | Rel.  | Fact. | Flu.  | Coh.  | Rel.  | Fact. | Flu.  | Coh.  | Rel.  | Fact. |
| BART fine-tuned       | 4.24  | 4.15  | 4.03  | 3.96  | 4.16  | 3.96  | -     | -     | 3.80  | 3.87  | -     | -     |
| T5 fine-tuned         | 4.08  | 4.09  | 3.99  | 3.96  | 4.03  | 4.01  | -     | -     | 3.93  | 3.91  | -     | -     |
| PEGASUS-large fine-tuned | 4.10  | 4.13  | 3.99  | 3.87  | 4.07  | 3.96  | -     | -     | 3.95  | 3.79  | -     | -     |

4.4 Human Evaluation

To further assess the performance of abstractive baselines, human annotators were hired and asked to evaluate generated summaries across four dimensions: fluency, coherence, relevance, and factuality. Scores were assigned on a Likert scale from 1 to 5, with each example annotated by 3 judges and the scores averaged. Relevance and factuality were evaluated only on the paragraph-level since both dimensions require an understanding of the source text, which in the case of chapters and books is prohibitively long. Results are shown in Table 3.

Similarly to the study using automatic metrics, BART shows strong performance across all dimensions for the paragraph- and chapter-level subsets and slightly underperforms on full books. The results also show a general decrease in fluency and coherence across all models as the length of the source documents and summaries increases. This suggests that generating longer passages of fluent and coherent text poses a problem for existing neural models and could be addressed in future work.

4.5 Discussion

The generate & rank approach allowed us to overcome the limitations of existing models and apply the baselines to the chapter- and book-level data. We recognize that generating and scoring sentences independently has drawbacks, namely: 1) generated summaries may lack coherence, 2) content of selected sentences may overlap or be of low significance, which could negatively affect the overall relevance of the summary. However, the experiments discussed in this section were intended to be groundwork for the introduced task and we leave developing more tailored methods for future work.

The experiment results also show that BOOKSUM poses challenges not only for existing summarization models, but also for evaluation metrics. The abstractive nature of reference summaries makes lexical overlap measured by ROUGE an inadequate metric for model evaluation (Fabbri et al., 2021). Other recently introduced metrics, such as BERTScore and SummaQA, leverage pre-trained neural models, which in turn makes them subject to the same input length limitations as the evaluated summarization models. While the model-based metrics can be individually applied to chunks of the data and then aggregated, as in the case of SummaQA, such use was not studied by the authors and could affect the reliability of returned scores. Human-based studies, which are often used to assess dimensions omitted by automatic metrics, are also problematic when conducted with long-form data included in BOOKSUM. For example, assessing factual consistency requires annotators to be familiar with the content of the source document, which in the case of chapters or books could span dozens of pages making such studies unreliable and prohibitively time consuming.

5 Conclusions

In this work we introduced BOOKSUM, a collection of datasets for long-form narrative summarization. BOOKSUM includes annotations on three levels of granularity of increasing difficulty: paragraph, chapter, and full-book. Through a quantitative analysis we compare our dataset to existing summarization corpora and show that BOOKSUM sets new challenges for summarization methods. We trained extractive and abstractive baseline models leveraging state-of-the-art pre-trained architectures to test the performance of current methods on the task of long-narrative summarization and to enable easy comparison with future methods. We hope our dataset will contribute to the progress made in the field of automatic text summarization.
6 Limitations

Data Collection Web data is subject to local copyright laws. For data that is no longer protected by copyright law, we understand the use described within the paper is legally permissible. For data that is subject to copyright, we understand that such use is allowed under U.S. copyright law’s fair use provision. Depending on how others use this data, the purpose of their use, the jurisdiction they are in, and other factors considered under copyright law, we understand that the decision on whether a specific use case is fair use involves a legal analysis. It is advisable to obtain legal counsel prior to using such data. All data described in this work was collected exclusively for the academic purpose of conducting research. The purpose of using the BOOKSUM data was only for training models and not for public display or any other use. No data was stored upon completion of the research process.

Data Biases The BOOKSUM dataset contains books written or translated into English. These books are also more than fifty years old and so representative of society in that era. The various pretrained models we evaluated on our dataset carry biases of the data they were pretrained on. However, we did not stress test these models for such ethical biases. We request our users to be aware of these ethical issues in our dataset that might affect their models and evaluations.

Model Evaluation In this work, we have used established metrics, such as ROUGE, as well as recently introduced metrics, such as BERTScore and SummaQA, to evaluate the introduced baseline models. However, such automatic metrics have not been evaluated for use with very long source documents and highly abstractive summaries. Thus, might not accurately reflect the true performance of the evaluated models. Reliable evaluation of highly abstractive summarization models trained on long source documents is an open problem and an area of active research. Authors using the BOOKSUM data are encouraged to consult appropriate literature whether more robust evaluation methods are available at the time of writing.

Computational Resources Considering the length of source documents included in the BOOKSUM dataset, training and evaluation of neural models might require substantial computational resources.

References

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. arXiv preprint arXiv:1805.11080.

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, NAACL-HLT, New Orleans, Louisiana, USA, June 1–6, 2018, Volume 2 (Short Papers), pages 615–621. 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 Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2–7, 2019, Volume 1 (Long and Short Papers), pages 4171–4186. Association for Computational Linguistics.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir R. Radev. 2021. Summeval: Re-evaluating summarization evaluation. Trans. Assoc. Comput. Linguistics, 9:391–409.

D. Gale and L. S. Shapley. 1962. College admissions and the stability of marriage. The American Mathematical Monthly, 69(1):9–15.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. CoRR, abs/1911.12237.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers), pages 708–719. Association for Computational Linguistics.

Hiroaki Hayashi, Wojciech Kryściński, Bryan McCann, Nazneen Fatema Rajani, and Caiming Xiong. 2020. What’s new? summarizing contributions in scientific literature. CoRR, abs/2011.03161.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2019. Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring. arXiv preprint arXiv:1905.01969.

Anna Kazantseva. 2006. An approach to summarizing short stories. In EACL 2006, 11st Conference of the
European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, April 3-7, 2006, Trento, Italy. The Association for Computer Linguistics.

Chris Kedzie, Kathleen R. McKeown, and Hal Daumé III. 2018. Content selection in deep learning models of summarization. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018, pages 1818–1828. Association for Computational Linguistics.

