# Controlled Text Reduction

**Aviv Slobodkin, Paul Roit, Eran Hirsch, Ori Ernst, Ido Dagan**

Bar-Ilan University

{lovodkin93,plroit,hirsch.eran,oriern}@gmail.com

dagan@cs.biu.ac.il

## Abstract

Producing a reduced version of a source text, as in generic or focused summarization, inherently involves two distinct subtasks: deciding on targeted content and generating a coherent text conveying it. While some popular approaches address summarization as a single end-to-end task, prominent works support decomposed modeling for individual subtasks. Further, semi-automated text reduction is also very appealing, where users may identify targeted content while models would generate a corresponding coherent summary.

In this paper, we focus on the second subtask, of generating coherent text given pre-selected content. Concretely, we formalize *Controlled Text Reduction* as a standalone task, whose input is a source text with marked spans of targeted content ("highlighting"). A model then needs to generate a coherent text that includes all and only the target information. We advocate the potential of such models, both for modular fully-automatic summarization, as well as for semi-automated human-in-the-loop use cases. Facilitating proper research, we crowdsource high-quality dev and test datasets for the task. Further, we automatically generate a larger "silver" training dataset from available summarization benchmarks, leveraging a pretrained summary-source alignment model. Finally, employing these datasets, we present a supervised baseline model, showing promising results and insightful analyses.[1]

## 1 Introduction

Abstractive text summarization takes one or more documents as input and aims at generating an accurate and coherent summary from it. It requires both locating salient information in the input and then generating a concise text covering it. While some modern state-of-the-art abstractive summarization models treat the task as a single end-to-end task, it has been common practice for summarization models to separate the salience detection phase from the text generation phase (Barzilay and McKeown, 2005; Oya et al., 2014; Banerjee et al., 2016; Vilca and Cabezudo, 2017), with renewed popularity in recent years (Lebanoff et al., 2019, 2020a,b; Xiao et al., 2022; Ernst et al., 2021a; Gehrmann et al., 2018a; Chen and Bansal, 2018; Cho et al., 2019). But, though those proposed techniques comprised distinguishable subtasks, evaluation was performed on the whole summarization pipeline, rather than optimizing each step separately.

In this paper, we focus on the text generation step, while addressing it as a standalone task at the sub-sentence level. To that end, we introduce a new task which we denote *Controlled Text Reduction*. The task takes as input a document with pre-chosen salient spans in it, which we will henceforth call *highlights*. A model is then expected to reduce the document to a smaller coherent text which covers all and only the highlighted content, i.e., consolidating the highlighted spans into a fluent and coherent passage, as exemplified in Figure 1. This task poses a challenge, as it requires generating fluent and grammatical text from non-consecutive spans while keeping it faithful to the source document. Hence, to balance the coherency and faithfulness constraints, models will be expected to use the context document to fill in implied details and to properly connect the different spans.

Focusing on this task can facilitate greater control over the generated text. It could lead to a modular summarization pipeline, where text-generation models can be trained once, and then used with different content selections to accommodate different needs. For example, we may envision a user (e.g., a student) pre-selecting the desirable textual content (either manually or via a designated model) while
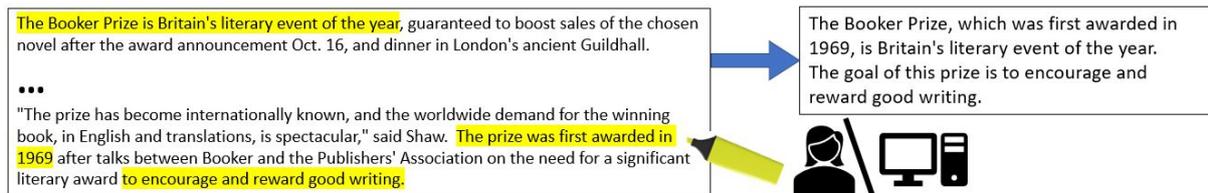
---

Figure 1: An example of an input, consisting of a source document and highlights (left), and the generated passage covering the highlighted content while preserving coherence (right). Such highlights in realistic use cases may be produced either by a human user or by a salience detection model.

focusing on personal needs, possibly interactively (Hirsch et al., 2021; Shapira et al., 2021). Then, an available controlled text reduction module would transform the pre-selected fragments into a concise summary. Also, separating the content selection and generation stages can lead to developing data-efficient systems, one to model salient content and another to generate the text. It could also lead to a more efficient characterization and research of each step separately without the need for probing, which is the prevailing approach in end-to-end models (Conneau et al., 2018; Tenney et al., 2019a,b; Slobodkin et al., 2021; Pandit and Hou, 2021).

To promote research on the advocated text reduction task, we first develop a suitable controlled crowdsourcing methodology, following Roit et al. (2020), and apply it to produce high-quality dev and test datasets (§4). Next, we automatically generate a larger training dataset, by aligning propositional units of information (Ernst et al., 2021b), extracted with OpenIE (Stanovsky et al., 2018), between source documents and their summaries (§5). We use this data to train an abstractive supervised model, and evaluate its performance against our testset while comparing it to an extractive reference baseline, which simply concatenates the highlights. We also perform analyses where we manipulate the highlights and show that the addition of highlights to a supervised model is helpful in steering the model toward the pre-selected content, in addition to improving overall faithfulness and fluency (§8).

Hence, the contribution of this paper is manifold:

1. Proposing the *"Controlled Text Reduction"* task as a standalone module in automated or semi-automated use cases.
2. Defining an intuitive and easy-to-reproduce crowd-sourcing method for the task.
3. Constructing the first data suite for the task, including crowd-sourced dev and test sets and an automatically-generated train set.
4. Developing a supervised baseline model for future work.

## 2 Background

In this section, we briefly review related work and discuss the limitations of their framing.

As mentioned above, much of the related previous work focused primarily on end-to-end summarization (Carbonell and Goldstein, 1998; Haghighi and Vanderwende, 2009; Nallapati et al., 2016c,b; Paulus et al., 2017; Gehrmann et al., 2018b), with the vast majority of related datasets aimed at end-to-end summarization (Fabbri et al., 2019; Kim et al., 2019; Ghalandari et al., 2020), with only a source document as input. On the other hand, research on leveraging control through the injection of pre-chosen (rather than learned) signals in the seq-to-seq scenario focused mostly on semantic and syntactic signals, and also almost exclusively targeted Machine Translation models (Bugliarello and Okazaki, 2020; Akoury et al., 2019; Sundararaman et al., 2019; Choshen and Abend, 2021; Slobodkin et al., 2022).

Attempts to leverage some control over the generation step in summarization received attention in recent years in the form of query-focused summarization (Baumel et al., 2018; Xu and Lapata, 2020, 2021; Wei and Zhizhuo, 2017) and keywords-focused summarization (Keskar et al., 2019; He et al., 2020), with a few recently published corresponding datasets (Pasunuru et al., 2021; Kulkarni et al., 2020; Baumel et al., 2016). A similar trend tried to leverage control through the addition of a planning step (Zhao et al., 2020; Narayan et al., 2021). Although these lines of research allowed for some control over salience, this control was limited and mostly focused on biasing the summary's topic, style, or structure.

