# Multi-Document Summarization with Centroid-Based Pretraining

**Ratish Puduppully**[1,2*] and **Parag Jain**[4] and **Nancy F. Chen**[1,2,3]and **Mark Steedman**[4]
[1]Institute for Infocomm Research (I$^2$R), A*STAR, Singapore
[2]CNRS@CREATE, Singapore
[3]Centre for Frontier AI Research (CFAR), A*STAR
[4]School of Informatics, University of Edinburgh
puduppully@i2r.a-star.edu.sg parag.jain@ed.ac.uk
nfychen@i2r.a-star.edu.sg steedman@inf.ed.ac.uk

## Abstract

In Multi-Document Summarization (MDS), the input can be modeled as a set of documents, and the output is its summary. In this paper, we focus on pretraining objectives for MDS. Specifically, we introduce a novel pretraining objective, which involves selecting the ROUGE-based centroid of each document cluster as a proxy for its summary. Our objective thus does not require human written summaries and can be utilized for pretraining on a dataset consisting solely of document sets. Through zero-shot, few-shot, and fully supervised experiments on multiple MDS datasets, we show that our model *Centrum* is better or comparable to a state-of-the-art model. We make the pretrained and fine-tuned models freely available to the research community[1].

## 1 Introduction

In Multi-Document Summarization (MDS), the input is a set of documents, and the output is a summary that describes important information in a coherent and non-redundant manner (McKeown and Radev, 1995; Radev and McKeown, 1998). In recent years, there have been significant improvements in MDS due to the availability of MDS datasets (Fabbri et al., 2019; Gholipour Ghalandari et al., 2020; Liu* et al., 2018) and advances in pretraining approaches (Lewis et al., 2020; Raffel et al., 2020; Zhang et al., 2020).

In particular, Xiao et al. (2022) introduced a pretraining approach called PRIMERA (Pyramid-based Masked Sentence Pretraining) adapted for MDS. To create synthetic summaries, they used the Pyramid scheme (Nenkova and Passonneau, 2004), incorporating a process of identifying and ranking entities, followed by grouping sentences containing these entities in the input documents. The sentences with the highest overlap with other documents (measured using ROUGE) in each group were masked in the input and integrated into the output, forming a synthetic summary. Xiao et al. (2022) show that an encoder-decoder model trained on such a corpus attains strong zero-shot, few-shot, and fully supervised results on multiple datasets.

However, these synthetic summaries may lack coherence as the sentences are derived from various positions within the input documents. Furthermore, there is potential for redundancy, as sentences encapsulating similar information could be selected for inclusion in the summary.

In this paper, we propose Centrum, a pretraining objective for MDS, which is conceptually simple and overcomes these problems. The key intuition is that among a set of documents in a news cluster, the document which shares the most content with the other documents in the cluster can serve as a proxy for the summary of the document set. Such a cluster centroid is inherently coherent as it is a human-written document. Furthermore, because it isn't artificially assembled, it avoids content repetition.

In this paper, we pretrain Centrum on NewSHead (Gu et al., 2020) corpus and perform zero-shot, few-shot and fully-supervised experiments on various MDS datasets. We show that Centrum performs favorably compared to PRIMERA, especially in the zero-shot and few-shot settings, where there are none or very few training examples available.

## 2 Centroid-based Selection of Document as Summary

**Background on PRIMERA** Xiao et al. (2022) leveraged the NewSHead corpus (Gu et al., 2020), a compilation of 369,940 news clusters, for pretraining. Using the Pyramid scheme (Nenkova and Passonneau, 2004), they created synthetic summaries through a multi-step procedure. They gathered the entity mentions in the input documents and rank the entities by the count of documents in which

*Part of the work was done when the author was at the University of Edinburgh

[1]https://github.com/ratishsp/centrum



*Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*
*Volume 2: Short Papers*, pages 128–138

an entity is mentioned. Next, they divide the sentences from the documents into distinct groups, such that the sentences containing an entity belong to the same group. They then extracted the sentence with the highest overlap (as quantified by ROUGE (Lin, 2004)) with other documents from each group. This sentence was replaced with a mask token in the input, and copied to the output document. The idea here was to leverage information from other documents to reconstruct the masked sentence. The sentences thus obtained were concatenated to form a synthetic summary.

Xiao et al. (2022) applied the method to the NewSHead corpus (Gu et al., 2020) containing news articles clustered by topic. To accommodate long document lengths, they use Longformer Encoder-Decoder (LED) architecture (Beltagy et al., 2020). LED supports sparse global attention along with dense local attention on the input. PRIMERA employs global attention on specialized tokens (<doc-sep>), which act as separators between the documents within the input cluster. The pretrained PRIMERA model was then used for zero-shot evaluation, few-shot or full finetuning across multiple MDS datasets.