Wojciech Kryściński, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. CoRR, abs/1908.08960.

Faisal Ladhak, Bryan Li, Yaser Al-Onaizan, and Kathleen R. McKeown. 2020. Exploring content selection in summarization of novel chapters. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, pages 5043–5054. Association for Computational Linguistics.

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

Jiwei Li, Minh-Thang Luong, and Dan Jurafsky. 2015. A hierarchical neural autoencoder for paragraphs and documents. CoRR, abs/1506.01057.

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.

Chunyi Liu, Peng Wang, Jiang Xu, Zang Li, and Jieping Ye. 2019. Automatic dialogue summary generation for customer service. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019, pages 1957–1965. ACM.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. arXiv preprint arXiv:1908.08345.

Rada Mihalcea and Hakan Ceylan. 2007. Explorations in automatic book summarization. In EMNLP-CoNLL 2007, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, June 28-30, 2007, Prague, Czech Republic, pages 380–389. ACL.

Ramesh Nallapati, Bowen Zhou, Cicero Nogueira dos Santos, Çağlar Gülčehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. In Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016, Berlin, Germany, August 11-12, 2016, pages 280–290. ACL.

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 Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium.

Haojie Pan, Junpei Zhou, Zhou Zhao, Yan Liu, Deng Cai, and Min Yang. 2018. Dial2desc: End-to-end dialogue description generation. CoRR, abs/1811.00185.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andres Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, pages 8024–8035.

Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. In 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. CoRR, abs/1910.10683.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics.

Evan Sandhaus. 2008. The New York Times Annotated Corpus.

Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2019. Answers unite! unsupervised metrics for reinforced summarization models. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, pages 3244–3254. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics.
Eva Sharma, Chen Li, and Lu Wang. 2019. **BIG-PATENT**: A large-scale dataset for abstractive and coherent summarization. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019*, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, pages 2204–2213. Association for Computational Linguistics.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William Merrill, Paul Mooney, Dewey Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Chris Wilhelm, Boya Xie, Douglas Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. **CORD-19**: the covid-19 open research dataset. *CoRR*, abs/2004.10706.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. **Huggingface’s transformers**: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Manzil Zaheer, Guru Guruganesh, 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. *CoRR*, abs/2007.14062.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2019a. **PEGASUS**: pre-training with extracted gap-sentences for abstractive summarization. *CoRR*, abs/1912.08777.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. **Bertscore**: Evaluating text generation with BERT. In 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net.

Weiiwei Zhang, Jackie Chi Kit Cheung, and Joel Oren. 2019b. Generating character descriptions for automatic summarization of fiction. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 7476–7483. AAAI Press.

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2020. Extractive summarization as text matching. *arXiv preprint arXiv:2004.08795*. 

6546
A Further Implementation Details

Model hyperparameters followed the best configurations described by the original authors of the models. Models were trained for 10 epochs using a batch size of 16. Many of the baselines presented in this work leveraged pre-trained checkpoints to initialize weights before fine-tuning on the BOOKSUM data. Table 4 lists the checkpoints used for each of the baselines and the approximate number of parameters of each model. Experiments were conducted using 4 NVidia A100 GPUs, all studies described in this paper took an approximate 8 GPU hours.

| Model                  | Checkpoint | Parameters |
|------------------------|------------|------------|
| Bi-Encoder (pretrained) | sentence-transformer/gpt2-medium  | 6547M       |
| Cross-Encoder          | cross-encoder/stsb-roberta-large  | 737M        |
| Bi-Encoder (roBERTa)   | sentence-transformers/roberta-base-xl | 568M        |
| Cross-Encoder          | cross-encoder/stsb-roberta-large  | 406M        |
| Bi-Encoder (paraphrase)| sentence-transformers/paraphrase-distilroberta-base-xl | 109M        |
| Cross-Encoder          | sentence-transformers/paraphrase-distilroberta-base-xl | 109M        |

Table 4: Hugginface Model Hub checkpoints used to initialize baseline and similarity score models

B Data Alignment Process

Alignments between book paragraphs and sentences from associated summaries were computed using heuristic methods. The alignment processed followed two steps described by Ladhak et al. (2020): 1) similarity scores were computed for all paragraph-sentence pairs, 2) based on the similarity scores paragraph and sentence were aligned using a stable matching algorithm. Similarity scores between paragraphs and sentences can be computing using different metrics. In our study, we focused on lexical overlap methods and neural embedding methods. The first computed a token overlap between paragraphs and sentences using the ROUGE toolkit and treated that as a similarity score. The second utilized neural networks to embed the text spans into dense vector representations and next computed the similarity score as the cosine distance between such vectors.

To choose the best similarity score metric we conducted a pilot study on a subset of 100 paragraph-sentence pairs sampled from the validation set. The sampled examples were matched using the procedure described above with different neural models used for embedding the text spans. The following similarity score methods were considered:

- **ROUGE-wtd** (Ladhak et al., 2020) computes an average of token-weighted ROUGE-{1,2,L} scores between the sentence and paragraph texts. Token weights approximate the saliency of words and are computed as an inverse frequency of word occurrences in the document.

- **ROUGE-avg** (Ladhak et al., 2020) computes an average of (unmodified) ROUGE-{1,2,L} scores between the sentence and paragraphs.

- **BERTScore** (Zhang et al., 2020) measures semantic overlap between the words in the sentences and paragraphs. It aligns words in both text spans by maximizing the cosine similarity between BERT representations of the tokens.

- **Cross-Encoder** (Humeau et al., 2019) performs self-attention over the sentence and paragraph text passed together through a Transformer network to generate a similarity score between the input pair.

- **Bi-Encoder** (Reimers and Gurevych, 2019) uses a Transformer architecture to independently encode the sentence and paragraph texts into a dense vector representation. The similarity score is calculated using cosine similarity between the sentence and paragraph representations. We evaluate two checkpoints for the Bi-Encoders as described in Table 4.

The quality of data alignments obtained during the pilot study was assessed by human judges hired through the Amazon Mechanical Turk platform. Workers were hired from English speaking countries and offered a wage of approximately 12 USD per hour. Annotators were shown paragraphs which were aligned with a shared summary sentence using the different methods. For each alignment the annotators were asked to label whether the paragraph-sentence pair is related, somewhat related, or unrelated. Each example was evaluated by three judges, related and somewhat related labels were merged into a single positive label and the majority vote was computed. Results of the study are presented in Table 5 and show the number of times a method was assigned a positive label. The best performing strategy which used a Bi-Encoder fine-tuned on paraphrase detection data.

