Echoes from Alexandria: A Large Resource for Multilingual Book Summarization

Alessandro Scirè[1,2] **Simone Conia**[2] **Simone Ciciliano**[3*] **Roberto Navigli**[2]

Babelscape, Italy
[1]scire@babelscape.com

Sapienza NLP Group
Sapienza University of Rome
[2]{first.lastname}@uniroma1.it

Free University of Bozen
[3]sciciliano@unibz.it

Abstract

In recent years, research in text summarization has mainly focused on the news domain, where texts are typically short and have strong layout features. The task of full-book summarization presents additional challenges which are hard to tackle with current resources, due to their limited size and availability in English only. To overcome these limitations, we present "Echoes from Alexandria", or in shortened form, "Echoes", a large resource for multilingual book summarization. Echoes features three novel datasets: i) Echo-Wiki, for multilingual book summarization, ii) Echo-XSum, for extremely-compressive multilingual book summarization, and iii) Echo-FairySum, for extractive book summarization. To the best of our knowledge, Echoes – with its thousands of books and summaries – is the largest resource, and the first to be multilingual, featuring 5 languages and 25 language pairs. In addition to Echoes, we also introduce a new extractive-then-abstractive baseline, and, supported by our experimental results and manual analysis of the summaries generated, we argue that this baseline is more suitable for book summarization than purely-abstractive approaches. We release our resource and software at https://github.com/Babelscape/echoes-from-alexandria in the hope of fostering innovative research in multilingual book summarization.

1 Introduction

Recent research in Automatic Text Summarization – the task of shortening a text while preserving its meaning – has mainly focused on news stories. News texts are usually short documents; for example, 99.3% and 98.6% of the articles in XSum (Narayan et al., 2018) and CNN/DailyMail (Nallapati et al., 2016), respectively, are shorter than 2048 tokens. Additionally, news stories are characterized by strong layout features, such as the "lead bias", in which the first sentences usually contain the most relevant information for a summary. Accordingly, the Lead-3 baseline, which uses the first three sentences of a news item as its summary, performs competitively on news summarization benchmarks (Gehrmann et al., 2018; Zhu et al., 2019). Although recent approaches have achieved high performance, it is still unclear how they behave on longer documents and whether they can generalize across domains and genres. For this reason, the research community has been shifting toward more challenging settings, which include interviews (Zhu et al., 2021) and scientific articles (Gupta et al., 2021; Cohan et al., 2018).

One setting that has been attracting growing attention is full-book summarization (Kryscinski et al., 2021), i.e., the task of producing the plot of a book from its full text. Summarizing a book is hard not only because of its average text length – currently not processable in a single forward pass even by architectures for long-form text processing (Beltagy et al., 2020; Guo et al., 2022) – but also due to other critical aspects, such as the presence of dialogues, rich discourse structures, parallel and non-linear lines of plot, and long-distance dependencies between entities, among others. Therefore, we deem book summarization a complex testbed to challenge current approaches and investigate their capabilities and limitations.

Although the first small-scale datasets for the task were introduced several years ago (Mihalcea and Ceylan, 2007), the area has recently regained traction thanks to larger-scale resources, such as BookSum (Kryscinski et al., 2021) and NarrativeQA (Kočiský et al., 2017). However, despite this recent progress, current resources for book summarization are still, i) limited in size, making them difficult to use for proper training and evaluation, and ii) monolingual (usually English-only).

[*] Work carried out while at Sapienza University of Rome.

Findings of the Association for Computational Linguistics: ACL 2023, pages 853–867
July 9-14, 2023

To overcome these issues, we introduce "Echoes from Alexandria" (Echoes), the largest resource to date for book summarization and the first one providing books and summaries in multiple languages. We use Echoes to investigate how current summarization approaches perform on a large-scale multilingual summarization dataset, concluding that current purely-abstractive approaches still struggle in our setting. We additionally devise a new baseline, showing that the extractive-then-abstractive paradigm represents a promising direction for future research.

The main contributions of our work are the following:

- We introduce Echoes, the first multilingual resource for book summarization, with thousands of texts and plots in 5 languages, for a total of 25 language pairs. Echoes is also the largest resource among current English datasets for full-book summarization.

- We release the three datasets of Echoes: i) Echo-Wiki, for multilingual abstractive summarization, ii) Echo-XSum, for extremely-compressive multilingual book summarization, and iii) Echo-FairySum, an English dataset for evaluating extractive book summarization.

- We leverage BookSum and Echoes to evaluate state-of-the-art systems, both in zero-shot and fine-tuning settings, bringing to light their inadequate generalization capabilities in book summarization.

- Our experiments demonstrate that an *extractive-then-abstractive* baseline outperforms the purely-abstractive counterpart on our datasets while achieving state-of-the-art results on BookSum.

- We provide a comprehensive manual evaluation of the automatically generated summaries and release the dataset with our human judgments.

We hope our work will foster research in multilingual long document understanding and summarization. We release Echoes and our software for research purposes at `https://github.com/Babelscape/echoes-from-alexandria`.

2 Related Work

Resources for summarization. Research efforts to create summarization resources have steadily increased in numbers over recent years. For the news domain, XSum (Narayan et al., 2018) and CNN/DailyMail (Nallapati et al., 2016) are the *de-facto* standard datasets for training and evaluating summarization systems. XSum comprises 226k news articles accompanied by a one-sentence abstractive summary. In CNN/DailyMail, the authors retrieved 93k articles from CNN[1] and 220k articles from DailyMail[2] newspapers. Both publishers supplement their articles with a list of bullet points containing the main information of the news text.