**Problems with PRIMERA pretraining** PRIMERA's reference summaries consist of sentences extracted from varying positions within different documents in a cluster. This method can yield incoherent summaries, as it can be unclear which entities the sentences refer to. We illustrate this with an example of a synthetic summary created using PRIMERA in Table 1. The first sentence about Lady Gaga originates from the first document, while the second sentence mentioning Donald Trump and Elton John comes from the second document. The lack of entity mentions within these sentences disrupts the overall coherence of the summary. We also note occurrences of redundant information in the synthetic summary. Our hypothesis is that pretraining using such noisy synthetic summaries could negatively impact model performance, particularly in zero-shot or few-shot experiments.

**Our Model** We propose an alternate pretraining objective for MDS called as Centrum. We hypothesize that a document exhibiting the highest similarity with the rest of the documents in a cluster could serve as a proxy for its summary. This method inherently filters out documents that bear only a distant relation to other documents in the cluster. Furthermore, it addresses potential noise present in automatically created multi-document cluster datasets (Gu et al., 2020), for example, a document falsely associated with a cluster. The Centrum pretraining objective excludes such noise, as a mismatched document would not be chosen as the cluster centroid. Among the documents, a document may have more relevant content than others. The Centrum objective will select the more relevant document as the summary.

Drawing inspiration from Gu et al. (2020), we designate a document as the summary if it maximizes the semantic match with other documents in the cluster. Specifically, from each document set $\mathcal{D}$ in an instance, the best candidate summary $\hat{y}$ is selected as:

$$\hat{y} = \arg\max_{x \in \mathcal{D}} 1/|\mathcal{D}| \sum_{x' \in D \setminus \{x\}} f(x, x') \quad (1)$$

where $f(x, x')$ represents the semantic match of summary $x$ with document $x'$. A model can be trained to learn this function $f$. In our approach, we employ the average of ROUGE1, ROUGE2, and ROUGEL as this function. Our pretraining corpus is constructed by treating $D \setminus \{\hat{y}\}$ as the

| |
|---|
| She's a fantastic person, solid as a rock and I'm very proud of her success because I really believe I had at least something to do with it." It was unclear exactly what type of records he was referring to — the attendance of 6,500 fell far short of many Elton John concerts. . . . (4 sent) Donald Trump made his way to Great Falls, Montana, on Thursday (July 5), primarily to slam Democratic Sen. Jon Tester and accuse him of failing to live up to his promises in Washington. "I've broken more Elton John [attendance] records, and I don't have a musical instrument," he boasted. This is my only musical instrument-the mouth-and hopefully the brain is attached to the mouth. During a rally in Great Falls, Montana, where President Trump derided the #MeToo movement and attacked individual Democratic lawmakers, the president once again bragged about the size of his supporter turnout. "I've broken more Elton John [attendance] records, and I don't have a musical instrument," Trump said according to Yahoo News. "I don't have a guitar, or an organ. . . . This is my only musical instrument - the mouth - and hopefully the brain is attached to the mouth. The brain is so much more important." . . . |

Table 1: Example of a synthetic reference summary in PRIMERA (Xiao et al., 2022). We see that the reference summaries in PRIMERA can contain instances of incoherence and repetition. In this summary, the first sentence is about Lady Gaga, and the second is about Donald Trump and Elton John. The subjects of the first two sentences (highlighted in orange) are unclear due to the lack of named entities. Additionally, the sentences in brown and red contain repetitive information.

input and $\hat{y}$ as the output.

Vogler et al. (2022) recently employed a comparable strategy for unsupervised MDS. However, our approach differs from theirs by applying this strategy for MDS task-specific pretraining. Moreover, following Xiao et al. (2022), we employ the LED architecture for handling long document context in the input.

## 3 Experimental Setup

**Model** We utilize the Transformers (Wolf et al., 2020) library to conduct our experiments. Similar to Xiao et al. (2022), we adopt the large configuration of LED, comprising 459M parameters. We finetune the LED model on the NewSHead corpus (Gu et al., 2020) with our Centrum pretraining objective. Documents within a cluster are concatenated into a single text sequence, with <doc-sep> tokens employed as separators. We apply global attention to the <doc-sep> tokens, while local attention is used for the remaining tokens. Further details about the hyperparameter settings can be found in Appendix C.

**Datasets** We conduct our evaluation on the Multi-News (Fabbri et al., 2019), WCEP (Gholipour Ghalandari et al., 2020), and DUC 2007 datasets, comparing zero-shot, few-shot, and fully-supervised results. DUC 2007 comprises 45 examples, 20 of which we designate as the test set (Xiao et al., 2022).