Using the selected scoring function, paragraph-summary sentence scores were computed between all paragraph-sentence pairs. Next, this data was input into a stable matching algorithms (Gale and Shapley, 1962) to obtain the final alignments. The stable matching procedure creates alignments...
Table 5: Number of times an alignment method received a positive label.

where no paragraph would prefer to be matched with a different summary sentence to which it is already matched, and no summary sentence would prefer to be matched to another paragraph than it is already matched with.

### C Alignment Quality

The quality of alignments obtained using the process described in Section 3.1 and Appendix B was also evaluated quantitatively, results are presented in Table 6 To measure the semantic similarity of source paragraphs and paired summary sentences, the cosine similarity between their embeddings was computed. To measure lexical overlap between the paragraph-summary pairs ROUGE-1 (R-1), ROUGE-2 (R-2), and ROUGE-L (R-L) scores were computed. Results are presented in Table 6.

The cosine similarity of 0.412 indicates strong semantic overlap between the paired sentences and source paragraphs, suggesting high quality pairings. In comparison, the relatively low lexical overlap of 17.39 R-1 between the mentioned fragments highlights the high abstractiveness of the data.

| Model                  | # selected |
|------------------------|------------|
| ROUGE-wtd              | 74         |
| ROUGE-avg              | 66         |
| BERTscore              | 68         |
| Cross Encoder          | 72         |
| Bi-Encoder (paraphrase)| 78         |
| Bi-Encoder (roBERTa)   | 74         |

Table 6: Alignment scores between source paragraphs and paired summary sentences. Semantic similarity evaluated by means of cosine similarity between the embedded text fragments. Lexical similarity evaluated using ROUGE-1, ROUGE-2, and ROUGE-L f-scores.

### D Data Creation Pipeline

The data creation process is visualized in Figure 3.

### E Source examples

Examples of chapter-level summaries of "Sense and Sensibility" collected from different sources are shown in Table 7.

### F Human Evaluation UI

Screenshots of the user interface, including evaluation instructions, used in the human studies of abstractive baselines on the paragraph-level are presented in Figure 4, and on the chapter- and book-level in Figure 5.

### G Model outputs

Example summaries generated on the paragraph-, chapter-, and book-level by the baseline models discussed in our work are presented in Tables 8, 9, 10, 11, 12, 13, 14, 15.
Figure 3: The data collection and pre-processing pipeline used to create the BOOKSUM collection.

Text from "Sense and Sensibility", Chapter 1

The family of Dashwood had long been settled in Sussex. Their estate was large, and their residence was at Norland Park, in the centre of their property, where, for many generations, they had lived in so respectable a manner as to engage the general good opinion of their surrounding acquaintance. The late owner of this estate was a single man, who lived to a very advanced age, and who for many years of his life, had a constant companion and housekeeper in his sister. But her death, which happened ten (…)

Summary from Gradesaver

The Dashwood family is introduced; they live at Norland Park, an estate in Sussex, which has been in their family for many years. Henry Dashwood has a son by a previous marriage, who is well-off because of his long-deceased mother’s fortune; Mr. Dashwood also has three daughters by his present wife, who are left with very little when he dies and the estate goes to his (…)

Summary from Shmoop

We begin with a history of the Dashwood family of Sussex, England: the head of the family, old Mr. Dashwood, dies and distributes his estate among his surviving relatives: his nephew, Henry Dashwood, and his children. The children include one son, John, from a first marriage, and three daughters, Elinor, Marianne, and Margaret, from his second. Even though John and his (…)

Summary from CliffNotes

For many years, Henry Dashwood and his family had lived at Norland Park and cared for its owner, Henry’s aged uncle. On the old man’s death, Henry inherited the estate. He had always expected that he would be free to leave it, in turn, to be shared among his wife and three daughters. John, his son by a previous marriage, was amply provided for. His mother had left him a large (…)

Summary from SparkNotes

Old Mr. Dashwood is the owner of a large estate in Sussex called Norland Park. Following the death of his sister, Mr. Dashwood invites his nephew Mr. Henry Dashwood to come live with him at Norland. The younger Mr. Dashwood brings John Dashwood, his son from a previous marriage, as well as the three daughters born to his present wife. John Dashwood is grown and (…)

Summary from Novelguide

Sense and Sensibility opens by introducing the Dashwood family, whose fortunes the novel follows. The Dashwoods have for many generations owned and occupied the country estate of Norland Park in Sussex, England. The recent owner, Henry Dashwood, inherited the estate from a Dashwood uncle, referred to as “the old Gentleman.” Henry Dashwood has a son, (…)

Summary from BarronBooks

Mr. Henry Dashwood is leading a comfortable and happy life with his family at Norland Estate, which belongs to his uncle. He is the rightful heir to the property. However, after his uncle’s death, it is revealed that his son, John Dashwood, and his grandson, Harry, are to inherit the estate. Mr. Henry Dashwood is obviously disappointed. He is concerned about the welfare of his (…)

Table 7: Examples of chapter-level summaries of "Sense and Sensibility" collected from different sources. Text spans underlined with the same color highlight the high semantic and low lexical overlap between the summaries indicating that the summaries are highly abstractive.
Instructions

In this task you will evaluate the quality of summaries of paragraphs of a book. To correctly solve this task, follow these steps:

1. Carefully read each paragraph, be aware of the information they contain.
2. Carefully read the summaries, be aware of the information they contain.
3. Rate each summary on a scale from 1 (worst) to 5 (best) by its fluency, coherence, relevance, and factuality.

Definitions

Fluency

This rating measures the grammatical quality of the sentences of the summary text. It evaluates whether sentences are well-written and grammatically correct.

You should focus on the quality of individual sentences.

Coherence:

This rating measures the grammatical quality of the whole summary text. It evaluates how well sentences flow together.

You should focus on the quality of the whole text.

Relevance:

This rating measures how relevant the facts included in the summary are to the summary of the paragraph. It evaluates whether the summary does not contain unnecessary information.

You should take into account the summary content and the paragraph text.

Factuality:

This rating measures whether the facts included in the summary are aligned with the facts from the paragraph. It evaluates whether the summary does not contain fabricated facts.