The prevailing way to treat summarization in earlier works was to separate the salience detection phase from the text generation phase (Barzilay and McKeown, 2005; Oya et al., 2014; Banerjee et al., 2016; Vilca and Cabezudo, 2017), yet the evaluation was performed on the whole pipeline.
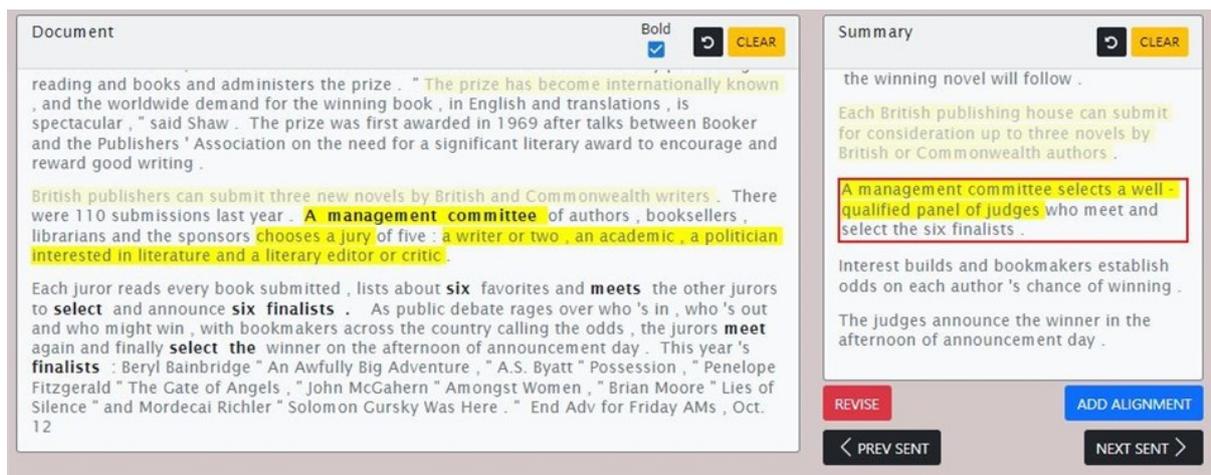
Figure 2: The Highlighting Annotation UI, presenting a document and its corresponding summary. Saved alignments have a faded yellow background, whereas currently selected alignments (which haven't been saved yet) have a normal yellow background. The current summary sentence is marked in a red box. Also, the bold feature is activated, meaning the document words which are related to those in the summary sentence are boldfaced (see §4.1).

Some recent work focused on salience detection (Ernst et al., 2021a,b; Gehrmann et al., 2018a; Chen and Bansal, 2018; Cho et al., 2019), whereas the generation step has mostly been explored in a full-sentence-fusion setting (Geva et al., 2019; Lebanoff et al., 2019, 2020b; Xiao et al., 2022), rather than in a sub-sentence level. Lebanoff et al. (2020a) took it one step further, leveraging sentence fusion through a fine-grained content selection algorithm. But, though they did perform some analysis of this additional step by comparing different salience detection strategies, his evaluation focused on the full pipeline, similarly to his predecessors.

There has also been some work on extracting salient information in source documents in the form of highlights (Cho et al., 2020; Arumae et al., 2019). Yet, though acknowledging the full potential of using highlights to mark salient information in the source document, it mainly focused on the process of obtaining these highlights, overlooking its actual usage in subsequent generation tasks, and in summarization in particular. Moreover, these lines of work focused solely on automatic highlight detection, lacking any crowdsourced annotation scheme. There has also been work that pre-identified salient parts as input to the generation phase (Chen and Bansal, 2018; Xu et al., 2020; Liu et al., 2021; Deutsch and Roth, 2021) But, contrary to our work, the salience detection and generation tasks were addressed and evaluated jointly, without assessing the quality of each individual task.

All those research directions recognized the potential of separating the summarization task into subtasks and performing each subtask explicitly. However, they all evaluated the subtasks jointly, and in doing so overlooked the potential laying in the optimization and characterization of each task individually, and specifically the generation task given content-selection. In this work, we propose to isolate the generation task given pre-selected content, treating it as a stand-alone task, thus promoting focused evaluation and model designing.

## 3 Task Definition

We define the *controlled Text Reduction* task as follows. Given a document and a set of marked spans within that document, denoted as *highlights*, produce a coherent output text encompassing only the information provided within those highlights (see Figure 1). The desirable output should adhere to two requirements beyond coherency: (1) Its content has to be derived from the highlights alone, keeping any additional document premises to the minimum required for coherency; (2) The output has to retain all of the details covered by the highlighted spans.

Such requirements give rise to many interesting challenges, such as recognizing the connecting thread between disparate spans and faithfully representing the information contained within them. We forgo a strict definition for a highlighted span and allow possibly marking sub-sentence elements: an entity or a clause, even discontinuous descriptions of these (e.g., the last two highlights in Figure 1). Hence, the input highlights may be disconnected in both their surface realization (i.e. grammatically
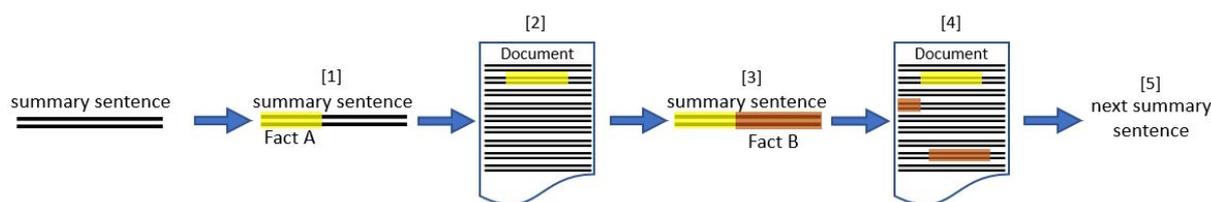
Figure 3: Illustration of Highlighting Annotation process for a summary sentence: [1] A *summary* fact is located and highlighted; [2] The matching *document* spans are highlighted, and the alignment is saved; [3] Another *summary* fact is identified and highlighted; [4] The matching *document* spans are highlighted, and the alignment is saved; [5] When the summary sentence is fully highlighted, we proceed to the next sentence, and so on. In this example, the summary consists of two facts, but steps 1 and 2 can be repeated as needed per sentence, until all its propositions (facts) are covered.

unsuitable), and semantic fluency.

Figure 1 features an input-output example. The output covers exclusively and completely the highlighted information while using the source document's context to connect the disparate spans.

# 4 Gold Dataset for Evaluation

We leverage different summarization datasets to annotate a high-quality dataset for the evaluation of controlled-reduction systems. In summarization, every summary arises from a set of salient document spans. Exploiting this in our annotation process, we wish to "reverse-engineer" each summary and locate the spans in the document that led to its construction. This significantly reduces the annotation complexity and load, instead of compiling a new text given a set of highlighted spans, an annotator has to highlight document spans given the output text (i.e. the summary).

To create our development and test partitions we sample 121 and 108 unique documents from DUC 2001 and 2002 Single-Document-Summarization (SDS) datasets[2] respectively. Each document is accompanied by up to 4 different reference summaries (with an average of 2.14 summaries per document), resulting in a total of 488 unique document-summary pairs (see Table 1 for full statistics and §A for preprocessing details).

We build an intuitive and convenient annotation tool for extracting highlights from document-summary pairs [3], designed to be embedded into crowdsourcing platforms (see §4.1 and Figure 2). Given the complexity of our task, we follow Roit et al. (2020)'s *controlled crowdsourcing* setup, including principled steps of annotator recruitment

and training, leading to a trusted and qualified annotators group, employed for the annotation process.

## 4.1 Annotation Process

To annotate document spans, whose content corresponds to the summary content, we build a web-based user interface that is published on Amazon Mechanical Turk[4] and used by crowd-workers (see Figure 2). An annotator is presented with a document and its reference summary side-by-side and is instructed to highlight all of the phrases in the document whose content corresponds to the summary (see yellow background in Figure 2). To facilitate accurate and systematic processing of each instance, workers are asked to align spans from the summary that comprise a single fact to minimal spans in the document which cover them. Thus, annotators create a series of alignments that cover every piece of information in the summary (see Figure 3 for illustration of the annotation flow).