**Preprocessing of NewSHead Dataset** We apply the following criteria when preprocessing the dataset:

- **Minimum Document Count in a Cluster**: We require that a news cluster must contain a minimum of three documents, allowing a document to serve as a summary for the remaining documents in the cluster. Clusters not meeting this requirement are excluded.
- **Minimum Summary Size**: We hypothesize that a significant variance in summary lengths during pretraining could hurt performance. Therefore, we ensure that candidate summaries during pretraining are not too short, setting a minimum requirement of 250 tokens. Clusters not meeting this requirement are also excluded. In contrast, Xiao et al. (2022) can control the length of their synthetic reference summaries, ensuring that the sentence count in the synthetic summary constitutes at least 30% of the total sentences in the cluster.

Additional preprocessing steps are outlined in Appendix D. After applying these criteria, we retain 172K clusters, approximately 45% of the total clusters in the NewSHead corpus (Gu et al., 2020).

**Comparison models** In addition to the reported scores of PRIMERA (denoted as PRIMERA* in Table 2), we independently reproduce the PRIMERA model scores by running inference using the PRIMERA checkpoints available in the Transformers (Wolf et al., 2020) library. Similar to the findings of Giorgi et al. (2022), we note that our reproduced scores are lower than those reported by Xiao et al. (2022), an exception being the zero-shot results for the WCEP dataset. For a broader comparison, we also consider the Pegasus model proposed by Zhang et al. (2020). Pegasus is a pretrained model focusing on single-document summarization (SDS), which obtains strong results on multiple SDS datasets such as XSum (Narayan et al., 2018) and CNN-DailyMail (Hermann et al., 2015).

## 4 Results

We conduct experiments in three settings: zero-shot, few-shot, and fully supervised.

**Zero-shot** In the zero-shot setting, we evaluate our pretrained Centrum model on the test datasets of Multi-News, WCEP, and DUC 2007. Following Xiao et al. (2022), the output length of the summary is set as the average length of the gold summaries of the test datasets. As Table 2 illustrates, Centrum outperforms the PRIMERA model in terms of ROUGE scores across all three datasets.

**Few-shot** In the few-shot setting, we follow the approach of Xiao et al. (2022) by conducting two sets of experiments. We randomly select 10 and 100 examples from the training set for model finetuning, and an equivalent number of examples from the validation set. To account for potential variance in scores due to example selection, we repeat this process five times with different seeds.

We observe that the summaries generated by Centrum are, on average, longer than those produced by PRIMERA. This is primarily a result of the Centrum pretraining objective, which imposes a minimum summary length of 250 tokens. In contrast, PRIMERA synthetic summaries are restricted

| | Zero Shot | | | 10 Examples | | | 100 Examples | | |
|---|---|---|---|---|---|---|---|---|---|
| System | R1 | R2 | RL | R1 | R2 | RL | R1 | R2 | RL |
| Multi-News (256) | | | | | | | | | |
| Pegasus* | 32.0 | 10.1 | 16.7 | 39.0 | 12.1 | 20.3 | 43.0 | 13.5 | 21.1 |
| PRIMERA* | 42.0 | 13.6 | 20.8 | 44.0 | 15.5 | 22.0 | 46.0 | 16.8 | 22.9 |
| PRIMERA | 41.6 | 13.1 | 19.9 | 43.4 | 15.3 | 21.6 | 45.2 | 16.3 | 22.7 |
| Centrum | **43.5** | **15.7** | **22.4** | **43.4** | **16.6** | **22.2** | **45.7** | **16.8** | **23.2** |
| WCEP (50) | | | | | | | | | |
| Pegasus* | 33.2 | 12.7 | 23.8 | 35.6 | 14.8 | 26.8 | 42.1 | 19.9 | 33.0 |
| PRIMERA* | 28.0 | 10.3 | 20.9 | 39.0 | 17.6 | 30.6 | 43.0 | 20.5 | 33.9 |
| PRIMERA | 32.9 | 12.1 | 23.4 | 37.0 | 15.8 | 28.2 | **42.4** | **20.5** | **33.4** |
| Centrum | **35.7** | **14.2** | **25.8** | **38.2** | **17.0** | **29.5** | 42.0 | 20.1 | 33.0 |
| DUC2007 (250) | | | | | | | | | |
| Pegasus | 22.7 | 4.2 | 12.8 | 23.1 | 3.5 | 15.2 | - | - | - |
| PRIMERA | 31.9 | 5.4 | 14.2 | 34.6 | 6.6 | 15.2 | - | - | - |
| Centrum | **32.7** | **5.7** | **15.0** | **35.3** | **7.7** | **16.8** | - | - | - |