You should take into account the summary content and the paragraph text.

---

Figure 4: Screenshot of the User Interface used to evaluate summaries on the paragraph-level.

---

Instructions

In this task you will evaluate the quality of summaries of a part of a book. To correctly solve this task, follow these steps:

1. Carefully read the summaries, be aware of the information they contain.
2. Rate each summary on a scale from 1 (worst) to 5 (best) by its fluency and coherence.

Definitions

Fluency

This rating measures the grammatical quality of the sentences of the summary text. It evaluates whether sentences are well-written and grammatically correct.

You should focus on the quality of individual sentences.

Coherence:

This rating measures the grammatical quality of the whole summary text. It evaluates how well sentences flow together.

You should focus on the quality of the whole text.

---

Figure 5: Screenshot of the User Interface used to evaluate summaries on the chapter- and book-level.
The family of Dashwood had long been settled in Sussex. Their estate was large, and their residence was at Norland Park, in the centre of their property, where, for many generations, they had lived in so respectable a manner as to engage the general good opinion of their surrounding acquaintance. The late owner of this estate was a single man, who lived to a very advanced age, and who for many years of his life, had a constant companion and housekeeper in his sister. But her death, which happened ten years before his own, produced a great alteration in his home; for to supply her loss, he invited and received into his house the family of his nephew Mr. Henry Dashwood, the legal inheritor of the Norland estate, and the person to whom he intended to bequeath it. In the society of his nephew and niece, and their children, the old Gentleman’s days were comfortably spent. His attachment to them all increased. The constant attention of Mr. and Mrs. Henry Dashwood to his wishes, which proceeded not merely from interest, but from goodness of heart, gave him every degree of solid comfort which his age could receive; and the cheerfulness of the children added a relish to his existence.

Reference Summary - Paragraph-level (Shmoop)
We begin with a history of the Dashwood family of Sussex, England: the head of the family, old Mr. Dashwood, dies and distributes his estate among his surviving relatives: his nephew, Henry Dashwood, and his children. Tragically, this arrangement doesn’t last long - Henry Dashwood dies, and his estate, including the money he’d recently inherited from his uncle, is re-distributed amongst his wife and children.

| Model Output: CNN-LSTM | the family of dashwood had long been settled in sussex . |
|------------------------|--------------------------------------------------------|
| Model Output: BertExt  | the family of dashwood had long been settled in sussex . |
| Model Output: MatchSum| But her death, which happened ten years before his own, produced a great alteration in his home; for to supply her loss, he invited and received into his house the family of his nephew Mr. Henry Dashwood, the legal inheritor of the Norland estate, and the person to whom he intended to bequeath it. |
| Model Output: BART     | The Dashwood family has been settled in Sussex for generations. The late owner of the estate was a single man, who lived to a very advanced age and had a constant companion and housekeeper in his sister. His sister died ten years before his own, and he invited the family of his nephew, Henry Dashwood, into his house. |
| Model Output: BART     | The story of the death of Henry Dashwood, 1st Earl of Norland, is told in a series of letters written by his son and daughter-in-law. |
| Model Output: T5       | The Dashwood family has been settled in Sussex for many generations. |
| Model Output: T5       | the late owner of the norland estate was a single man, who lived to a very advanced age . to supply her loss, he invited and received into his house the family of his nephew . the old Gentleman’s days were comfortably spent with his nephew and niece . |
| Model Output: PEGASUS  | The Dashwood family has been settled in Sussex for many generations. Their estate is large, and their residence is at Norland Park, in the center of their property. The late owner of this estate was a single man, who lived to a very advanced age, and who for many years of his life had a constant companion and housekeeper in his sister. |
| Model Output: PEGASUS  | The late owner of this estate was a single man, who lived to a very advanced age, and who for many years of his life, had a constant companion and housekeeper in his sister. |

Table 8: Examples of decoded summaries for one paragraph of “Sense and Sensibility, Chapter 1”.

6551
Paragraph from “Sense and Sensibility”, Chapter 1

The family of Dashwood had long been settled in Sussex. Their estate was large, and their residence was at Norland Park, in the centre of their property; where, for many generations, they had lived in so respectable a manner as to engage the general good opinion of their surrounding acquaintance. The late owner of this estate was a single man, who lived to a very advanced age, and who for many years of his life, had a constant companion and housekeeper in his sister. But her death, which happened (...)

Reference Summary - Chapter-level (Shmoop)

We begin with a history of the Dashwood family of Sussex, England: the head of the family, old Mr. Dashwood, dies and distributes his estate among his surviving relatives: his nephew, Henry Dashwood, and his children. The children include one son, John, from first marriage, and three daughters, Elinor, Marianne, and Margaret, from his second. Even though John and his (...)

Model Output: CNN-LSTM

the late owner of this estate was a single man, who lived to a very advanced age, and who for many years of his life, had a constant companion and housekeeper in his sister. the family of dashwood had long been settled in sussex. no sooner was his father’s funeral over, than mrs. john dashwood, without sending any notice of her intention to her mother-in-law, arrived with her child and their attendants; by a former marriage, mr. henry dashwood had one son; by his present lady; three daughters; so acutely did mrs. dashwood feel this ungracious behaviour, and so earnestly did she despise her daughter-in-law for it, that, on the arrival of the latter, she would have quitted the house for ever, had not the entreaty of her eldest girl induced her first to reflect on the propriety of going, and her own tender love for all her three children determined her afterwards to stay, and for their sakes avoid a breach with their brother. the resemblance between her and her mother was strikingly great. he was not an ill-disposed young man, unless to be rather cold-hearted and rather selfish is to be ill-disposed: but he was, in general, well respected; for he conducted himself with propriety in the discharge of his ordinary duties: elinor, this eldest daughter, whose advice was so effectual, possessed a strength of understanding, and coolness of judgment, which qualified her, though only nineteen, to be the counsellor of her mother, and enabled her frequently to counteract, to the advantage of all, the eagerness of mind in mrs. dashwood which must generally have led to imprudence. but her death, which happened ten years before his own, produced a great alteration in his home; for to supply her loss, he invited and received into his house the family of his nephew mr. henry dashwood, the legal heir of the norland estate, and the person to whom he intended to bequeath it. he meant not to be unkind, however, and, as a mark of his affection for the three girls, he left them a thousand pounds a-piece. mrs. dashwood had not the strong feelings of the rest of the family: but he was affected by a recommendation of such a nature at such a time, and he promised to do everything in his power to make them comfortable. she was sensible and clever; but eager in everything: her sorrows, her joys, could have no moderation.