More recently, summarization resources have been shifting towards more challenging scenarios, i.e., where the documents of interest are longer and belong to different domains. Notably, Cohan et al. (2018) released two large-scale datasets of long and structured scientific papers obtained from arXiv[3] and PubMed[4]. In these datasets, paper abstracts are used as ground truth summaries. Another relevant example is MediaSum (Zhu et al., 2021), a collection of interview transcriptions from National Public Radio (NPR)[5] and CNN, where overview and topic descriptions are employed as summaries.

In long-form text summarization research, a task that is attracting growing attention is book summarization. Although this task was originally introduced several years ago by Mihalcea and Ceylan (2007), who released the first small-scale evaluation resource, book summarization regained traction thanks to a few notable endeavors. The most important example is BookSum (Kryscinski et al., 2021), which provides a collection of resources for book summarization at three levels of granularity: paragraph, chapter, and full book. Book texts are collected from Project Gutenberg, while summaries are obtained from the Web Archive.[6] BookSum features 222 unique book titles with a total of 6,987 book chapters and 142,753 paragraphs. Relatedly, NarrativeQA (Kočiský et al., 2017) is a collection of 1572 stories retrieved from Project Gutenberg (783 books and 789 movie scripts) associated with summaries from Wikipedia. The annotators were required to generate questions and answers based

[1] https://www.edition.cnn.com/
[2] https://www.dailymail.co.uk/
[3] https://arxiv.org/
[4] https://pubmed.ncbi.nlm.nih.gov/
[5] https://www.npr.org/
[6] https://web.archive.org/

on the summaries. Even if NarrativeQA is primarily intended for Question Answering, it can also be used for book summarization. Due to their limited size, however, BookSum (in the full-book setting) and NarrativeQA can be more useful for evaluating models on the task rather than for training purposes. It is also worth noting that these resources are monolingual, i.e., English-only, limiting their usefulness for researchers seeking to evaluate multilingual summarization models. Despite the great work carried out so far, we argue that there is still ample room to improve book summarization resources.

Approaches to book summarization. Kryscinski et al. (2021) conducted experiments on full-book summarization using a generate&rank strategy. This approach involves training a system to generate paragraph-level summaries, which are then sorted by perplexity and concatenated to form a full-book summary. More recently, Wu et al. (2021) proposed an approach where passages are recursively summarized and concatenated to form a full summary. However, generated summaries are affected by the errors accumulated from previous stages (Wu et al., 2021). Recursively generating a summary is a paradigm that has also been used by other works for long-document summarization (Zhang et al., 2021; Gidiotis and Tsoumakas, 2020). Another family of approaches is that of *extractive-then-abstractive* approaches. This family of approaches first extracts key sentences from the input document and then uses such sentences as input to an abstractive model, which is tasked with generating a summary that captures the main ideas and themes of the source. While it was successfully employed in previous works for short (Li et al., 2021) and long-form summarization (Chen and Bansal, 2018), this paradigm has never been explored for summarizing books. In this paper, we aim to fill this gap by presenting a new, simple extractive-then-abstractive model and showing its effectiveness for book summarization.

3 Echoes

Echoes is the first collection of resources for book summarization in 5 languages: English, French, German, Italian, and Spanish. With Echoes, we introduce the following three novel datasets:

- **Echo-Wiki**, in which we pair book texts with plots retrieved from a hand-curated list of Wikipedia page sections.
- **Echo-XSum**, in which we pair book texts with extremely-compressive summaries, manually created starting from the lead section of Wikipedia pages.
- **Echo-FairySum**, an evaluation dataset for extractive summarization of short stories and fairy tales, composed of 197 English manually-annotated extractive summaries.

We provide an overview of the main differences between Echoes and existing resources in Table 1.

3.1 Text collection

We collect the book texts that comprise Echoes from two main sources: Project Gutenberg and Wikisource. Project Gutenberg is a digital library that provides free access to public-domain books and features over 60k texts. We collect all the available books from Project Gutenberg by following their robot-access policies.[7] While often considered one of the most reliable sources of copyright-free books, Project Gutenberg provides only very limited coverage of non-English books and non-English translations of English books. This is one of the reasons why we also rely on Wikisource. Part of the Wikimedia Foundation, Wikisource contains a huge number of texts from a wide range of domains, e.g., books, and legal and historical documents, in various languages. Therefore, for Echoes, we rely on Wikisource in English, French, German, Spanish, and Italian to retrieve other book texts and expand the coverage of books already available from Project Gutenberg.[8] We call this set of full-text books B. We note that Wikisource can also be used to expand Echoes to other languages. Given the limited amount of work in multilingual summarization, we focus on the five above high-resource languages. We defer the expansion of Echoes to future work.

While Project Gutenberg has already been used as a source of books in previous resources, such as BookSum and NarrativeQA, the use of Wikisource is what enables Echoes to become the largest resource for book summarization in English and the first resource for multilingual book summarization.

[7] https://www.gutenberg.org/help/mirroring.html

[8] Wikisource dumps are freely available to download at https://dumps.wikimedia.org/<l>wikisource/ where <l> ∈ {EN, FR, DE, ES, IT}. Last accessed: July 1, 2022.