Table 2: This table presents the ROUGE scores for zero-shot and few-shot evaluations on the Multi-News, WCEP, and DUC datasets. PRIMERA* represents the scores reported by Xiao et al. (2022), while PRIMERA corresponds to the scores we reproduced using their provided checkpoints. The figures in parentheses denote the maximum length set during inference. Due to the DUC 2007 dataset's total size of 45 examples, results for few-shot evaluations with 100 examples are not provided. Our proposed model, Centrum, surpasses PRIMERA in zero-shot and few-shot (10 examples) settings, and performs comparably in the few-shot (100 examples) setting.

| System | R1 | R2 | RL |
|---|---|---|---|
| PRIMERA* | 49.9 | 21.1 | 25.9 |
| PRIMERA | **50.0** | **20.6** | **25.5** |
| Centrum | 49.0 | 20.4 | 25.4 |

Table 3: Comparison of fully-supervised models based on ROUGE scores on the Multi-News dataset. Our proposed model, Centrum, demonstrates performance on par with PRIMERA.

to a maximum length equating to 30% of the input set. To ensure a fair comparison, we truncate the summaries in the few-shot setting to match the lengths assigned in the zero-shot setting.

Table 2 presents the average scores obtained over the five seeds. Given that the DUC 2007 dataset contains only 45 examples, results are reported for training and validation with 10 examples. From the results, we see that Centrum outperforms PRIMERA across all datasets when finetuned with 10 examples. Furthermore, Centrum maintains performance parity with PRIMERA when finetuned using 100 examples.

**Fully supervised** In this setting, the pretrained models are finetuned on the training split of the Multi-News dataset. As reported in Table 3, the results from the fully-supervised experiments demonstrate that Centrum performs on par with PRIMERA on the Multi-News dataset.

**Human Evaluation** To complement the automatic evaluation results, we conduct a human evaluation study. Three professional linguists are tasked with comparing the outputs of Centrum, PRIMERA, and Pegasus using the DUC 2007 dataset, and are compensated at rates higher than local minimum wages. The evaluation focuses on three metrics as outlined by Angelidis et al. (2021): informativeness (which assesses the consistency between model output and the human reference summary), coherence (which evaluates the ordering of information in the summary), and non-repetition (where a higher-quality summary exhibits fewer repetitions of words, phrases, or sentences).

The evaluators are presented with three summaries from the three models, randomly ordered, along with the reference summary. They are then instructed to rank the summaries from best (+1) to worst (-1) for each of the three metrics. These rankings are summed and scaled by the number of examples (20), producing scores that range from 100% (best) to –100% (worst). The results of this human evaluation are presented in Table 4.

Our findings indicate that Centrum significantly outperforms Pegasus across all three metrics, as confirmed by a one-way ANOVA with a post-hoc Tukey test ($p \leq 0.05$). In comparison to PRIMERA, Centrum is significantly better in terms of informativeness and performs comparably in terms of coherence and non-repetition. Pegasus, on the other hand, is marked by heavy repetition within its summaries, which likely accounts for its lower scores.

| | Inform | Coh | Rep |
|---|---|---|---|
| Pegasus | -100* | -100* | -100* |
| PRIMERA | 34.5* | 46.6 | 58.6 |
| Centrum | 65.5 | 53.4 | 41.4 |

Table 4: Human evaluation results for the DUC2007 dataset, with higher scores being preferable. We compare the Pegasus, PRIMERA, and Centrum models across three metrics: informativeness (Inform), coherence (Coh), and avoidance of repetition (Rep). Results that are statistically significantly different from Centrum are marked with an asterisk (*).

## 5 Conclusion

We propose a centroid-based pretraining objective for multi-document summarization. Through experiments, we see that our model Centrum outperforms the existing state-of-the-art model PRIMERA on zero-shot settings and is comparable with PRIMERA in few-shot and supervised settings.

## 6 Limitations

As mentioned in the main paper, one of the limitations of our Centrum model is that it tends to produce longer outputs in comparison to PRIMERA. This necessitates controlling the length of the summary by truncating to a desired length. Moreover, due to our requirement of at least three documents in a cluster for centroid computation, we are unable to utilize clusters of only two documents present in Gu et al. (2020). This constraint significantly reduces the utilizable corpus size, leading us to work with roughly 45% of the corpus size used by PRIMERA. Future research could explore the possibility of initializing Centrum with the gap sentence generation-based Pegasus (Zhang et al., 2020) single document summarization objective, potentially allowing for full utilization of the corpus size of Gu et al. (2020).

## Acknowledgements

This research was supported by funding from the Institute for Infocomm Research (I2R) under A*STAR ARES, Singapore, and by the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme. The work was supported in part by ERC Advanced Fellowship GA 742137 SEMANTAX and the University of Edinburgh Huawei Laboratory. Parag is supported by Huawei and the UKRI Centre for Doctoral Training in Natural Language Processing (grant EP/S022481/1). We thank the anonymous reviewers for their constructive feedback.