Model Output: BertEx

Margaret, the other sister, was a good-humoured, well-disposed girl; but as she had already imbibed a good deal of Marianne’s romance, without having much of her sense, she did not, at thirteen, but fear to equal her sisters at a more advanced period of life. Marianne’s abilities were, in many respects, quite equal to elinor’s: when he gave his promise to his father, he mediated within himself to increase the fortunes of his sisters by the present of a thousand pounds a-piece. Elinor saw, with concern, the excess of her sister’s sensibility; but by mrs. dashwood it was valued and cherished: by a former marriage, mr. henry dashwood had one son; by his present lady; three daughters; so acutely did mrs. dashwood feel this ungracious behaviour, and so earnestly did she despise her daughter-in-law for it, that, on the arrival of the latter, she would have quitted the house for ever, had not the entreaty of her eldest girl induced her first to reflect on the propriety of going, and her own tender love for all her three children determined her afterwards to stay, and for their sakes avoid a breach with their brother. he was not an ill-disposed young man, unless to be rather cold-hearted and rather selfish is to be ill-disposed: but he was, in general, well respected; for he conducted himself with propriety in the discharge of his ordinary duties: elinor, this eldest daughter, whose advice was so effectual, possessed a strength of understanding, and coolness of judgment, which qualified her, though only nineteen, to be the counsellor of her mother, and enabled her frequently to counteract, to the advantage of all, the eagerness of mind in mrs. dashwood which must generally have led to imprudence. but her death, which happened ten years before his own, produced a great alteration in his home; for to supply her loss, he invited and received into his house the family of his nephew mr. henry dashwood, the legal heir of the norland estate, and the person to whom he intended to bequeath it. he meant not to be unkind, however, and, as a mark of his affection for the three girls, he left them a thousand pounds a-piece. mrs. dashwood had not the strong feelings of the rest of the family: but he was affected by a recommendation of such a nature at such a time, and he promised to do everything in his power to make them comfortable. she was sensible and clever; but eager in everything: her sorrows, her joys, could have no moderation.

Model Output: MatchSum

Margaret, the other sister, was a good-humored, well-disposed girl, but as she had already imbibed a good deal of Marianne’s romance, without having much of her sense, she did not, at thirteen, but fear to equal her sisters at a more advanced period of life. No sooner was his father’s funeral over, than Mrs. John Dashwood, without sending any notice of her intention to her mother-in-law, arrived with her child and their attendants. When he gave his promise to his father, he intended not to be unkind, however, and, as a mark of his affection for the three girls, he left them a thousand pounds a-piece. Mr. John Dashwood had never been a favourite with any of her husband’s family; but she had had no opportunity, till the present, of shewing them with how little attention to the comfort of other people she could act when occasion required it. no one could dispute her right to come; the house was her husband’s from the moment of his father’s death; but the indulgence of her conduct was so much the greater, and to a woman in Mrs. Dashwood’s situation, with only common feelings, must have been highly impraising—but in HER mind there was a sense of honor so keen, a generosity so romantic, that any offence of the kind, by whosoever given or received, was to her a source of incommunicable disgust. Mr. John Dashwood had not the strong feelings of the rest of the family: but he was affected by a recommendation of such a nature at such a time, and he promised to do everything in his power to make them comfortable. His son was sent for as soon as his danger was known, and to him Mr. Dashwood recommended, with all the strength and urgency which illness could command, the interest of his mother-in-law and sisters. So acutely did Mrs. Dashwood feel this ungracious behaviour, and so earnestly did she despise her daughter-in-law for it, that, on the arrival of the latter, she would have quitted the house for ever, had not the entreaty of her eldest girl induced her first to reflect on the propriety of going, and her own tender love for all her three children determined her afterwards to stay, and for their sakes avoid a breach with their brother. The prospect of four thousand a-year, in addition to his present income, besides the remaining half of his own mother’s fortune, warmed his heart, and made him feel capable of generosity—Elinor, this eldest daughter, whose advice was so effectual, possessed a strength of understanding, and coolness of judgment, which qualified her, though only nineteen, to be the counsellor of her mother, and enabled her frequently to counteract, to the advantage of all, the eagerness of mind in mrs. dashwood which must generally have led to imprudence. but her death, which happened ten years before his own, produced a great alteration in his home; for to supply her loss, he invited and received into his house the family of his nephew mr. henry dashwood, the legal heir of the norland estate, and the person to whom he intended to bequeath it. Mr. Dashwood’s disappointment was, at first, severe; but his temper was cheerful and sangueous; and he might reasonably hope to live many years, and by living economically, lay by a considerable sum from the produce of an estate already large, and capable of almost immediate improvement. Mrs. John Dashwood had never been a favourite with any of her husband’s family; but she had had no opportunity, till the present, of shewing them with how little attention to the comfort of other people she could act when occasion required it.

Table 9: Examples of decoded summaries of the Chapter 1 of “Sense and Sensibility”, part 1.