Dataset	Languages	# Documents	Coverage	Density	C. Ratio	Avg. length (# Tokens)	
						Source	Summary
XSum	EN	226,677	0.66	1.09	19.3	438.4	23.9
CNN/DailyMail	EN	311,971	0.85	3.47	14.9	803.7	59.7
ArXiv/PubMed	EN	346,187	0.87	3.94	31.2	5,179.2	257.4
MediaSum	EN	463,596	0.80	1.86	116.3	1,925.8	16.6
BookSum (full)	EN	405	0.89	1.83	126.2	112,885.2	1,167.2
Echo-Wiki	EN, FR, DE, ES, IT	5,001	0.79	2.08	103.7	75,600.9	729.4
Echo-Wiki$_{EN}$	EN	2,375	0.84	2.24	117.1	83,724.1	678.0
Echo-XSum	EN, FR, DE, ES, IT	3,383	0.78	1.67	1624.0	86,040.0	53.0
Echo-XSum$_{EN}$	EN	1,828	0.81	1.78	1706.1	90,971.9	53.0
Echo-FairySum	EN	197	1.00	1.00	2.8	4,438.8	1,506.2

Table 1: Comparison of Echoes (Echo-Wiki, Echo-XSum, and Echo-FairySum) with existing resources for summarization. **Coverage and density:** measures of the "extractiveness" of a summary. **Compression Ratio:** micro-average ratio between the lengths of the source and the summary.

3.2 Pairing books with Wikipedia summaries

Book summaries from Wikipedia follow a standard set of guidelines[9] and are often of remarkable quality, as they are continuously refined over time by the Wikipedia community. Therefore, once we have collected our set of full-book texts (see Section 3.1), we iterate over the Wikipedia dumps[10] in English, French, German, Italian, and Spanish. Given our set B of full-book texts, and W, the set of Wikipedia pages, our objective is to uniquely associate a book $b \in B$ to a page $w \in W$, such that w is the Wikipedia page of book b. We obtain a set of potential matches by finding Wikipedia pages whose contents contain a hyperlink to a book in B. To improve the accuracy of our mapping, we first apply a string distance metric[11] to compare the titles of the books and their associated Wikipedia pages. We then check if the lead section of the Wikipedia page in question mentions the surname of the author of the associated book. This additional step helps us further refine and ensure the validity of our associations.

After our matching process, we manually inspect the cases in which books are associated with multiple Wikipedia pages. We discover that the pages in excess refer to adaptations of the book in other mediums, such as movies and theatrical plays. To resolve this ambiguity, we utilize the mapping between Wikipedia pages and Wikidata nodes to obtain metadata about the medium, e.g., *book, movie, play*, and retain only the Wikipedia page that corresponds to the book.

At this point, given the Wikipedia page content, our goal is to extract only the book summary and discard other information, such as the biography of the author, historical background, prizes and accolades, and critical reception, among others. To achieve this, we employ native speakers to manually identify a list of section names that, in the different languages, only contain plot information, aiming for high precision rather than coverage. We use the content of these identified sections as summaries and provide our list of section names in Appendix A for reference. We name the resulting set of (Wikipedia summary, full-text book) pairs **Echo-Wiki**.

We note that the average number of unique editors (220.6), revisions (421.4), and year of creation (2008) of the Wikipedia pages we select for the Echo-Wikidataset are large: this indicates that their book summaries have been curated over time and suggests that they are of high quality. Table 1 shows how Echo-Wiki compares against BookSum, the previous largest existing dataset for book summarization, to the best of our knowledge. Besides being multilingual, it is worth noticing that Echo-Wiki is about 12 times larger than BookSum (5,001 vs. 405 books) while still featuring similar compression ratios (103.7 vs. 126.2).

3.3 Enabling extreme summarization of books

Inspired by the work of Narayan et al. (2018) on the news domain with XSum, which showcases the capabilities of highly-abstractive summarization, we

[9] https://en.wikipedia.org/wiki/Wikipedia:How_to_write_a_plot_summary

[10] Wikipedia dumps are freely available to download at https://dumps.wikimedia.org/<l>wiki/ where <l> ∈ { EN, FR, DE, ES, IT}. Last accessed: July 1, 2022.

[11] We used the Edit distance to retain only those pairs whose titles were highly similar, by setting a stringent threshold (0.2).

introduce **Echo-XSum**, a new dataset for training and evaluating systems for extreme summarization of books. In Echo-XSum, we pair full-text books with very short summaries. These summaries contain the minimum number of sentences required to provide an overview of the main contents of a book, typically one to three sentences. The main challenge posed by Echo-XSum is dealing with the great disparity between the size of the input and the size of the output. Indeed, as we can observe in Table 1, the compression ratio of Echo-XSum (1624.0) is unprecedented in the field of summarization, being an order of magnitude greater than those of Echo-Wiki (103.7) and BookSum (126.2).

The extreme summaries in Echo-XSum are the result of a manual annotation process, which involved an expert linguist who is a fluent speaker in all 5 languages of Echoes. The annotator was explicitly contracted for this task. Given a book and its previously-identified Wikipedia page (see Section 3.1), the annotator was tasked with extracting portions of text from the introduction that described the essential plot of a book. An excerpt of a book text with the corresponding multilingual summaries from Echo-XSum can be found in Appendix B. Notice that the portions of text extracted by the annotator are not necessarily contiguous, as long as the extracted text can be read independently of its original text. As a rule of thumb for the annotation process, the linguist followed the definitions of Consistency, Relevance, Fluency, and Coherence of a summary (Fabbri et al., 2021). The annotator spent an average of 5 minutes per sample. We provide an example of the annotations produced in Appendix C. At the end of the manual creation of our extreme summaries, the resulting Echo-XSum is still about 8 times larger than BookSum (3,383 vs. 405 books).[12]

3.4 Classifying books into genres

Differently from existing resources, such as BookSum, which is limited by its relatively small size, the thousands of books in Echoes give us the opportunity to investigate book summarization more in-depth. Indeed, books in Echoes cover a wide range of genres, including novels, theatrical plays, and poems, among others. We argue that developing a strategy to automatically identify book genres provides valuable insights into the dataset and enables a fine-grained evaluation of current and future summarization approaches. An analysis by genre can help us determine which genres are the most challenging to summarize.