## References

Stefanos Angelidis, Reinald Kim Amplayo, Yoshihiko Suhara, Xiaolan Wang, and Mirella Lapata. 2021. Extractive opinion summarization in quantized transformer spaces. *Transactions of the Association for Computational Linguistics*, 9:277–293.

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

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.

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. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1302–1308, Online. Association for Computational Linguistics.

John Giorgi, Luca Soldaini, Bo Wang, Gary Bader, Kyle Lo, Lucy Lu Wang, and Arman Cohan. 2022. Exploring the challenges of open domain multi-document summarization.

Xiaotao Gu, Yuning Mao, Jiawei Han, Jialu Liu, Hongkun Yu, You Wu, Cong Yu, Daniel Finnie, Jiaqi Zhai, and Nicholas Zukoski. 2020. Generating Representative Headlines for News Stories. In *Proc. of the the Web Conf. 2020*.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 1693–1701.

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.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Peter J. Liu*, Mohammad Saleh*, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences. In *International Conference on Learning Representations*.

Kathleen McKeown and Dragomir R. Radev. 1995. Generating summaries of multiple news articles. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '95, pages 74–82, New York, NY, USA. Association for Computing Machinery.

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*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Ani Nenkova and Rebecca Passonneau. 2004. 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.

Dragomir R. Radev and Kathleen R. McKeown. 1998. Generating natural language summaries from multiple on-line sources. *Computational Linguistics*, 24(3):469–500.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Nikolai Vogler, Songlin Li, Yujie Xu, Yujian Mi, and Taylor Berg-Kirkpatrick. 2022. An unsupervised masking objective for abstractive multi-document news summarization. *CoRR*, abs/2201.02321.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. 2022. PRIMERA: Pyramid-based masked sentence pre-training for multi-document summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5245–5263, Dublin, Ireland. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.

## A Potential Risks

Despite our model's potential, there is a risk that the generated summaries might not accurately represent the input document due to noise present in the training and finetuning examples. At the same time, we believe that our Centrum pretraining strategy doesn't affect the factuality of the model either positively or negatively compared to Xiao et al. (2022). Future research will aim to explicitly evaluate and improve the factuality of our model's output.

## B Details of the Datasets

Table 5 provides detailed information about the datasets used in our study. The NewSHead, MultiNews, and DUC 2007 datasets all originate from the news domain, while the WCEP dataset is derived from the Wikipedia Current Events Portal.

## C Hyperparameter Details

Our hyperparameters are similar to Xiao et al. (2022). We train for 100K steps with a learning rate of 3e-5. We evaluate every 500 steps and early-stop on the validation perplexity with a patience of

| Name | #Ex | #Doc/C | #L$_{doc}$ | #L$_{summ}$ |
|---|---|---|---|---|
| NewSHead (2020) | 177K | 4.2 | 1692 | 484 |
| Multi-News (2019) | 56K | 2.8 | 1793 | 217 |
| WCEP (2020) | 10K | 9.1 | 3866 | 28 |
| DUC 2007 | 45 | 25 | 540 | 250 |

Table 5: Characteristics of the datasets utilized in this paper. The notations are as follows: #Ex represents the number of examples, #Doc/C is the average number of documents per cluster, #L$_{doc}$ signifies the average token count in the input, and #L$_{summ}$ indicates the average token count in the summary. Values associated with the Multi-News and WCEP datasets are sourced from Xiao et al. (2022).

50. Pretraining Centrum on a 4-node A100 GPU took around 4 days. We computed the results using ROUGE (Lin, 2004) library [2] with the default settings and '–use_stemmer' argument.

## D Additional Preprocessing Steps

- **Removing boilerplate text from summaries**: We remove boilerplate text such as "Sorry, this video isn't available any more.", "Advertisement Story continues below" from the summary sentences using regular expression based cleaning.
- **Truncation of documents**: We truncate each document in the cluster to the maximum length of source context allowed in LED divided by the count of the documents in the cluster. Thus, each document has a proportional representation in the cluster, similar to Xiao et al. (2022).