| Paragraph from “Sense and Sensibility”, Chapter 1 | Model Output: NVIDIA R2022 with copy-attention | Model Output: MatchSum | Model Output: CNN-LSTM |
|--------------------------------------------------|-----------------------------------------------|------------------------|------------------------|
| The family of Dashwood had long been settled in Sussex. Their estate was large, and their residence was at Norland Park, in the centre of their property; where, for many generations, they had lived in so respectable a manner as to engage the general good opinion of their surrounding acquaintance. The late owner of this estate was a single man, who lived to a very advanced age, and who for many years of his life, had a constant companion and housekeeper in his sister. But her death, which happened (...) | Margaret, the other sister, was a good-humored, well-disposed girl, but as she had already imbibed a good deal of Marianne’s romance, without having much of her sense, she did not, at thirteen, but fear to equal her sisters at a more advanced period of life. | Margaret, the other sister, was a good-humored, well-disposed girl, but as she had already imbibed a good deal of Marianne’s romance, without having much of her sense, she did not, at thirteen, but fear to equal her sisters at a more advanced period of life. | the late owner of this estate was a single man, who lived to a very advanced age, and who for many years of his life, had a constant companion and housekeeper in his sister. the family of dashwood had long been settled in sussex. no sooner was his father’s funeral over, than mrs. john dashwood, without sending any notice of her intention to her mother-in-law, arrived with her child and their attendants; by a former marriage, mr. henry dashwood had one son; by his present lady; three daughters; so acutely did mrs. dashwood feel this ungracious behaviour, and so earnestly did she despise her daughter-in-law for it, that, on the arrival of the latter, she would have quitted the house for ever, had not the entreaty of her eldest girl induced her first to reflect on the propriety of going, and her own tender love for all her three children determined her afterwards to stay, and for their sakes avoid a breach with their brother. the resemblance between her and her mother was strikingly great. he was not an ill-disposed young man, unless to be rather cold-hearted and rather selfish is to be ill-disposed: but he was, in general, well respected; for he conducted himself with propriety in the discharge of his ordinary duties: elinor, this eldest daughter, whose advice was so effectual, possessed a strength of understanding, and coolness of judgment, which qualified her, though only nineteen, to be the counsellor of her mother, and enabled her frequently to counteract, to the advantage of all, the eagerness of mind in mrs. dashwood which must generally have led to imprudence. but her death, which happened ten years before his own, produced a great alteration in his home; for to supply her loss, he invited and received into his house the family of his nephew mr. henry dashwood, the legal heir of the norland estate, and the person to whom he intended to bequeath it. he meant not to be unkind, however, and, as a mark of his affection for the three girls, he left them a thousand pounds a-piece. mrs. dashwood had not the strong feelings of the rest of the family: but he was affected by a recommendation of such a nature at such a time, and he promised to do everything in his power to make them comfortable. she was sensible and clever; but eager in everything: her sorrows, her joys, could have no moderation. |

6552
The Dashwood family has been settled in Sussex for generations. The late owner of the estate was a single man, who lived to a very advanced age and had a constant companion and housekeeper in his sister. His sister died ten years before his own, and he invited the family of his nephew, Henry Dashwood, into his house. Mr. Dashwood has three daughters, one of whom is his son, and the other two are his sisters. Margaret is a good-humored, well-disposed girl, but as she had already imbibed a good deal of Marianne’s romance without having much of her sense, she did not, at thirteen, bid fair to equal her sisters at a more advanced age. Mr. Dashwood dies and leaves his estate to his nephew, a four-year-old boy. Mr. Dashwood’s health begins to fail, and he is told that he will not live long. He was not an ill-disposed young man, unless to be rather cold hearted and rather selfish, but he was, in general, well respected. He promises to give his sisters a thousand pounds a year. Mrs. John Dashwood arrives at her husband’s funeral without telling her mother-in-law. The eldest daughter, Elinor, is nineteen years old, and she has a “strength of understanding, and correctness of judgment, which qualified her, though only nineteen, to be the counsellor of her mother.” Marianne is described as “sensible and clever, but eager in everything; her sorrows, her joys, could have no moderation.” Elinor and her sister-in-law, Mrs. Dashwood, are deeply affected by their mother’s death.

Table 10: Examples of decoded summaries of the Chapter 1 of “Sense and Sensibility”, part 2.
The family of Dashwood had long been settled in Sussex. Their estate was large, and their residence was at Norland Park, in the centre of their property, where, for many generations, they had lived in so respectable a manner as to engage the general good opinion of their surrounding acquaintance. The late owner of this estate was a single man, who lived to a very advanced age, and with forty years of his life, had a constant companion and housekeeper in his sister. But her death, which happened ...

Reference Summary - Book-level (Shimoop)
We begin with a summary of the Dashwood family of Sussex, England: the head of the family, old Mr. Dashwood, dies and distributes his estate among his surviving relatives. His nephew, Henry Dashwood, and his children. The children include one son, John, from a first marriage, and three daughters, Elinor, Marianne, and Margaret, from his second. Even though John and his (...

Model Output: CNN-LSTM

...they were interrupted by the entrance of margaret , and elinor was then at liberty to think over the representations of her mother , to acknowledge the probability of many , and hope for the justice of all . mrs. palmer ’s information respecting willoughby was not very material ; but any testimony in her favour , however small , was pleasing to her . elinor was alternately shamed and disappointed , but marianne preserved , and saw every night in the brightness of the fire , and every morning in the appearance of the atmosphere , the certain symptoms of approaching frost . “ it is not every one , “ said elinor , “ who has your passion for dead leaves . mrs. dashwood again assured him that no alteration of the kind should be attempted . mrs. dashwood , persuaded that such an excursion would be productive of much amusement to both her daughters , and perceiving through all her affectionate attention to herself , how much the...

Model Output: BertExt

...and myself–and i resolved therefore on calling at the cottage, in my way to honiton. yet, though smiling within herself at the mistake, she honoured her sister for that blind partiality to edward...

Table 11: Examples of decoded summaries of the full text of “Sense and Sensibility”, part 1.