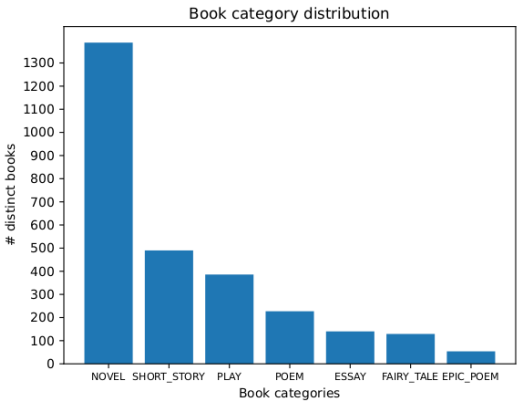

Figure 1: Distribution of the genres – novels, short stories, play, poems, essays, fairy tales, and epic poems – in the English partition of Echo-Wiki.

Similarly to what was described in Section 3.2, we rely on a graph-based heuristic on the knowledge graph of Wikidata to identify genres. More specifically, given a Wikipedia article of a book, we retrieve its corresponding Wikidata node, and analyze its relations (e.g., *genre* and *form_of_creative_work*) with its neighboring nodes. This process is able to distinguish between 7 main genres: novels, plays, poems, epic poems, short stories, fairy tales, and essays. Note that our heuristic may assign more than one genre to a single book. Figure 1 illustrates the distribution of the genres in the English partition of Echo-Wiki, showing that novels are the most represented genre, followed by short stories and plays.

3.5 Digging up extractive summarization

Over the past few years, the attention of the research community has gradually shifted from extractive to abstractive summarization, especially thanks to the advent of flexible sequence-to-sequence models, which have proven effective for summarizing short documents. Thanks to genre classification (see Section 3.4), we are able to perform a small-scale investigation of extractive book summarization on two genres in Echoes. More specifically, we construct **Echo-FairySum**, the first evaluation dataset for extractive summarization of fairy tales and short stories.

To create extractive summaries for Echo-FairySum, we set up the following manual annotation process: given the text of a book, and its

[12]Echo-XSum includes fewer book/summary pairs than Echo-Wiki because the annotator was not able to find an extreme summary in the Wikipedia pages of some books.

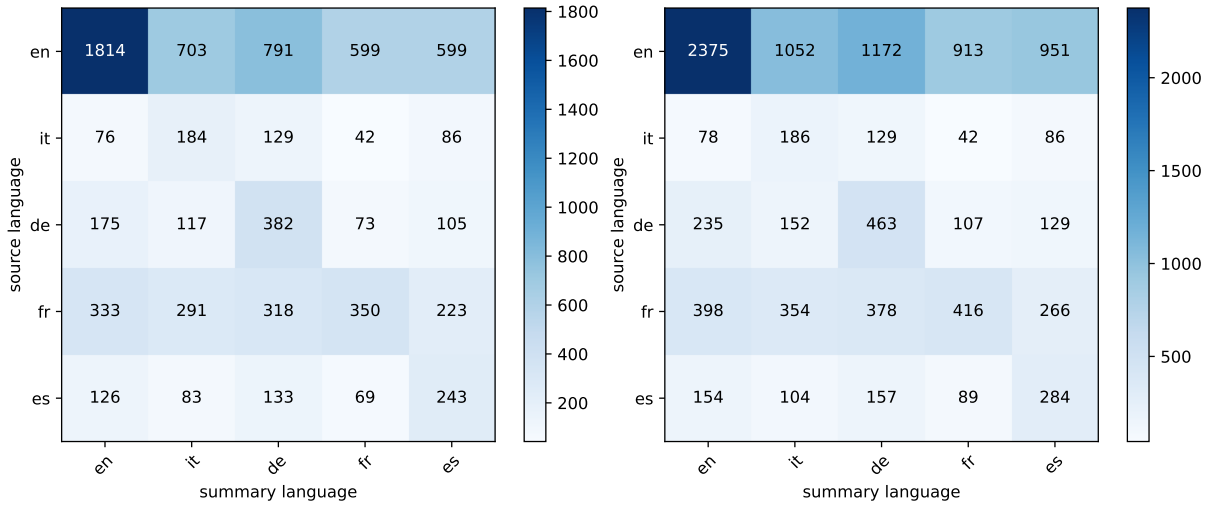

Figure 2: Number of *book-summary* (left) and *version-summary* pairs (right) for all language pairs in Echo-Wiki. Best seen in color.

abstractive summary from Wikipedia (Section 3.2), annotators are required to extract relevant sentences from the book text. A sentence is relevant if it provides a piece of information that is also contained in the abstractive summary. The annotators were asked to adhere as closely as possible to the concepts of Consistency, Relevance, and Coherence defined by Fabbri et al. (2021). The annotators were drawn from a pool of fifty-eight Master-level students from the 'Narrative Understanding and Storytelling' minicourse held at the Sapienza University of Rome by the last co-author, as part of the AI and Robotics degree. The selected students carried out the task as part of their course assignments. On average, each student annotated 3 texts, resulting in multiple annotations for each text. The annotation agreement was measured using Cohen's Kappa coefficient, which indicated substantial agreement (0.71). A subset of annotations was further validated by our contracted annotator to ensure that the students were adhering to the guidelines. Overall, Echo-FairySum provides extractive summaries for 197 documents, about 4 times the size of the test set of BookSum.