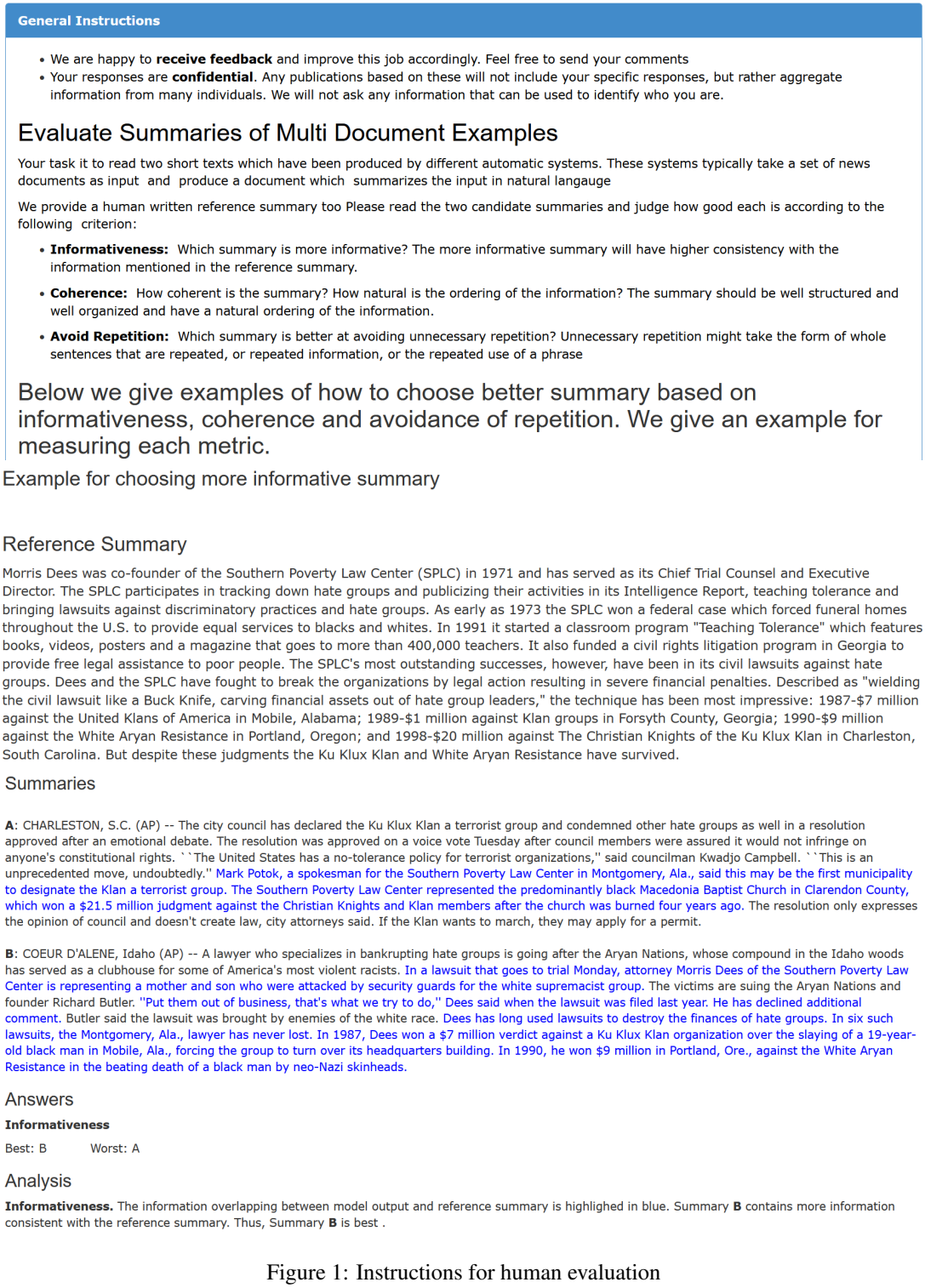
## E Software and Licenses

Our model relies on datasets downloaded from HuggingFace datasets (Lhoest et al., 2021) (Apache 2.0). We release our models under the Apache 2.0 license.

## F Human Evaluation

Figures 1 and 2 show the screenshots of the user interface presented to the raters.

[2] https://github.com/google-research/google-research/tree/master/rouge



**General Instructions**

- We are happy to **receive feedback** and improve this job accordingly. Feel free to send your comments
- Your responses are **confidential**. Any publications based on these will not include your specific responses, but rather aggregate information from many individuals. We will not ask any information that can be used to identify who you are.

## Evaluate Summaries of Multi Document Examples

Your task it to read two short texts which have been produced by different automatic systems. These systems typically take a set of news documents as input and produce a document which summarizes the input in natural langauge

We provide a human written reference summary too Please read the two candidate summaries and judge how good each is according to the following criterion:

- **Informativeness:** Which summary is more informative? The more informative summary will have higher consistency with the information mentioned in the reference summary.
- **Coherence:** How coherent is the summary? How natural is the ordering of the information? The summary should be well structured and well organized and have a natural ordering of the information.
- **Avoid Repetition:** Which summary is better at avoiding unnecessary repetition? Unnecessary repetition might take the form of whole sentences that are repeated, or repeated information, or the repeated use of a phrase

## Below we give examples of how to choose better summary based on informativeness, coherence and avoidance of repetition. We give an example for measuring each metric.