| Example | Summary |
|---------|---------|
| ...the shock of colonel brandon’s death at Barton was much softened to mrs. dashwood by her own previous alarm; for so great was her unsoundness about marianne, that she had already determined not to set out for clapham on that very day, without waiting for any further intelligence... | mrs. dashwood looked at elinor, and said, “i am sorry you are not more accustomed to understand how much you may be beloved, before you come to be beloved. she, who was, on her own account, much less applied than an unpleasant as she now were, would represent in the strongest manner the necessity of being engaged with the full expectations of the party. mrs. palmer appeared quite well, and i am commissioned to tell you, that you will certainly see her to-morrow.” i am persuaded that mrs. smith suspects her regard for marianne, disapproves of it, (perhaps because she has other views for him,) and on that account is eager to get him away—and that the business which she sends him off to transact is invented as an excuse to dismiss him. i confess, “replied mrs. ferrars, ‘that every circumstance except one is in favour of that engagement, but that one is the total silence of both on the subject, and with me it is overweighs every other’ “she is not well, she has had a nervous complaint on several occasions for which she will be more hurt by it, for rob always was her favourite—she will be more hurt by it, and on the same principle will forgive him much sooner.” the parties stood thus: the two mothers, though each really believed that her own son was the tallest, politely decided in favour of the other. mrs. ferrars is to be the man. well, said elinor, who, though pitying her, grew impatient for his departure; “and this is all!” i have entered many a shop to avoid your sight, as the carriage drove by. she fell into violent hysterics immediately, with such screams as reached your brother’s ears, as he was sitting in his own dressing-room door stairs, thinking about writing a letter to his steward in the country: affecting that air of playfulness, therefore, which is delightful in a woman one loves, she opened the letter directly, and read its contents, it came from town, and is merely a letter of business.” lady edwills was delighted with the thought. he wished a good morning, and, addressed by sir john, left the room. it has been, and is, and will probably always be a heavy misfortune to me, that i have had no necessary business to engage me, no profession to give employment, or affright me any thing like independence has; not his behaviour to marianne and to all of us, for at least the last fortnight, declared that he loved and considered her as his future wife, and that he felt for the attachment of the nearest relation? well; then, he is fairly dead, marianne, for i am sure there was such a man once, and his name begins with an “f” mrs. ferrars, with the utmost liberality, will come forward, and settle on her a thousand, if the match takes place. sir john’s confidence in his own horse was confirmed by this spirited praise, and he set out accordingly, for the cottage to tell the miss dashwoods of the miss ferrars’ arrival, and to invite her to...
The morning was chiefly spent in leaving cards at the houses of Mrs. Jennings's acquaintance to inform them of her being in town; and Marianne was all the time busy in observing the direction of the wind, the variations of the sky, and imagining an alteration in the air: "Oh!" cried Miss Steele, "I dare say Lucy will be as wise as Miss Dashwood." An opportunity was soon given to Marianne of refreshing her acquaintance with the Dowagers of Chadleigh who, at last, long after it was observable to everybody else—on her—what could she do? She had heard her mother's voice, but Elton did not feel very compassionate. Again he stopped to recover himself, and Elton spoke her feelings in an explanation of tender concern, at the fate of his unfortunate friend. They were interrupted by the entrance of Margaret, and Elton was then at liberty to think over the transactions of the morning. To plai...
A continuance in a place where everything reminded her of former delight, was exactly what Mrs. Dashwood wanted. She essayed it, and found it; the provision of security.

Table 13: Examples of decoded summaries of the full text of “Sense and Sensibility”, part 3.

“but I have not known you long to be sure, personally at least, but I have known you and all your family by description a great while; and as soon as I saw you, I felt almost as if you was an old woman of seven and twenty can never hope to feel or inspire affection again, and if her home be uncomfortable, or her fortune small, she might bring herself to submit to the offices of a nurse, which will confirm it. Marianne is astonished to find how much the imagination of her mother and herself had outstrip the truth. Mrs. Dashwood decides that it would be better for Marianne to superintend the progress of the Parsonage. She was born to discover the falsehood of her own opinions, and to counteract, by her conduct, her most favourite maxims. Marianne says that a happiness, of a man who, to every advantage of person and talents, united a disposition naturally open and honest, and a feeling, affectionate temper. Extravagance and vanity had made him of illness, fancied that all relief might soon be in vain, that everything had been delayed too long, and pictured to herself her suffering mother arriving too late to see this darling child. At seven years of age, thoughts are still fixed incoherently on her mother, and whenever she mentions her name, it gives a pang to the heart of poor Elinor, who, reproaching herself for having trifled with so many days next morning, Elinor receives a letter from Lucy. Marianne’s pianoforte is unpacked and properly disposed of; Elinor’s drawings are affixed to the walls of their sitting room. Marianne’s thoughts are still fixed on her mother, and whenever she mentions her name, it gives a pang to the heart of poor Elinor, who, reproaching herself for having trifled with so many days of illness, fancied that all relief might soon be in vain, that everything had been delayed too long, and pictured to herself her suffering mother arriving too late to see this darling child. At seven years of age, thoughts are still fixed incoherently on her mother, and whenever she mentions her name, it gives a pang to the heart of poor Elinor, who, reproaching herself for having trifled with so many days next morning, Elinor receives a letter from Lucy. Marianne’s pianoforte is unpacked and properly disposed of; Elinor’s drawings are affixed to the walls of their sitting room. Marianne’s thoughts are still fixed on her mother, and whenever she mentions her name, it gives a pang to the heart of poor Elinor, who, reproaching herself for having trifled with so many days of illness, fancied that all relief might soon be in vain, that everything had been delayed too long, and pictured to herself her suffering mother arriving too late to see this darling child. At seven years of age, thoughts are still fixed incoherently on her mother, and whenever she mentions her name, it gives a pang to the heart of poor Elinor, who, reproaching herself for having trifled with so many days...
a gentleman carrying a gun, with two pointers playing round him, was passing up the hill and within a few yards of Marianne, when her accident happened. he put down his gun and ran to her assistance; she had raised herself from the ground, but her foot had been twisted, and carried her down the hill. "in defence of your protege you can even be saucy." "My protege, as you call him, is a sensible man," said Elinor, "he has been at the courts of which she could not be only too grateful, though he must have seen enough to be sensible he is very capable of making a woman sincerely attached to him." "you are right, my love; it will be better that there should be no untruth in the case, whatever I may give them occasionally will be of far greater assistance than a yearly allowance." "a present of fifty pounds, now and then, will prevent their ever being distressed for money, and will, I think, be amply disappointing her to the father." "if you were to see them at the altar, you would advise two, to set off for town, when you were tired of barton, without saying a word to miss dashwood about it." "then would have thought myself at liberty to bestow my own on another, and have no need to make any resolutions on remaining at nadolf no longer than was unavoidable. to separate Edward and Elizer was as far from being her object as ever; and she wished to show Mrs. John dashboard how totally she disregarded her dissatisfaction of the match. my love, i am convinced that if it was in my power, but i had just set him down a week before it was my opinion that there would be no annuity in the case; whatever I may give them occasionally will be of far greater assistance than a yearly allowance. a present of fifty pounds, now and then, will prevent their ever being distressed for money, and will, I think, be amply disappointing her to the father."