We observed that processing summary text one fact at a time substantially focuses the annotators' attention and expedites the search for relevant spans in the document. This is exemplified when a single sentence in the summary is comprised of details that are mentioned in different locations spread out across the source document (e.g., the first summary sentence in Figure 1). Further, to streamline the process, we segment the document into paragraphs and bolden content words in the document that share the same lemma with words in the current summary sentence (see document side in Figure 2 and also §A for details). This method helps the human annotator to skim quickly through the document and is relatively bias-free. It is our assumption that a trained worker will not predominantly use same-lemma words for highlighting, as it is discouraged

---

[2] https://duc.nist.gov/
[3] https://github.com/lovodkin93/ highlights-extract-app

[4] www.mturk.com

5702

| | #unique docs | #summaries/doc (average) | #summary-doc pairs | mean input/output size (tkns) | max input/output size (tkns) | mean input/output size (sentences) | summary sentences aligning to multiple doc sentences |
|---|---|---|---|---|---|---|---|
| Train (DUC) | 893 | 2.14 | 1911 | 849.13/115.34 | 8311/153 | 35.73/4.60 | 41.87 % |
| Dev (DUC) | 57 | 2.26 | 129 | 790.95/121.05 | 3079/164 | 27.68/4.44 | 40.62 % |
| Test (DUC) | 172 | 2.09 | 359 | 876.35/120.59 | 3384/161 | 30.84/4.34 | 40.71 % |
| Overall (DUC) | 1122 | 2.14 | 2399 | 850.40/116.44 | 8311/164 | 34.58/4.56 | 41.63 % |
| Train (CNN-DM) | 285073 | 1 | 285073 | 810.77/56.91 | 2934/2100 | 40.07/2.72 | 71.29 % |

Table 1: Statistics of our dataset, including the number of unique documents, the average number of summaries per document, the number of summary-document pairs (a unique document creates a pair with each of its summaries), the mean input/output size (in tokens and in sentences), the maximum input/output size (in tokens) and the percentage of sentences whose alignments span across more than one document sentence.

in our guidelines (see §4.2).

After carefully assembling our trained worker pool, (see later §4.3), each document-summary instance is annotated by a single worker. To supervise the resulting quality, we randomly sample submissions, supplying additional feedback if needed.

## 4.2 Guidelines

We instruct our workers to process the text systematically and align facts from each summary sentence to the corresponding phrases in the document.

**Summary-related Guidelines** We provide guidelines for the annotator to break up the summary sentence into the facts that it is comprised of. We target facts encoded in main or embedded clauses, appositions, copular phrases, conjunctions, and more. §B.1 covers the full summary-related guidelines provided to the annotator.

**Document-related Guidelines** Once a summary fact was identified and highlighted, the crowdworkers are instructed to find its corresponding spans in the document. We define those spans as the minimal set of phrases that fully describe the current highlighted fact in the summary and nothing else. We define minimal in the sense that removing a content word from the document span would necessarily render some detail as not covered. For example, omitting anything from the first summary sentence in Figure 1, e.g., "in 1969", would result in an overlooked highlighted fact. Notably, the annotators may highlight multiple document spans portraying the same fact (redundantly in the document). Finally, we elaborate on the guidelines to touch down on issues such as paraphrasing, inconsecutive highlights, and highlighting in context. A more comprehensive overview of the guidelines and examples appears in §C.

## 4.3 Annotator Training

We follow the Controlled Crowdsourcing Methodology (Roit et al., 2020) to detect a group of qualified annotators, using two open qualification rounds for an initial selection, and proceeding with closed qualification rounds (for selected annotators) for further training and refining. In each round, the annotator is instructed to read a short description of the task and annotate a trial instance.[5] The closed qualification rounds proceeded with a 20-minute video explaining the different features of our annotation tool (see §4.1). Each round is followed by a thorough review of the authors for further feedback. The qualification rounds are fully paid, take up to 30 minutes to complete, and consist of 3 summary-document pairs and reading relevant feedback. Upon completion, we remained with 11 annotators who successfully completed the training session, out of 15 who began the training round.

**Cost** We price every annotation instance, that takes on average 10 minutes to complete, at 2$. We also compensate the workers for the time spent watching the 20-minute video during training with a 4$ bonus upon completion of the video. The total dataset cost amounted to approximately 1400$.

## 4.4 Dataset Quality

To assess the quality of the resulting dataset we calculate different agreement scores between crowdworkers and experts. Given the same summary-document pair annotated separately by two annotators, we calculate Intersection-over-Union (*IOU*) of the tokens' indices[6] between the highlighted document spans that are aligned to the same summary sentence, similarly to Ernst et al. (2021b). We collect the sentence-wise IOU scores across 3 summary-document pairs, annotated by 11 workers to calculate the Inter-Annotator-Agreement and

---

[5]For the open rounds, the instance is simplified with a single summary sentence to focus on.

[6]We consider only content words.

find that our workers exhibit a high agreement of 82.09, suggesting that our annotation protocol is well-defined and stable. Likewise, we calculate the agreement between the annotators to references made by two of the authors and find it to be also high (78.23), indicating a good quality of our annotated data.

From analyzing all disagreements ($IoU <$ $90\%$), we find that the main factor for disagreement stems from two separate spans in the document entailing the same event, resulting in each of the annotators highlighting a different mention of it or in one of them highlighting both mentions. This does not harm the quality of our data, as both options are fitting for the task. Another prevalent reason for disagreement arises from one of the annotators highlighting extra phrases that overall add only insignificant details on top of the summary. For examples, see §D. Finally, an interesting characteristic of our dataset is that for $> 40\%$ our annotated data, a summary sentence is aligned with non-consecutive phrases originating in different document sentences (see Table 1), representing the challenges faced by a text reduction model in a realistic setting.

## 5   Train Dataset

To acquire a larger dataset for training supervised models, we opt for an automatic approach to extract highlights. For that, we employ the superPAL model (Ernst et al., 2021b), a proposition-based summary-source alignment model trained on a sentence alignment dataset (Copeck and Szpakowicz, 2005; Copeck et al., 2006, 2007, 2008) based on the Pyramid evaluation method (Nenkova and Passonneau, 2004b). The model extracts propositions from the document and the summary, and then uses a RoBERTa encoder fine-tuned on MNLI and augmented with a binary classification layer to determine which propositions are aligned.

We run the pre-trained superPAL model on the SDS DUC 2001 and 2002 document-summary pairs that were not already manually annotated (see §4), consisting of 1911 such pairs (see Table 1), and the pairs of the CNN-DM train split (Nallapati et al., 2016a), consisting of 285073 such pairs (see Table 1). For each pair, we collect only document highlights with an alignment probability of 0.5 or more, similarly to Ernst et al. (2021b). This way, we perform automatically the task that was manually performed in §4.

| P | R | F1 |
|---|---|---|
| 66.17 | 68.35 | 65.24 |

Table 2: Token-wise macro-averaged precision, recall, and F1 scores when comparing the manually and automatically annotated document-summary pairs (dev&test).

### 5.1   Evaluation of Automatic Annotation

Next, we wish to assess the quality of the automatically-generated data, and especially its correlation to the manually annotated dataset. For that, we first use SuperPal to extract potential highlights in the document-summary pairs annotated by our annotators (see §4). Next, for every data point, we compare all its automatically-extracted highlights with their crowd-sourced counterparts.

Table 2 presents the tokenwise[7] macro-averaged precision, recall, and F1 values, with the crowd-sourced highlights as the gold data (the micro-averaged values show similar trends - see §E). These results suggest that our automatically-generated highlights cover a substantial portion of the highlights, with reasonable precision, making them useful for large-scale training. However, these figures also stress the necessity of our manual annotation for the dev and test sets.

## 6   Baseline Models

We experiment with two methods for the controlled text reduction task: a supervised model, whose input is the full document, supplemented with indications of the highlighted spans (§6.1) and another supervised model that receives as input only a concatenation of the highlights, without the surrounding context (§6.2). Both models are trained on our automatically-generated train dataset (§5).

### 6.1   Highlights in Context

Considering the length requirements of our data (see Table 1), we opt for a model designated for long inputs. We employ the Longformer Encoder-Decoder base model (LED$_{base}$; Beltagy et al., 2020), with the standard configurations.[8] The Longformer is an adaption of BART (Lewis et al., 2020) for longer inputs, replacing BART's encoder with a combination of a local and a (optional) global attention mechanism. The local attention, which comes in the form of a sliding window, is mostly used to build contextual representations, by

---

[7]We consider only content words.
[8]For details, see this colab notebook.

enabling each token to attend to its neighbors. Alternatively, a global attention, which is given to a few pre-selected input tokens, enables those tokens to attend to all the tokens in the input (and not only its neighbors), and also allows all input tokens to attend to the global ones. LED has demonstrated state-of-the-art results when evaluated on the arXiv long document summarization dataset (Cohan et al., 2018), making it a suitable choice for our experiments. We denote this model LED$_H$.