### Example for choosing more informative summary

### Reference Summary

Morris Dees was co-founder of the Southern Poverty Law Center (SPLC) in 1971 and has served as its Chief Trial Counsel and Executive Director. The SPLC participates in tracking down hate groups and publicizing their activities in its Intelligence Report, teaching tolerance and bringing lawsuits against discriminatory practices and hate groups. As early as 1973 the SPLC won a federal case which forced funeral homes throughout the U.S. to provide equal services to blacks and whites. In 1991 it started a classroom program "Teaching Tolerance" which features books, videos, posters and a magazine that goes to more than 400,000 teachers. It also funded a civil rights litigation program in Georgia to provide free legal assistance to poor people. The SPLC's most outstanding successes, however, have been in its civil lawsuits against hate groups. Dees and the SPLC have fought to break the organizations by legal action resulting in severe financial penalties. Described as "wielding the civil lawsuit like a Buck Knife, carving financial assets out of hate group leaders," the technique has been most impressive: 1987-$7 million against the United Klans of America in Mobile, Alabama; 1989-$1 million against Klan groups in Forsyth County, Georgia; 1990-$9 million against the White Aryan Resistance in Portland, Oregon; and 1998-$20 million against The Christian Knights of the Ku Klux Klan in Charleston, South Carolina. But despite these judgments the Ku Klux Klan and White Aryan Resistance have survived.

### Summaries

**A**: CHARLESTON, S.C. (AP) -- The city council has declared the Ku Klux Klan a terrorist group and condemned other hate groups as well in a resolution approved after an emotional debate. The resolution was approved on a voice vote Tuesday after council members were assured it would not infringe on anyone's constitutional rights. ``The United States has a no-tolerance policy for terrorist organizations,'' said councilman Kwadjo Campbell. ``This is an unprecedented move, undoubtedly.'' Mark Potok, a spokesman for the Southern Poverty Law Center in Montgomery, Ala., said this may be the first municipality to designate the Klan a terrorist group. The Southern Poverty Law Center represented the predominantly black Macedonia Baptist Church in Clarendon County, which won a $21.5 million judgment against the Christian Knights and Klan members after the church was burned four years ago. The resolution only expresses the opinion of council and doesn't create law, city attorneys said. If the Klan wants to march, they may apply for a permit.

**B**: COEUR D'ALENE, Idaho (AP) -- A lawyer who specializes in bankrupting hate groups is going after the Aryan Nations, whose compound in the Idaho woods has served as a clubhouse for some of America's most violent racists. In a lawsuit that goes to trial Monday, attorney Morris Dees of the Southern Poverty Law Center is representing a mother and son who were attacked by security guards for the white supremacist group. The victims are suing the Aryan Nations and founder Richard Butler. ''Put them out of business, that's what we try to do,'' Dees said when the lawsuit was filed last year. He has declined additional comment. Butler said the lawsuit was brought by enemies of the white race. Dees has long used lawsuits to destroy the finances of hate groups. In six such lawsuits, the Montgomery, Ala., lawyer has never lost. In 1987, Dees won a $7 million verdict against a Ku Klux Klan organization over the slaying of a 19-year-old black man in Mobile, Ala., forcing the group to turn over its headquarters building. In 1990, he won $9 million in Portland, Ore., against the White Aryan Resistance in the beating death of a black man by neo-Nazi skinheads.

### Answers

**Informativeness**

Best: B Worst: A

### Analysis

**Informativeness.** The information overlapping between model output and reference summary is highlighed in blue. Summary **B** contains more information consistent with the reference summary. Thus, Summary **B** is best .

Figure 1: Instructions for human evaluation

Example on choosing a better summary for coherence

Below are two model generated summaries. The example is about choosing a more coherent summary

**A**: MOBILE, Ala. (AP) -- A white supremacist arrested after buying hand grenades from an undercover agent said he wanted to send mail bombs to Washington and Montgomery, authorities said. Chris Scott Gilliam said he didn't want to be like the Unabomber, who killed three person and wounded several others, a federal agent testified at a federal court hearing. ``He wanted to kill everybody,'' David Pasqualotto, a special agent with the Bureau of Alcohol, Tobacco and Firearms, said Friday. There was no immediate indication what agencies or people might have been targeted in the national and state capital cities. An ATF representative did not immediately return calls for comment Saturday. Gilliam, 27, was charged with possessing an unregistered firearm found at his home in Foley along with what agents said was apparently a silencer.