Table 14: Examples of decoded summaries of the full text of “Sense and Sensibility”, part 4.
The Dashwood family has been settled in Sussex for many generations. Their estate is large, and their residence at Norland Park, in the center of their property. The late owner of this estate was a simple man, who lived to a very advanced age, and who for many years of his life had a constant companionship and housekeeper in his sisters. The Dashwoods are now settled at Barton with two of the oldest sisters. The house and the property are now become familiar. Mrs. Palmer is a woman of uncommon attraction, that open, affable, and lively manner, which it was no merit to possess; and by that still ardent love for Marianne. Colonel Brandon, the friend of Sir John, seemed no more adapted by resemblance of manner to be his friend than Dashwood could have been described as the only happiness worth a wish. Edward was allowed to retain the privilege of first comer, and the baronet, with the utmost sang-froid, and go away without making her any reply! Sir John's satisfaction in society is much more real. He is a blessing to all the juvenile part of the family party which would then be restored, of their mutual pursuits and cheerful society, as the only happiness worth a wish. Edward was allowed to retain the privilege of first comer, and the baronet, with the utmost sang-froid, and go away without making her any reply! Sir John's satisfaction in society is much more real. He is a blessing to all the juvenile part of the family party which would then be restored, of their mutual pursuits and cheerful society, as the only happiness worth a wish. Edward was allowed to retain the privilege of first comer, and the baronet, with the utmost sang-froid, and go away without making her any reply! Sir John's satisfaction in society is much more real. He is a blessing to all the juvenile part of the family party which would then be restored, of their mutual pursuits and cheerful society, as the only happiness worth a wish. Edward was allowed to retain the privilege of first comer, and the baronet, with the utmost sang-froid, and go away without making her any reply! Sir John's satisfaction in society is much more real. He is a blessing to all the juvenile part of the family party which would then be restored, of their mutual pursuits and cheerful society, as the only happiness worth a wish. Edward was allowed to retain the privilege of first comer, and the baronet, with the utmost sang-froid, and go away without making her any reply! Sir John's satisfaction in society is much more real. He is a blessing to all the juvenile part of the family party which would then be restored, of their mutual pursuits and cheerful society, as the only happiness worth a wish. Edward was allowed to retain the privilege of first comer, and the baronet, with the utmost sang-froid, and go away without making her any reply! Sir John's satisfaction in society is much more real. He is a blessing to all the juvenile part of the family party which would then be restored, of their mutual pursuits and cheerful society, as the only happiness worth a wish. Edward was allowed to retain the privilege of first comer, and the baronet, with the utmost sang-froid, and go away without making her any reply! Sir John's satisfaction in society is much more real. He is a blessing to all the juvenile part of the family party which would then be restored, of their mutual pursuits and cheerful society, as the only happiness worth a wish. Edward was allowed to retain the privilege of first comer, and the baronet, with the utmost sang-froid, and go away without making her any reply! Sir John's satisfaction in society is much more real. He is a blessing to all the juvenile part of the family party which would then be restored, of their mutual pursuits and cheerful society, as the only happiness worth a wish. Edward was allowed to retain the privilege of first comer, and the baronet, with the utmost sang-froid, and go away without making her any reply! Sir John's satisfaction in society is much more real. He is a blessing to all the juvenile part of the family party which would then be restored, of their mutual pursuits and cheerful society, as the only happiness worth a wish. Edward was allowed to retain the privilege of first comer, and the baronet, with the utmost sang-froid, and go away without making her any reply! Sir John's satisfaction in society is much more real. He is a blessing to all the juvenile part of the family party which would then be restored, of their mutual pursuits and cheerful society, as the only happiness worth a wish. Edward was allowed to retain the privilege of first comer, and the baronet, with the utmost sang-froid, and go away without making her any reply! Sir John's satisfaction in society is much more real. He is a blessing to all the juvenile part of the family party which would then be restored, of their mutual pursuits and cheerful society, as the only happiness worth a wish. Edward was allowed to retain the privilege of first comer, and the baronet, with the utmost sang-froid, and go away without making her any reply! Sir John's satisfaction in society is much more real. He is a blessing to all the juvenile part of the family party which would then be restored, of their mutual pursuits and cheerful society, as the only happiness worth a wish. Edward was allowed to retain the privilege of first comer, and the baronet, with the utmost sang-froid, and go away without making her any reply! Sir John's satisfaction in society is much more real. He is a blessing to all the juvenile part of the family party which would then be restored, of their mutual pursuits and cheerful society, as the only happiness worth a wish. Edward was allowed to retain the privilege of first comer, and the baronet, with the utmost sang-froid, and go away without making her any reply! Sir John's satisfaction in society is much more real. He is a blessing to all the juvenile part of the family party which would then be restored, of their mutual pursuits and cheerful society, as the only happiness worth a wish. Edward was allowed to retain the privilege of first comer, and the baronet, with the utmost sang-froid, and go away without making her any reply! Sir John's satisfaction in society is much more real. He is a blessing to all the juvenile part of the family party which would then be restored, of their mutual pursuits and cheerful society, as the only happiness worth a wish. Edward was allowed to retain the privilege of first comer, and the baronet, with the utmost sang-froid, and go away without making her any reply! Sir John's satisfaction in society is much more real. He is a blessing to all the juvenile part of the family party which would then be restored, of their mutual pursuits and cheerful society, as the only happiness worth a wish. Edward was allowed to retain the privilege of first comer, and the baronet, with the utmost sang-froid, and go away without making her any reply! Sir John's satisfaction in society is much more real. He is a blessing to all the juvenile part of the family party which would then be restored, of their mutual pursuits and cheerful society, as the only happiness worth a wish. Edward was allowed to retain the privilege of first comer, and the baronet, with the utmost sang-froid, and go away without making her any reply! Sir John's satisfaction in society is much more real. He is a blessing to all the juvenile part of the family party which would then be restored, of their mutual pursuits and cheerful society, as the only happiness worth a wish. Edward was allowed to retain the privilege of first comer, and the baronet, with the utmost sang-froid, and go away without making her any reply! Sir John's satisfaction in society is much more real. He is a blessing to all the juvenile part of the family party which would then be restored, of their mutual pursuits and cheerful society, as the only happiness worth a wish. Edward was allowed to retain the privilege of first comer, and the baronet, with the utmost sang-froid, and go away without making her any reply! Sir John's satisfaction in society is much more real. He is a blessing to all the juvenile part of the family party which would then be restored, of their mutual pursuits and cheerful society, as the only happiness worth a wish. Edward was allowed to retain the privilege of first comer, and the baronet, with the utmost sang-froid, and go away without making her any reply! Sir John's satisfaction in society is much more real. He is a blessing to all the juvenile part of the family party which would then be restored, of their mutual pursuits and cheerful society, as the only happiness worth a wish. Edward was allowed to retain the privilege of firstS