## 6.2 Only Highlights

To demonstrate the necessity of the document context, we also train a variant of the LED model where the input consists of a concatenation of the supplied document spans, without the surrounding context.[9] We denote it LED$_{only-H}$. We use the same configurations as in §6.1 while omitting the global attention (given it is not needed in this setting).

## 7 Experimental Setup

**Baseline Models** We use our training dataset (§5) to finetune our two LED variants (§6). We employ the CNN-DM dataset together with our DUC trainset for initial fine-tuning, which is then followed by further finetuning on the DUC trainset alone. We avoid using the CNN-DM dataset in the latter finetuning phase since its quality is notably lower compared to the DUC dataset. Specifically, CNN-DM was generated automatically, in comparison to the expert-written summaries in DUC, and it consists of standalone bullet points, lacking the desired discourse properties and flow of natural text. To avoid overfitting on the CNN-DM dataset, which is much larger than DUC, we experimented with using only fractions of the CNN-DM data. Optimal performance was achieved when using the full CNN-DM data for the initial finetuning of the LED$_H$ model (§6.1), while for the LED$_{only-H}$ model it was best to finetune only on the DUC data, avoiding the CNN-DM data altogether.

In the LED$_{only-H}$, we preprocess our input, extracting the highlights and then using a dot (followed by a space) to separate spans originating in different sentences, and a white space otherwise. To model the highlights in the LED$_H$ setting, we follow Deutsch and Roth (2021) and add to the vocabulary two special tokens, *<highlight_start>* and *<highlight_end>*, which are inserted as vec-

tors into the source documents' embedding layer at the beginning and end of each highlighted span. Also, we combine LED's local attention with its global attention mechanism. As the global attention adds bias to the designated tokens, we mark all *<highlight_start>* and *<highlight_end>* tokens as global tokens. Our motivation stems from the assumption that allowing all the highlight tokens to attend to one another (through the symmetry of the global attention) will encourage the model to fuse the information they are attached to, under the assumption that the highlighted spans are related. Though LED supports inputs with up to 16384 tokens, for our purposes we limit it to 4096 tokens (see Table 1).

We also examined other techniques to represent the highlights (Chen and Bansal, 2018; Xu et al., 2020; Liu et al., 2021), but as they introduced dependencies between their salience detection and generation components, we found them less fitting in our setting.

As a reference point, we compare the abstractive models to an extractive text generated by simply concatenating the highlights, as described previously (i.e., the input to LED$_{only-H}$). This version serves to demonstrate the necessity of our new abstractive task formulation, since without a system that bridges disparate texts, the concatenated spans are often unintelligible.

**No Highlights** In addition to the two baseline models for our text reduction task, we also examine LED in a standard no-highlight summarization setting, where it is finetuned and evaluated on the original document without any highlights. In the absence of highlights, the global attention becomes unnecessary, hence this variant incorporates solely local attention. This no-highlight variant of LED, denoted LED$_{NH}$, matches the classic summarization setting and provides insights into the ability of the model to pick up the highlighting signals. When optimizing the amount of CNN-DM data to use in the initial finetuning phase, as described above for the baseline models, we found it optimal to use $5\%$ of the CNN-DM data.

**Highlights-Summary Mix** To investigate the extent of the highlights' impact, we create a variant of our highlighted test setting: For each document-summary pair, we assign highlighted spans that were extracted from another reference summary available for the same document. We use all the

---

[9]Given this input is short, we also experimented with Pegasus, which showed comparable results on the dev set.

| Concat. | LED$_{\text{only-H}}$ | LED$_{\text{H}}$ |
|---------|-----------------------|------------------|
| 2.76    | 3.12                  | **4.58**         |

Table 3: The (averaged) human ratings of fluency of the summaries generated by our two baseline models and the extractive reference model (Concat.).

document-summary pairs in our test set for this experiment. We then evaluate the finetuned LED$_{\text{H}}$ (see §6.1) on this setting, evaluating the generated summary for each highlighted document with the summary that may have different salient content than the highlights, denoting it LED$_{\text{H-mix}}$.

**Manual Fluency Evaluation**    To test our assumption that simply concatenating the highlights, or excluding the document context, results in less coherent text, we ask crowd-workers to rate the fluency of the generated texts for the suggested baseline models. Our group of crowd-workers consists of reliable workers that have shown a good grasp of different semantic tasks including summarization in past experiments. To evaluate, we randomly choose 100 documents from our test set, each is assigned a single set of highlights corresponding to some summary. We design a simple Amazon Mechanical Turk interface, where we present all three generated summaries of the same input (see §F). Inspired by Fabbri et al. (2021)'s judgment guidelines to crowd-workers, we use a 5-point Likert scale to evaluate the consistency and fluency of the generated summaries and add criteria explaining each score, to reduce ambiguity and enforce more consistent rating (see §F).

## 8    Analysis and Results

First, we present the fluency results to validate the necessity of our task setting. As expected, it arises from Table 3 that the Concat. approach generates highly incoherent summaries, as opposed to the supervised model. This shows that just copying from the highlights directly leads to incoherent text. We also see that removing the context from the input is also detrimental to the model's ability to generate a coherent text (LED$_{\text{H}}$ vs. LED$_{\text{only-H}}$), demonstrating the importance of context (see §G for example generated texts). To obtain further insight into context importance, we manually inspect the crowd-sourced datasets and find that for 74% of the document-summary pairs, context is indeed required to properly connect the disparate highlighted spans.

Next, we proceed to evaluate content preserva-

tion using ROUGE (Lin and Hovy, 2003), a lexical overlap metric (see Table 4). To measure content preservation we apply the metric between the generated text and the highlighted content aimed to be preserved (technically, the highlights are concatenated to apply the ROUGE measure).[10]    As may expected, it arises from Table 4 that passing only the highlights through a supervised model results in the best ROUGE scores (see LED$_{\text{only-H}}$), suggesting that, in the absence of additional content, the LED$_{\text{only-H}}$ model tends to preserve the original lexical content within its input highlights. Yet, as was seen in Table 3, avoiding the context yields unacceptably incoherent text, making this model irrelevant to the task. Adding context to the input (LED$_{\text{H}}$) downgrades the ROUGE score, which may be attributed to either desired or undesired behaviors of the LED$_{\text{H}}$ model. In some cases, the generated text does preserve the highlighted content, but deviates from it lexically in order to generate fluent text, possibly incorporating certain lexical elements from the context while preserving meaning. In other cases, however, the generated text does deviate from the highlighted content by erroneously adding to the output non-highlighted content from the surrounding context. Unfortunately, the ROUGE measure, being based solely on lexical matches, does not distinguish between these two cases. To that end, we add a manual faithfulness analysis in §8.1 (Table 5), which evaluates content preservation more precisely, with respect to both precision (faithfulness) and recall (coverage).

Finally, we observe an approximately 8 points decrease in all ROUGE metrics when removing the highlights (LED$_{\text{NH}}$), indicating that highlights do in fact play a major role in directing the model to focus on specific targeted content. We see a similar trend in LED$_{\text{H-mix}}$, suggesting that each set of highlights steers the model toward the specific content it focuses on. This further confirms the highlights' role in the model's content-related decisions.