3.6 Aggregating books across versions and languages

A book can be published in various editions after its original publication. Perhaps most importantly, the same version of a book can also be translated into multiple languages. Given the potentially large variety of versions and translations of a book, we argue that it is important to aggregate those versions. Indeed, aggregating books across versions and translations can allow Echoes to also be employed for machine translation, cross-lingual sentence alignment, and cross-lingual summarization.

To achieve this objective, we leverage two characteristics of Wikipedia. First, we aggregate all those book texts aligned to the same Wikipedia page (see Section 3.2). We increase the accuracy of this step by taking into account the information found on some Wikisource pages, which list the editions available for some books. Second, we navigate the Wikipedia interlanguage links, which connect pages that refer to the same concept/entity in different languages, to aggregate different translations and summaries (in different languages) of the same book. Figure 2 presents the number of *book-summary* and the *version-summary* pairs for all the language pairs in Echo-Wiki obtained after our aggregation process.

4 Experiments and Results

In recent years, two promising paradigms have emerged from previous work on long-document summarization: *recursive-abstractive* and *extractive-then-abstractive*. In this section, we evaluate and analyze their effectiveness on Echoes.

4.1 Recursive-abstractive approaches

Recursive-abstractive approaches consist in dividing the source document into smaller segments, referred to as chunks, and then using an abstractive summarization model to summarize each segment. If the concatenated output summaries are still larger

	Model	R-1	R-2	R-L	BERTScore
recursive-abs.	BART_{XSum}	18.02	2.91	13.81	0.438
	$\text{BART}_{MediaSum}$	13.95	5.11	12.72	0.416
	LED_{XSum}	18.86	2.99	14.83	0.440
	$\text{LED}_{MediaSum}$	14.69	4.26	12.79	0.421
	LongT5_{XSum}	14.53	2.31	12.05	0.413
	$\text{LongT5}_{MediaSum}$	16.54	5.47	14.35	0.429
extractive-abs.	BART	30.44	12.41	25.76	0.557
	BART_{XSum}	**30.78**	13.44	**26.73**	0.558
	LED	30.18	12.73	25.79	0.558
	LED_{XSum}	30.22	13.05	26.28	**0.560**
	LongT5	30.05	**13.52**	26.02	**0.560**
	LongT5_{XSum}	29.42	13.35	26.00	0.557

Table 2: Automatic evaluation of recursive-abstractive and extractive-then-abstractive approaches on Echo-XSum.

than a single chunk, the recursive-abstractive approach repeats the process by treating the concatenation as a new source document and summarizing it in the same way. The recursive process continues until the concatenated output summaries are short enough to be considered as the final summary, i.e., until their size is shorter than the maximum size of a single chunk.

Experimental setting. In its simplest form, a recursive-abstractive approach requires a model trained on a standard summarization dataset; this model is then employed recursively, as described above. For our experiments, we consider three sequence-to-sequence Transformer-based models – BART-large (Lewis et al., 2020), LED-base (Beltagy et al., 2020), and LongT5-base (Guo et al., 2022) – and train them on XSum (short documents, news) and MediaSum (long documents, interviews). Then, we evaluate our trained models on the test set of Echo-XSum,[13] whose summaries feature an average length similar to that of the summaries in XSum and MediaSum but belong to a different genre (books). For the evaluation, we adopt standard summarization metrics, such as ROUGE-1, ROUGE-2, ROUGE-L, and BERTScore (Zhang et al., 2019).

Results. Table 2 (top) provides an overview of the results obtained by our recursive-abstractive baseline using different language models and trained on different summarization datasets. Overall, we can observe that, independently of the language model and training dataset employed, the baseline does not achieve good results on Echo-XSum. Indeed, the best configuration (LED_{XSum}) obtains only 14.83 points in ROUGE-L on Echo-XSum. By comparison, the same configuration achieves 30.24 points on XSum. Therefore, i) Echo-XSum is empirically more challenging than XSum, ii) a simple recursive-abstractive approach is not sufficient to obtain acceptable results on Echo-XSum, and, iii) different pretrained language models and different summarization datasets (from different genres/domains) do not significantly affect the results of a recursive-abstractive approach on our book summarization dataset.

4.2 Extractive-then-abstractive approaches

Since recursive-abstractive approaches yield unsatisfying results on Echo-XSum (see Table 2), we propose a simple, novel baseline based on the extractive-then-abstractive paradigm. Our model is composed of two submodules: the *extractor* extracts key sentences from the input text, while the *abstractor* uses the concatenation of these key sentences to generate an abstractive plot of the book. Given an input text $T = (s_1, s_2, \ldots, s_{|T|})$ where each s_i is a sentence, the extractor produces a score in $[0.0, 1.0]$ for each s_i, quantifying its degree of importance for the target summary. More formally:

$$\mathbf{e}_i^s = \textsc{SentenceEncoder}(s_i)$$
$$\textsc{Score}(s_i) = \sigma(W\mathbf{e}_i + \mathbf{b})$$

where $\mathbf{e}_i^s$ is the sentence representation of s_i from a $\textsc{SentenceEncoder}$.[14] Then, the abstractor takes the subset T^* composed of the k sentences with higher scores according to the extractor, and uses T^* to generate the final summary. To make the abstractor aware of the relative importance of each sentence, we multiply the embedding of each token by the score of its sentence, as follows:

$$\mathbf{e}_{i,j}^t = \textsc{Score}(s_i) \cdot \textsc{Embedding}(t_{i,j})$$

where $\mathbf{e}_{i,j}^t$ is the encoding of the j-th token of the i-th sentence, for each sentence in T^*.