**B**: MOBILE, Ala. (AP) -- There was no immediate indication what agencies or people might have been targeted in the national and state capital cities. An ATF representative did not immediately return calls for comment Saturday. Chris Scott Gilliam said he didn't want to be like the Unabomber, who killed three person and wounded several others, a federal agent testified at a federal court hearing. A white supremacist arrested after buying hand grenades from an undercover agent said he wanted to send mail bombs to Washington and Montgomery, authorities said. Gilliam, 27, was charged with possessing an unregistered firearm found at his home in Foley along with what agents said was apparently a silencer. ``He wanted to kill everybody,'' David Pasqualotto, a special agent with the Bureau of Alcohol, Tobacco and Firearms, said Friday.

Answers

**Coherence**

Best: A Worst: B

Analysis

**Coherence.** Summary **A** contains the information in a coherent manner. In Summary **B** , in contrast, the information is ordered in a less natural way Thus, Summary **A** is best .

Below are two model generated summaries. The example is about choosing a summary which is less repetitive

**A**: COEUR D'ALENE, Idaho (AP) -- A jury on Thursday awarded $6.3 million to a woman and her son who were attacked by Aryan Nations guards outside the white supremacist group's north Idaho headquarters. The jury found that Aryan Nations leader Richard Butler, the group and its corporate entity, Saphire, Inc., were negligent in the selection, training and supervision of the security guards who assaulted Victoria and Jason Keenan two years ago. The Keenans' attorney, Morris Dees, had asked the jury to award more than $11 million in punitive damages. Dees, of the Montgomery, Ala.-based Southern Poverty Law Center, has said he hoped the penalty would be severe enough to bankrupt the Aryan Nations. The 82-year-old Butler spoke briefly outside the courtroom, comparing himself to some biblical figures and also declaring northern Idaho a haven for racists. ''This is nothing,'' Butler said. ''We have planted seeds. Most of north Idaho now is filled with the people who escaped multiculturalism or diversity or whatever you want to call it.'' Jurors set $6 million as a punitive damage award, with $330,000 in compensatory damages to the Keenans, who were chased, shot at and assaulted after they stopped to search for a dropped wallet near the entrance to the Aryan Nations' compound.

**B**: COEUR D'ALENE, Idaho (AP) -- A jury on Thursday awarded $6.3 million to a woman and her son who were attacked by Aryan Nations guards outside the white supremacist group's north Idaho headquarters. The jury found that Aryan Nations leader Richard Butler, the group and its corporate entity, Saphire, Inc., were negligent in the selection, training and supervision of the security guards who assaulted Victoria and Jason Keenan two years ago.(i) The Keenans' attorney, Morris Dees, had asked the jury to award more than $11 million in punitive damages. Dees, of the Montgomery, Ala.-based Southern Poverty Law Center, has said he hoped the penalty would be severe enough to bankrupt the Aryan Nations. The jury found that Aryan Nations leader Richard Butler was negligent in the selection of the security guards who assaulted Victoria and Jason Keenan two years ago. (i) ''This is nothing,'' Butler said. ''We have planted seeds. (ii)Most of north Idaho now is filled with the people who escaped multiculturalism or diversity or whatever you want to call it.'' ''This is nothing,'' Butler said. ''We have planted seeds. (ii) Jurors set $6 million as a punitive damage award, with $330,000 in compensatory damages to the Keenans, who were chased, shot at and assaulted after they stopped to search for a dropped wallet near the entrance to the Aryan Nations' compound.

Answers

**Avoiding Repetition**

Best: A Worst: B

Analysis

**Avoiding repetition.** Summary **A** is the best as Summary **B** contains **repetitive information** ↓ such as phrases **(i)** and **(ii)**.

Figure 2: Instructions for human evaluation (continued)

## ACL 2023 Responsible NLP Checklist

### A For every submission:

☑ A1. Did you describe the limitations of your work?
*Section 6*

☑ A2. Did you discuss any potential risks of your work?
*Appendix A*

☑ A3. Do the abstract and introduction summarize the paper's main claims?
*Section 1*

☒ A4. Have you used AI writing assistants when working on this paper?
*Left blank.*

### B ☑ Did you use or create scientific artifacts?

*Section 3 Models and Datasets*

☑ B1. Did you cite the creators of artifacts you used?
*Section 3 Models and Datasets, and Section B of Appendix*

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
*Section E*

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
*Not applicable. Left blank.*

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
*Not applicable. Left blank.*

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
*Section B of Appendix*

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
*Section B*

### C ☑ Did you run computational experiments?

*Section 3*

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
*Section 3, Section C of Appendix*

---

*The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.*

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
*Section 3, Section C of Appendix*

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
*Section 4*

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
*Section C of Appendix*

**D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?**

*Section 4 (Human evaluation)*

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
*Section 4 (Human evaluation) and Section F in Appendix*

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
*Section 4 (Human evaluation)*

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
*Section F in Appendix*

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
*Not applicable. Left blank.*

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
*Not applicable. Left blank.*