In conclusion, to evaluate future progress on the text reduction task, we firstly propose combining manual evaluation of fluency, requiring sufficient fluency to make models acceptable, along with automatic evaluation of content preservation via common measures for this purpose such as ROUGE. While we also inspected less standard automatic evaluation measures, for both fluency

---

[10]It is worth mentioning that we observed similar trends when comparing the generated texts to the original summaries - see §H for further details.

|  | R-1 | R-2 | R-L |
|---|---|---|---|
| LED$_{only-H}$ | **79.37** | **66.71** | **69.74** |
| LED$_H$ | 70.15 | 53.14 | 57.87 |
| LED$_{NH}$ | 49.98 | 28.89 | 36.55 |
| LED$_{H-mix}$ | 67.17 | 49.40 | 55.62 |

Table 4: ROUGE-1, -2 and -L content preservation results, comparing model output to the (concatenated) highlights in the input. We evaluate our baseline models (LED$_H$ and LED$_{only-H}$), along with the alternative compared configurations (LED$_{NH}$ and LED$_{H-mix}$).

|  |  | Faithfulness (P) | Coverage (R) |
|---|---|---|---|
| Full Doc | LED$_{NH}$ | 80.89 | N/A |
|  | LED$_H$ | **85.11** | N/A |
| Highlights | LED$_{NH}$ | 29.94 | 27.33 |
|  | LED$_H$ | **52.48** | **45.68** |

Table 5: Fact-wise faithfulness (P) and coverage (R) scores for LED$_{NH}$ and LED$_H$, once between generated summaries and the full source document and once between the generated summaries and the highlight.

(Mutton et al., 2007) and semantic-oriented content matching (Honovich et al., 2021; Laban et al., 2022), we found them to be not sufficiently reliable for our setting. That said, future progress in the quality of automatic evaluation of summary fluency and content matching would be highly applicable, and desired, for our text reduction task as well, particularly given the known deficiencies of the lexical-matching-based ROUGE measure. Further, reliable crowdsourcing methods for human evaluation of content matching may be considered as well (Shapira et al., 2019), as we illustrate in our limited-scale analysis in the next subsection.

### 8.1 Performance Analysis

To further evaluate the highlights' effect, we manually assess LED$_H$ and LED$_{NH}$ on two levels: (1) faithfulness of the generated text and (2) coverage of the highlighted spans in the system summary.

To determine the amount of system summary spans that are entailed by the source, we compare each summary span to the source. We conducted two manual experiments, one with respect to the full document, and one with respect to the highlighted spans only. To that end, we randomly select 10 unique documents from our test set, with one of their set of highlights. Then, following the notion of Summary Content Unit (SCU) in the Pyramid method for summarization evaluation (Nenkova and Passonneau, 2004a), we extract such units from both the summary and the source text us-

ing the Summary Evaluation Environment (SEE) described in that paper. Then, for each summary unit, we manually search for a matched document unit conveying the same information, to determine whether the summary unit is mentioned in the document (TP) or not (FP). Lastly, we calculate the micro-precision, which represents the faithfulness of both models' outputs. Table 5 shows an almost 5% improvement in faithfulness to the source document when adding highlights. This implies that the highlights not only steer the model towards specific content but also help it keep focused on the source. Interestingly, we find that one-third of the faithfulness errors (FP) stem from disparate highlights that were incorrectly combined, which is typical for summarization hallucinations.

We also evaluate the highlights' coverage by the summaries. For that, we calculate the number of False Negative (FN) summary facts, compared to the facts in the highlights, and compute the micro-recall value, representing the summaries' coverage of the highlights. Table 5 shows a clear advantage to including highlights, with almost twice as big faithfulness (P) and coverage (R) of the highlighted facts. With that said, we note that the highlight-related faithfulness is still only a little over 50%, indicating that the model included non-highlighted facts, which further exhibits the challenge to devise models that better focus only on the highlights.

## 9 Conclusion

In this paper, we promote the separation of the summarization task into the salience-detection and text-generation steps. We foresee applications where salient phrases will be highlighted by an avid reader, or selected by a model specialized in some domain, while a more general-purpose model would reformulate the disparate pieces into a coherent text. Thus, we argue that *Controlled Text Reduction*, the second step of summarization, is an interesting and useful research goal in its own right. To bolster the task, we release a high-quality evaluation dataset and a heuristically-generated training data, evaluation protocol, and the first baseline model. The latter clearly shows how the generated summary text benefits from the added salient span signals. Future works may expand this to include multi-document settings in order to accommodate the task to a broader range of applications, and also focus on designing better evaluation metrics for the task.

## 10 Limitations

In this work, we construct the first-of-its-kind Controlled Text Reduction dataset, by aligning text spans in existing summaries to their correlated document spans. This poses a limitation on the highlights chosen, whereas in a more general setting users are free to highlight whatever they find interesting. On the contrary, in our setting, the highlights contain general salient information (that was extracted by the former human summarizer) rather than specific details.

Also, our train dataset was derived automatically using the SuperPAL model. Hence, it is likely that some of the highlights in the training dataset are not perfectly aligned with the summary.

Finally, the dataset is based on a news corpus, which might limit its applicability to other applications that have different structures, such as medical or legal documents, or meeting summaries.

## Acknowledgements

## References

Nader Akoury, Kalpesh Krishna, and Mohit Iyyer. 2019. Syntactically supervised transformers for faster neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1269–1281, Florence, Italy. Association for Computational Linguistics.

Kristjan Arumae, Parminder Bhatia, and Fei Liu. 2019. Towards annotating and creating summary highlights at sub-sentence level. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 64–69, Hong Kong, China. Association for Computational Linguistics.

Siddhartha Banerjee, Prasenjit Mitra, and Kazunari Sugiyama. 2016. Multi-document abstractive summarization using ilp based multi-sentence compression.

Regina Barzilay and Kathleen R. McKeown. 2005. Sentence fusion for multidocument news summarization. *Computational Linguistics*, 31(3):297–328.

Tal Baumel, Raphael Cohen, and Michael Elhadad. 2016. Topic concentration in query focused summarization datasets. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Tal Baumel, Matan Eyal, and Michael Elhadad. 2018. Query focused abstractive summarization: Incorporating query relevance, multi-document coverage, and summary length constraints into seq2seq models. *arXiv preprint arXiv:1801.07704*.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer.

Emanuele Bugliarello and Naoaki Okazaki. 2020. Enhancing machine translation with dependency-aware self-attention. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1618–1627, Online. Association for Computational Linguistics.

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336.

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–686, Melbourne, Australia. Association for Computational Linguistics.

Sangwoo Cho, Logan Lebanoff, Hassan Foroosh, and Fei Liu. 2019. Improving the similarity measure of determinantal point processes for extractive multi-document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1027–1038, Florence, Italy. Association for Computational Linguistics.

Sangwoo Cho, Kaiqiang Song, Chen Li, Dong Yu, Hassan Foroosh, and Fei Liu. 2020. Better highlighting: Creating sub-sentence summary highlights. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6282–6300, Online. Association for Computational Linguistics.

Leshem Choshen and Omri Abend. 2021. Transition based graph decoder for neural machine translation. *arXiv preprint arXiv:2101.12640*.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In

*Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Terry Copeck, Diana Inkpen, Anna Kazantseva, Alistair Kennedy, Darren Kipp, Vivi Nastase, and Stan Szpakowicz. 2006. Leveraging duc. In *proceedings of DUC*.

Terry Copeck, Diana Inkpen, Anna Kazantseva, Alistair Kennedy, Darren Kipp, and Stan Szpakowicz. 2007. Catch what you can. *Proceedings of DUC 2007*.

Terry Copeck, Anna Kazantseva, Alistair Kennedy, Alex Kunadze, Diana Inkpen, and Stan Szpakowicz. 2008. Update summary update. In *TAC*.

Terry Copeck and Stan Szpakowicz. 2005. Leveraging pyramids.

Daniel Deutsch and Dan Roth. 2021. Question-based salient span selection for more controllable text summarization.

Ori Ernst, Avi Caciularu, Ori Shapira, Ramakanth Pasunuru, Mohit Bansal, Jacob Goldberger, and Ido Dagan. 2021a. Proposition-level clustering for multi-document summarization.