The model is trained in an end-to-end fashion, i.e., the extractor and abstractor are trained jointly, by minimizing the cross-entropy loss between the reference summary and the generated summary.

Experimental setting. We follow the experimental setting we used for our recursive-abstractive approach. We train and evaluate 3 models – BART-large, LED-base, and LongT5-base – on Echo-XSum. Since pretraining on XSum results in

[13] We split Echo-Wiki and Echo-XSum into train/dev/test sets using the standard 80/10/10 split.

[14] We adopt a SentenceTransformer based on Distil-RoBERTa from https://www.sbert.net/.

Model	R-1	R-2	R-L	BERTScore
BART	16.64	4.07	13.09	0.517
LED	19.13	4.89	14.74	0.532
LongT5	**27.20**	**6.87**	**19.74**	**0.561**

Table 3: Automatic evaluation of extractive-then-abstractive approaches on Echo-Wiki.

	Model	Cons.	Fluency	Rel.	Coher.
recursive-abs.	$BART_{XSum}$	2.19	3.81	1.62	3.58
	LED_{XSum}	1.65	3.96	1.31	2.92
	$LongT5_{XSum}$	1.23	2.88	1.19	2.34
	$BART_{MediaSum}$	1.73	2.46	1.62	2.19
	$LED_{MediaSum}$	1.61	2.23	1.46	1.92
	$LongT5_{MediaSum}$	1.11	1.38	1.12	1.38
extractive-abs.	BART	1.69	4.38	1.76	4.42
	$BART_{XSum}$	1.61	3.06	1.35	2.71
	LED	1.84	4.34	1.84	4.23
	LED_{XSum}	1.72	3.97	1.55	3.66
	LongT5	**2.73**	**4.50**	**2.73**	**4.62**
	$LongT5_{XSum}$	2.04	3.85	1.74	3.52

Table 4: Human evaluation of recursive-abstractive approaches on Echo-XSum.

slightly improved performance for the recursive-abstractive approach, we also evaluate how pre-training on XSum affects the performance of our extractive-then-abstractive approach. Finally, we also train and evaluate our approach on Echo-Wiki and on BookSum (the latter to directly compare performance with the current state of the art).

Results. Table 2 (bottom) provides an overview of the results obtained by our extractive-then-abstractive approach on Echo-XSum. We can immediately notice that each configuration significantly outperforms the recursive-abstractive baselines by a large margin. For example, the best extractive-then-abstractive model ($BART_{XSum}$) improves over the best recursive-abstractive model (LED_{XSum}) by 11.90 points in ROUGE-L (26.73 vs. 14.83), and this is true for all the metrics we consider (ROUGE-1, ROUGE-2, ROUGE-L, and BERTScore). It is interesting to note that, while there is little difference in the results on Echo-XSum of different model configurations, there is a significant difference between BART, LED, and LongT5 when evaluated on Echo-Wiki, as shown in Table 3. We hypothesize that such a variance in performance is due to several factors, but the inadequacy of current non-semantic metrics plays a large role, as supported by our human evaluation (see Section 5).

Model	Cons.	Fluency	Rel.	Coher.
BART	2.06	**3.73**	1.65	**3.08**
LED	2.02	3.63	1.61	3.07
LongT5	**2.15**	3.62	**1.72**	3.06

Table 5: Human evaluation of extractive-then-abstractive approaches on Echo-Wiki.

Finally, we further assess the effectiveness of our extractive-then-abstractive approach on the standard test set of BookSum (Table 6). In particular, our approach outperforms the system of Kryscinski et al. (2021) using 33% of its parameters, and is competitive with the system of Wu et al. (2021) using only 0.1% of its parameters.

5 Analysis and Discussion

Human evaluation. Following common practice in the field of summarization, we set up a human evaluation process to assess the quality of the system-generated summaries. The annotation task, performed by an expert English speaker, consists of reading the source text and rating the summaries using a Likert scale for Consistency, Relevance, Fluency, and Coherence, as outlined in Fabbri et al. (2021). To make this experiment feasible in terms of time and resources, we focus our evaluation on fairy tales and short stories, which can be read by a human in a short time. Interestingly, but not surprisingly (Fabbri et al., 2021), the results of our human evaluation experiment tell a story that is different from ROUGE, as shown in Tables 4 and 5. However, the evaluation still highlights the effectiveness of our extractive-then-abstractive model compared to the recursive-abstractive baseline. It is clear, however, that future work should focus in particular on improving the Consistency and Relevance of the summaries generated.

Challenges. Echoes opens the door to several other analyses and experiments that were not possible with previous datasets. For example, we can leverage Echo-FairySum to perform an analysis of the behavior of the extractor submodule of our extractive-then-abstractive approach, as we show in Appendix D. In Section 3.4, we examined the different book genres in Echoes; LongT5 model performances are detailed for each genre in Figure 3. We notice that epic poems are the hardest to summarize in this setting, while our model performs reasonably well on fairy tales.

Approach	R-1	R-2	R-L	# Params.
Kryscinski et al. (2021)	39.87	8.01	13.99	737M
Wu et al. (2021)	43.19	10.63	17.10	175,000M
Ours (LED/extractive-abs.)	42.13	10.53	16.75	243M

Table 6: Results of our approach compared to the state of the art on the BookSum test set.