Ori Ernst, Ori Shapira, Ramakanth Pasunuru, Michael Lepioshkin, Jacob Goldberger, Mohit Bansal, and Ido Dagan. 2021b. Summary-source proposition-level alignment: Task, datasets and supervised baseline. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 310–322, Online. Association for Computational Linguistics.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. 2018a. Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4098–4109, Brussels, Belgium. Association for Computational Linguistics.

Sebastian Gehrmann, Yuntian Deng, and Alexander M Rush. 2018b. Bottom-up abstractive summarization. *arXiv preprint arXiv:1808.10792*.

Mor Geva, Eric Malmi, Idan Szpektor, and Jonathan Berant. 2019. DiscoFuse: A large-scale dataset for discourse-based sentence fusion. In *Proceedings of*

the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 3443–3455, Minneapolis, Minnesota. Association for Computational Linguistics.

Demian Gholipour Ghalandari, Chris Hokamp, Nghia The Pham, John Glover, and Georgiana Ifrim. 2020. A large-scale multi-document summarization dataset from the wikipedia current events portal. *arXiv preprint arXiv:2005.10070*.

Aria Haghighi and Lucy Vanderwende. 2009. Exploring content models for multi-document summarization. In *Proceedings of human language technologies: The 2009 annual conference of the North American Chapter of the Association for Computational Linguistics*, pages 362–370.

Junxian He, Wojciech Kryściński, Bryan McCann, Nazneen Rajani, and Caiming Xiong. 2020. Ctrlsum: Towards generic controllable text summarization.

Eran Hirsch, Alon Eirew, Ori Shapira, Avi Caciularu, Arie Cattan, Ori Ernst, Ramakanth Pasunuru, Hadar Ronen, Mohit Bansal, and Ido Dagan. 2021. iFacetSum: Coreference-based interactive faceted summarization for multi-document exploration. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 283–297, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. $q^2$: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7856–7870, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation.

Byeongchang Kim, Hyunwoo Kim, and Gunhee Kim. 2019. Abstractive summarization of Reddit posts with multi-level memory networks. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2519–2531, Minneapolis, Minnesota. Association for Computational Linguistics.

Sayali Kulkarni, Sheide Chammas, Wan Zhu, Fei Sha, and Eugene Ie. 2020. Aquamuse: Automatically

generating datasets for query-based multi-document summarization. *arXiv preprint arXiv:2010.12694*.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Logan Lebanoff, Franck Dernoncourt, Doo Soon Kim, Walter Chang, and Fei Liu. 2020a. A cascade approach to neural abstractive summarization with content selection and fusion. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 529–535, Suzhou, China. Association for Computational Linguistics.

Logan Lebanoff, Franck Dernoncourt, Doo Soon Kim, Lidan Wang, Walter Chang, and Fei Liu. 2020b. Learning to fuse sentences with transformers for summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4136–4142, Online. Association for Computational Linguistics.

Logan Lebanoff, John Muchovej, Franck Dernoncourt, Doo Soon Kim, Seokhwan Kim, Walter Chang, and Fei Liu. 2019. Analyzing sentence fusion in abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 104–110, Hong Kong, China. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin and Eduard H. Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *NAACL*.

Shuaiqi Liu, Jiannong Cao, Ruosong Yang, and Zhiyuan Wen. 2021. Highlight-transformer: Leveraging key phrase aware attention to improve abstractive multi-document summarization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5021–5027, Online. Association for Computational Linguistics.

Andrew Mutton, Mark Dras, Stephen Wan, and Robert Dale. 2007. GLEU: Automatic evaluation of sentence-level fluency. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 344–351, Prague, Czech Republic. Association for Computational Linguistics.

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

Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. 2016b. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*.

Ramesh Nallapati, Bowen Zhou, and Mingbo Ma. 2016c. Classify or select: Neural architectures for extractive document summarization. *arXiv preprint arXiv:1611.04244*.

Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simões, Vitaly Nikolaev, and Ryan McDonald. 2021. Planning with learned entity prompts for abstractive summarization. *Transactions of the Association for Computational Linguistics*, 9:1475–1492.

Ani Nenkova and Rebecca Passonneau. 2004a. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 145–152, Boston, Massachusetts, USA. Association for Computational Linguistics.

Ani Nenkova and Rebecca J Passonneau. 2004b. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the human language technology conference of the north american chapter of the association for computational linguistics: Hlt-naacl 2004*, pages 145–152.

Tatsuro Oya, Yashar Mehdad, Giuseppe Carenini, and Raymond Ng. 2014. A template-based abstractive meeting summarization: Leveraging summary and source text relationships. In *Proceedings of the 8th International Natural Language Generation Conference (INLG)*, pages 45–53, Philadelphia, Pennsylvania, U.S.A. Association for Computational Linguistics.

Onkar Pandit and Yufang Hou. 2021. Probing for bridging inference in transformer language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4153–4163, Online. Association for Computational Linguistics.

Ramakanth Pasunuru, Asli Celikyilmaz, Michel Galley, Chenyan Xiong, Yizhe Zhang, Mohit Bansal, and Jianfeng Gao. 2021. Data augmentation for abstractive query-focused multi-document summarization. In *AAAI*.

Romain Paulus, Caiming Xiong, and Richard Socher. 2017. A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*.

Paul Roit, Ayal Klein, Daniela Stepanov, Jonathan Mamou, Julian Michael, Gabriel Stanovsky, Luke Zettlemoyer, and Ido Dagan. 2020. Controlled crowdsourcing for high-quality QA-SRL annotation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7008–7013, Online. Association for Computational Linguistics.

Ori Shapira, David Gabay, Yang Gao, Hadar Ronen, Ramakanth Pasunuru, Mohit Bansal, Yael Amsterdamer, and Ido Dagan. 2019. Crowdsourcing lightweight pyramids for manual summary evaluation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 682–687, Minneapolis, Minnesota. Association for Computational Linguistics.

Ori Shapira, Ramakanth Pasunuru, Hadar Ronen, Mohit Bansal, Yael Amsterdamer, and Ido Dagan. 2021. Extending multi-document summarization evaluation to the interactive setting. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 657–677, Online. Association for Computational Linguistics.

Aviv Slobodkin, Leshem Choshen, and Omri Abend. 2021. Mediators in determining what processing BERT performs first. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 86–93, Online. Association for Computational Linguistics.

Aviv Slobodkin, Leshem Choshen, and Omri Abend. 2022. Semantics-aware attention improves neural machine translation. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 28–43, Seattle, Washington. Association for Computational Linguistics.

Gabriel Stanovsky, Julian Michael, Luke Zettlemoyer, and Ido Dagan. 2018. Supervised open information extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 885–895, New Orleans, Louisiana. Association for Computational Linguistics.

Dhanasekar Sundararaman, Vivek Subramanian, Guoyin Wang, Shijing Si, Dinghan Shen, Dong Wang, and Lawrence Carin. 2019. Syntax-infused transformer and bert models for machine translation and natural language understanding.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019a. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. 2019b. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Gregory César Valderrama Vilca and Marco Antonio Sobrevilla Cabezudo. 2017. A study of abstractive summarization using semantic representations and discourse level information. In *Text, Speech, and Dialogue*, pages 482–490, Cham. Springer International Publishing.

Yang Wei and Yang Zhizhuo. 2017. Query based summarization using topic background knowledge. In *2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, pages 2569–2572. IEEE.

Liqiang Xiao, Hao He, and Yaohui Jin. 2022. Fusionsum: Abstractive summarization with sentence fusion and cooperative reinforcement learning. *Knowledge-Based Systems*, 243:108483.

Song Xu, Haoran Li, Peng Yuan, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020. Self-attention guided copy mechanism for abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1355–1362, Online. Association for Computational Linguistics.

Yumo Xu and Mirella Lapata. 2020. Coarse-to-fine query focused multi-document summarization. In *Proceedings of the 2020 Conference on empirical methods in natural language processing (EMNLP)*, pages 3632–3645.

Yumo Xu and Mirella Lapata. 2021. Text summarization with latent queries. *arXiv preprint arXiv:2106.00104*.

Chao Zhao, Marilyn Walker, and Snigdha Chaturvedi. 2020. Bridging the structural gap between encoding and decoding for data-to-text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2481–2491, Online. Association for Computational Linguistics.

## A Preprocessing

In preprocessing, we begin by removing meaningless characters from the input. Then, we use spaCy (Honnibal and Montani, 2017) to parse the input and the reference summaries to get their token segmentation, sentence separation, and lemmatization. Next, we construct a matrix $M_{ij}$ for each document-summary pair:

$$M_{ij} = \begin{cases} 1, & Similarity_{Lemma}(t_i^s, t_j^d) \geq 0.86 \\ 0, & \text{otherwise} \end{cases}$$ (1)

where $t_i^s$ and $t_j^d$ are **summary** token $i$ and **document** token $j$, respectively, and the $Similarity_{Lemma}(t_i^s, t_j^d)$ is computed using the SequenceMatcher[11] module on $t_i^s$'s and $t_j^d$'s lemmas.

In addition, given that most of our dataset was not segmented into paragraphs, we devise a naive algorithm to divide the source documents of each data point in the dev and test datasets, in order to make them more presentable for our annotators and easier to read through. For that, we first apply the neuralcoref model[12] on the documents to get co-reference clusters, which we used together with the spaCy sentence segmentation to determine when paragraph-breaks should occur.

## B Annotation Full Guidelines

In this section, we provide the full annotation guidelines, presented to our workers.

### B.1 Summary-related Guidelines

As mentioned in 4.2, we provide guidelines for the annotator to segment the summary sentence into the facts that it is composed of. We target facts encoded in different grammatical structures, but to present them to the annotator in a simplified manner we show the following three variants:

• SIDE-BY-SIDE: Two events are realized adjacently without sharing participants (e.g., *"He worked"*, comprising of two independent events - *"He worked while I slept"* and *"I slept"*).

• SHARED ELEMENTS: Two events that share some phrases (e.g., *"He worked while smiling"*, which comprises of two events sharing a subject - *"He worked"* and *"He smiled"*).

• NO EXPLICIT VERB: An event is expressed without an explicit verb (e.g., *"John Doe, my good friend, has arrived"*, whose first fact, *"John Doe (is) my good friend"*, lacks an implicit verb).

## C Document-related Guidelines

In this section, we present a more in-depth overview of the document-related guidelines presented to our annotators during their training.[13]:

- Paraphrasing: We instruct our workers to not solely rely on phrases with shared words, as often the most suiting document phrase is a paraphrasing of its summary counterpart (for example, in Figure 2, "a well-qualified panel of judges" is a paraphrasing of its document counterpart).

- Consecutiveness: We guide our workers to avoid highlighting unnecessary details, i.e., that did not appear in the summary span, and keep the highlights inconsecutive if necessary; (e.g., in Figure 2, the nature of the committee's members was excluded from the highlight, to adjust to the summary span, resulting in a non-consecutive highlight).

- Missing Details: When the corresponding document phrase is missing some details, the annotators are instructed to highlight some other mention of the absent information. For example, in Figure 1, the equivalent document span of the summary fact *"The Booker Prize, which was first awarded in 1969"* is "The prize was first awarded in 1969". But, as the prize's "identity" is absent from this span, some mention of it should be highlighted as well (e.g., at the beginning of the document).

- hallucination: For the rare instances where the reference summary has hallucinations, we instruct our workers to leave these details unhighlighted in the summary.

- Context: We guide our workers to verify that the document highlights are used in the same context as the summary spans. For example, if in Figure 1 there was a mention of another prize that was awarded in 1969, highlighting it would be erroneous.

## D IAA disagreement Examples

Figure 4 exemplifies two disagreements between our annotators, which demonstrate the two main causes for disagreement (§4.4). In Figure 4a, we can see that one of the annotators highlighted an extra mention of the necessity to discuss business

---

[11]https://github.com/python/cpython/blob/main/Lib/difflib.py
[12]https://github.com/huggingface/neuralcoref

[13]The full guidelines presented during training upon publication.

5712

| P | R | F1 |
|---|---|---|
| 63.85 | 67.22 | 65.49 |

Table 6: Tokenwise micro-averaged precision, recall, and F1 scores when comparing the manually annotated document-summary pairs with the automatically-annotated pairs.

| | R-1 | R-2 | R-L |
|---|---|---|---|
| Concat. | **71.53** | **47.11** | **52.53** |
| $LED_{only-H}$ | 65.49 | 40.33 | 45.83 |
| $LED_H$ | 59.48 | 33.30 | 40.39 |
| $LED_{NH}$ | 45.94 | 19.68 | 28.86 |
| $LED_{H-mix}$ | 46.86 | 20.26 | 29.60 |

Table 7: ROUGE-1, -2 and -L content preservation results, comparing model output to the gold summaries. We evaluate our baseline models ($LED_H$ and $LED_{only-H}$), along with the alternative compared configurations ($LED_{NH}$ and $LED_{H-mix}$).

(dashed blue), which is allowed in our setup. In Figure 4b, we can see that one of the annotators included "a euphemism for" in the highlight (dashed blue), which has no effect on the overall meaning of the highlight.

## E  Train Data Micro-Averaged Evaluation

Table 6 shows the **micro**-averaged precision, recall, and F1 scores of the comparisons discussed in subsection 5.1.

## F  Fluency Human Evaluation API

Figure 5 present an example of our API designated for the human evaluation of the generated summaries' fluency and coherence.

## G  Generation Examples

Fig. 6 shows two examples of a highlighted source document and the text generated by the Concat. approach (*Naive concatentaion*) and our two baseline models.

## H  ROUGE results When Compared to the Gold Summaries

Table 7 features the ROUGE results of all our models (and also the Concat. extractive approach) when compared to the gold summaries.

---

**Source Document Paragraph**

"Even if it weren't tax deductible, it'd be worth it," says Tom Foley, chairman of Bibb Co., a Macon, Ga., maker of home furnishings and textile products that use the National Football League team logos. Bibb is flying about 20 people -- mostly customers -- to the game. "We get more earnings from the trip than the expense," says Mr. Foley.

Some companies say they are making adjustments or at least educating their executives about the new rules. A spokesman for Adolph Coors &amp; Co, the Golden, Colo., brewery, said its tax advisers instructed executives that "this can't be just a fun trip -- we have to discuss business." Air fare and lodging are still 100% deductible, as long as there is a planned business meeting in addition to the game. Meals are now 80% deductible as long as business is discussed. Under the old law, meals were fully deductible and all a company had to do was show a business relationship with the guest.

**Summary**

Businesses show few signs of cutting back on their activities despite the new tax law's tighter restrictions on what can be claimed as a business-entertainment expense.
One big change for tickets to sporting events like the Super Bowl: only 80% of the face value of such tickets can be deducted and there must be a "substantial and bona fide business discussion" before, during, or after the game.
**As long as business is discussed, airfare and lodging are still 100% deductible and meals are 80% deductible.**

---

(a)

---

**Source Document Paragraph**

A controversial videotape being shown among activists nationwide shows Los Angeles police officers intentionally hurting the nonviolent demonstrators they are arresting. They press fingers under their noses. They dig their knuckles into protesters' necks, and torque martial arts weapons around their wrists. At one point, two officers twist a woman's arm till she rises from the ground, her face wrenched in agony. In another scene, a young man winces as officers lead him along. His arm, contorted behind his back, finally snaps. The law enforcement name for these techniques is "pain-compliance." Police departments nationwide say it's a tried and true way to make uncooperative protesters cooperate. But opponents call the term a euphemism for torture. Demonstrators have alleged police brutality at least since Freedom Riders launched their sit-down strikes in Alabama almost 30 years ago.

**Summary**

A videotape being shown nationwide catches Los Angeles police officers intentionally hurting nonviolent anti-abortion demonstrators.
They press fingers under their noses, dig their knuckles into demonstrators' necks, and torque martial arts weapons around their wrists.
The police call these techniques "pain-compliance."
**Opponents call them torture.**

---

(b)

Figure 4: Two examples of disagreement between annotators. For each example, the bottom part is the summary (with the summary sentence over which there was disagreement in bold and underlined) and the top part is a single paragraph from the source document with both the annotators' highlights (marked with a red solid line and a blue dashed line to indicate each highlight).