Language	# Examples	R-1	R-2	R-L	BERTScore
de	24	21.219	6.808	17.742	0.641
fr	33	21.602	7.681	17.721	0.622
es	45	24.509	8.966	19.554	0.634
it	37	25.174	10.446	22.343	0.633

Table 7: *Summarize-then-translate* experiment. We translate the summaries generated by LongT5$_{base}$ model, fine-tuned on Echo-XSum, and compare them against gold standard references.

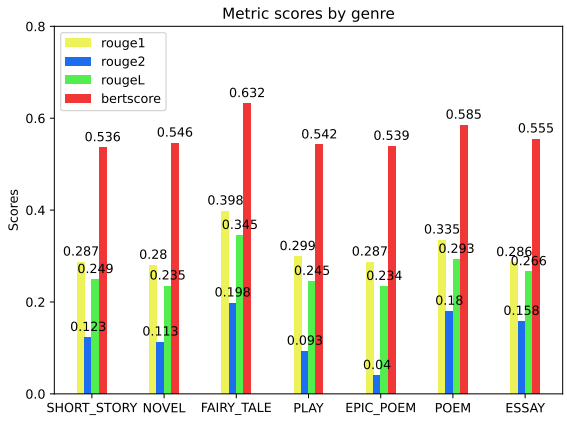

Figure 3: Genre-specific evaluation of LongT5$_{base}$ model fine-tuned on Echo-XSum. Best seen in color.

Cross-lingual book summarization. Additionally, Echoes can be employed as a multilingual and cross-lingual summarization benchmark, thanks to its coverage of 5 languages and 25 language pairs. In particular, we argue that cross-lingual book summarization is a very interesting challenge, as it requires a model to compress vast amounts of information while transferring knowledge across languages. Moreover, enabling cross-lingual book summarization is fundamental for all those cases in which we do not have the source text available in the language of interest, i.e., its translation may still be under copyright or may not exist at all. To move the first step in this direction, we propose a *summarize-then-translate* approach, a simple baseline for cross-lingual book summarization on Echo-XSum. As the name implies, our approach works by employing a monolingual model to produce a summary in the same language as the source text, and then it translates the summary from the source language to the desired target language. We report the results of this baseline in Table 7. While this is a strong baseline, it is still affected by two main issues: i) it requires two systems, a summarizer and a translator; ii) machine translation usually fails to translate language-specific items, e.g., character names may not be exact translations.

6 Conclusion

In this paper, we introduced Echoes, the first multilingual resource for book summarization and the largest among the English datasets. Echoes features three novel datasets, namely, Echo-Wiki, Echo-XSum, and Echo-FairySum, which address several limitations of existing book summarization resources, such as BookSum. Indeed, previous datasets for full-text book summarization are, i) limited in size, and, ii) monolingual, i.e., usually covering English only.

In addition, we leveraged Echoes to bring to light the unsatisfying capabilities of current approaches to generalize to book summarization. Finally, to mitigate this issue, we proposed a new *extractive-then-abstractive* baseline for book summarization, which outperforms its purely-abstractive counterpart on Echo-Wiki and Echo-XSum, achieving results on the standard BookSum test set that are comparable with the current state of the art while using a number of parameters that is only 0.1% compared to the best-performing method.

We believe that Echoes will foster future work on long-document summarization, especially in the multilingual and cross-lingual setting.

Limitations

Despite the multilinguality of our resource, there is still a strong bias towards the English language, as the majority of books are in English and many translations are from English. This may result in the values of English literature being reflected, and these may differ from those of other cultures; summarizing literature from different cultures and regions may not be fully accurate, as every region has had its own historical development.

Language models used in the experiments can inherit biases from the training data and the tools, such as the ones used for preprocessing, and have limitations that have not been fully evaluated and could impact the results of this study.

This study includes the use of Web data, which – while marked as public domain – may be subject to copyright laws. The data used in this study was collected for research purposes and was not intended for any other use. Additionally, it is worth noting that the majority of books used in our resource are copyright-free, and therefore, old. While this allowed us to include a large number of texts in our dataset, it also means that our resource may not fully capture contemporary literature and may not be representative of current linguistic trends and cultural values.

Acknowledgements

The authors gratefully acknowledge the support of the ERC Consolidator Grant MOUSSE No. 726487 under the European Union's Horizon 2020 research.

The last author gratefully acknowledges the support of the PNRR MUR project PE0000013-FAIR. This work was carried out while Alessandro Scirè was enrolled in the Italian National Doctorate on Artificial Intelligence run by Sapienza University of Rome. We would like to express our gratitude to Luigi Procopio and Edoardo Barba for their valuable insights on extractive-then-abstractive architectures, as well as to Fabrizio Brignone (Babelscape) for his exceptional support with the adaptation and use of Babelscape's keyword and phrase annotation interface.

References

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

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

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.

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. 2018. 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.

Alexios Gidiotis and Grigorios Tsoumakas. 2020. A divide-and-conquer approach to the summarization of long documents.

Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 2022. LongT5: Efficient text-to-text transformer for long sequences. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 724–736, Seattle, United States. Association for Computational Linguistics.

Vivek Gupta, Prerna Bharti, Pegah Nokhiz, and Harish Karnick. 2021. SumPubMed: Summarization dataset of PubMed scientific articles. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 292–303, Online. Association for Computational Linguistics.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2017. The narrativeqa reading comprehension challenge.

Wojciech Kryscinski, Nazneen Fatema Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir R. Radev. 2021. Booksum: A collection of datasets for long-form narrative summarization. *CoRR*, abs/2105.08209.