**Instructions**

In this task, you will evaluate the quality of 3 summaries (each in a separate tab).
To correctly solve this task, please follow these steps:

1. Read each summary.
2. Rate each summary on a scale from **1** (worst) to **5** (best) by its fluency\coherency.

**Definition of Fluency**

This rating measures the quality of the text - are the sentences well written and grammatically correct, do they fit together and sound natural.
Consider how legible, grammatical and coherent the summaries are.
The scale should be:

- **1** - The summary is incoherent, contains multiple grammatical errors and doesn't sound natural.
- **2** - Only small parts of the summary are coherent, containing many grammatical errors and its sentences fit together poorly.
- **3** - The summary is somewhat coherent, but it either contains several grammatical errors or its sentences don't fit very well.
- **4** - The summary is mostly coherent, contains little to no grammatical errors and sounds natural enough.
- **5** - The summary is very coherent, contains no grammatical errors and sounds natural.

| Summary A | Summary B | Summary C |

| Instructions | Shortcuts | How fluent is this summary? |

Debi Thomas' dream of Olympic gold turned into disappointment Saturday as East Germany's Katarina Witt won her second straight Olympic championship and Canadian Elizabeth Manley took home the silver. Thomas won the bronze despite three faulty landings. Thomas, of San Jose, Calif., the first black to win a U.S. figure skating crown and the 1986 world champion, skated poorly Saturday after doing well earlier in the Games. Thomas was for the United States. Brian Boitano won the men's crown a bronze in pairs went to. In addition figure skating the U.S. team had three speed-skating medals: one each gold, silver and bronze.

Select an option

| not fluent | 1 |
| mostly not fluent | 2 |
| partially fluent | 3 |
| mostly fluent | 4 |
| fluent | 5 |

Figure 5: Example of the data collection API used by crowd-source workers.

**Input:** TEAMS OF US and inter-national experts are being sent to southern Africa to assess the impact on food supplies of what in some areas is the worst drought of the century. Officials said millions of people will be affected, AP reports from Washington. Among the hardest hit of the 10 drought-stricken countries are Zimbabwe and South Africa, traditional food exporters which this year will have to import substantial quantities of grain. As the drought persists, estimates of the grain harvest throughout the region have been falling precipitously. The deteriorating situation in southern Africa adds a new dimension to the continent's overall food crisis. The north-east is the most deprived area in Africa, encompassing Sudan, Ethiopia, Somalia and Djibouti. The lives of 15m people are thought to be at risk. Officials in Zimbabwe said last week 60,000 cattle died from starvation during February alone and thousands more will die unless heavy rains come soon. The Rome-based Food and Agriculture Organisation warned late last month that the drought could lead to widespread famine. About 98m people live in the affected regions in southern Africa. 'Considerable donor assistance will be needed to avert a major humanitarian crisis in the region,' the US State Department said last week.

**Naive concatenation :** inter-national experts are being sent to southern Africa to assess the impact on food supplies of what in some areas is the worst drought of the century. Among the hardest hit of the 10 drought-stricken countries are Zimbabwe and South Africa traditional food exporters which this year will have to import substantial quantities of grain. the drought persists. The north-east Africa, encompassing Sudan, Ethiopia, Somalia and Djibouti. The lives of 15m people are thought to be at risk. Considerable donor assistance will be needed to avert a major humanitarian crisis in the region,' the US State Department said.

**LED$_{only-H}$ :** US inter-national experts are being sent to southern Africa to assess the impact on food supplies of what in some areas is the worst drought of the century. The hardest hit are Zimbabwe and South Africa traditional food exporters, which this year will have to import substantial quantities of grain. In the north-east Africa, encompassing Sudan, Ethiopia, Somalia and Djibouti, 15 million people are at risk. 'Considerable donor assistance will be needed to avert a major humanitarian crisis in the region,' the US State Department said.

**LED$_H$ :** A team of US and international experts is being sent to southern Africa to assess the impact of what in some areas is the worst drought of the century. Among the 10 drought-stricken countries are Zimbabwe and South Africa, traditional food exporters which this year will have to import substantial quantities of grain. As the drought persists, estimates of the grain harvest throughout the region have been falling precipitously. The north-east is the most deprived area in Africa, encompassing Sudan, Ethiopia, Somalia and Djibouti. 15 million people are thought to be at risk. Considerable donor assistance will be needed to avert a major humanitarian crisis in the region.

· · · · · · · · · · · · · · · · · · · · · · · · · · · · · · · · · · · · · · · · · · · · · · · · · · · · · · · · · ·

**Input:** The shortlist of six for the Pounds 20,000 Booker Prize for fiction, announced yesterday, immediately prompted the question 'Who?' from many in the publishing industry. According to one insider, some on the list 'are B-team writers at best'. The six include Alan Hollinghurst's The Folding Star (published by Chatto and Windus), a melancholy study of homosexual obsession which was tipped as a likely candidate from the initial 'long list' of 15, The Reef (Granta) by young Sri Lankan writer Romesh Gunesekera and How Late It Was, How Late (Secker and Warburg) by gritty Glasgow realist James Kelman, which was almost universally well-reviewed. As for the other three - Knowledge of Angels (Green Bay) a philosophical fable by children's author Jill Paton Walsh, Paradise (Hamish Hamilton) by Zanzibar-born writer Abdulrazak Gurnah and Beside the Ocean of Time (John Murray) by 72-year-old Orcadian poet George Mackay Brown - 'frankly, they don't make the grade'. The shortlist for the Booker, the UK's most hyped literary prize and one of the most lucrative, is all the more surprising in a bumper year for new fiction fulfilling the criteria - English language and non-American - for consideration for the award. Margaret Atwood's The Robber Bride seems an astonishing omission, as do new novels by Peter Ackroyd, Peter Carey, Candia McWilliam, William Trevor and Jim Crace. But if the shortlist of the final six candidates for the prize may be disappointing, the traditional controversy surrounding the award is as rife as ever. One unsurprising omission from the final selection was When The World Was Steady, a first novel by Claire Messud, which was on the 'long list' before it was pointed out that the author was the wife of James Wood, chief literary reviewer of The Guardian newspaper and a Booker judge. Professor John Bayley, chairman of the Booker panel and husband of former Booker Prize winner Dame Iris Murdoch, expressed surprise at Mr Wood's failure to reveal his relationship with Ms Messud. But it is the final list, rather than the controversy, that discredits the award according to some critics. 'This list,' said one man of letters,'must have dealt a final death blow to the Booker.' The winner, selected from an original 130, will be announced on October 11."

**Naive concatenation :** The shortlist of six for Booker Prize prompted the question 'Who. some on the list 'are B-team writers at best. surprising in a year for new fiction fulfilling the criteria English language and non-American. The winner will be announced on October.

**LED$_{only-H}$ :** The shortlist of six finalists for the Booker Prize has prompted the question, ""Who is the winner of the Booker?"" according to a Wall Street Journal editorial. The winner will be announced on October 11. Some of the six are B-team writers at best.
Others are non-fiction writers at worst. This is the first Booker Prize in a year for new fiction fulfilling the criteria English language and non-American. A shortlist has been announced. It is expected that the winner will not be announced until the end of October.

**LED$_H$ :** The shortlist of six candidates for the Pounds 20,000 Booker Prize for fiction, announced yesterday, immediately provoked the question ""Who?"" from many in the publishing industry. Some on the shortlist are B-team writers at best. Alan Hollinghurst's The Folding Star (published by Chatto and Windus), a melancholy study of homosexual obsession which was tipped as a likely candidate from the initial ""long list"" of 15, and How Late It Was, How Late (Secker and Warburg) by gritty Glasgow realist James Kelman, which was almost universally well-reviewed. The winner will be announced on October 11.

Figure 6: Example predictions from the various baseline models.