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.

Haoran Li, Arash Einolghozati, Srinivasan Iyer, Bhargavi Paranjape, Yashar Mehdad, Sonal Gupta, and Marjan Ghazvininejad. 2021. EASE: Extractive-abstractive summarization end-to-end using the information bottleneck principle. In *Proceedings of the Third Workshop on New Frontiers in Summarization*, pages 85–95, Online and in Dominican Republic. Association for Computational Linguistics.

Rada Mihalcea and Hakan Ceylan. 2007. Explorations in automatic book summarization. pages 380–389.

Ramesh Nallapati, Bowen Zhou, Cicero Nogueira dos santos, Caglar Gulcehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond.

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.

Jeff Wu, Long Ouyang, Daniel M. Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. 2021. Recursively summarizing books with human feedback.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert.

Yusen Zhang, Ansong Ni, Ziming Mao, Chen Henry Wu, Chenguang Zhu, Budhaditya Deb, Ahmed H. Awadallah, Dragomir Radev, and Rui Zhang. 2021. Summ^n: A multi-stage summarization framework for long input dialogues and documents.

Chenguang Zhu, Yang Liu, Jie Mei, and Michael Zeng. 2021. MediaSum: A large-scale media interview dataset for dialogue summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5927–5934, Online. Association for Computational Linguistics.

Chenguang Zhu, Ziyi Yang, Robert Gmyr, Michael Zeng, and Xuedong Huang. 2019. Leveraging lead bias for zero-shot abstractive news summarization.

A Wikipedia summary sections

In Table 8 we provide the list of Wikipedia section titles whose contents are used as summaries in Echo-Wiki.

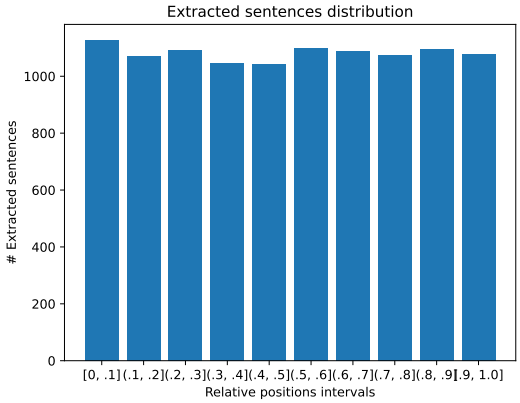

Figure 4: Number of extracted sentences for each relative position interval.

B Echo-XSum example

In Figure 5 we report an excerpt of the book text of the English version of "The Metamorphosis" by Franz Kafka, along with the multilingual extreme summaries from Echo-XSum.

C Echo-XSum annotation task

In Figure 6 we provide an example of a manually-annotated summary in Echo-XSum. The annotator was tasked to highlight portions of text containing information related to the plot from the Wikipedia introduction.

D Extractor analysis

We analyze the positions of the sentences selected by the extractor. This analysis is required to investigate the presence of any positional bias, e.g., the lead bias, which is known to affect systems trained on news stories. Figure 4 depicts the distribution of the relative positions of the extracted sentences on texts from Echo-FairySum, i.e., fairy tales and short stories. We deduce that the extractions are not affected by any bias. Thanks to Echo-FairySum extractive annotations, we are also able to evaluate the performance of the extractor component of the *extractive-then-abstractive* approaches. We aggregate multiple extractive annotations in Echo-FairySum by retaining the intersecting sentences; we refer to these sentences as the gold sentences. We measure the Extractor performance by computing the overlap between the sentences extracted by the model and the gold ones. We compute the *Precision@K* by comparing the topK-ranked sentences with the references. We report the Extractor

IT	EN	ES	FR	DE
trame	plot overview	resumen de la trama	trame	zusammenfassung
trama	subject	trama	résumé synthétique	synthese
trama del racconto	plots	argumento	résumé	handlung
sinossi	plot details	contenido	trame romanesque	inhalt
vicenda	structure and plot	resumen	synopsis	
riassunto	plot and structure	sinopsis	la trame romanesque	
racconto	abstracts		la trame de l'histoire	
il racconto	plot summary			
riassunti	synopsis			
	subjects			
	plot			
	story			
	summaries			
	abstract			
	the story			
	plot synopsis			
	plot introduction			
	summary			
	thematic summary			
	summary and themes			
	plot outline			

Table 8: Table of Wikipedia section titles utilized in the Echo-Wiki parsing process in multiple languages

performance in Table 9. We observe relatively low scores, meaning that the extractor is only partially able to discriminate relevant sentences from irrelevant ones. This aspect confirms that there is still large room for improving the Extractor and, consequently, the relevance of the summaries.

K	Precision
1	31.1
2	28.8
3	28.8
4	27.2
5	25.6

Table 9: Extractor evaluation: Precision@K

Book: The Metamorphosis

Text (EN): One morning, when Gregor Samsa woke from troubled dreams, he found himself transformed in his bed into a horrible vermin...*(21,897 words omitted)*

Summary (EN): Metamorphosis tells the story of salesman Gregor Samsa , who wakes one morning to find himself inexplicably transformed into a huge insect and subsequently struggles to adjust to this new condition.

Summary (DE): Die Verwandlung handelt von Gregor Samsa , dessen plötzliche Verwandlung in ein „ Ungeziefer “ die Kommunikation seines sozialen Umfelds mit ihm immer mehr hemmt , bis er von seiner Familie für untragbar gehalten wird und schließlich zugrunde geht .

Summary (FR): La Métamorphose décrit la métamorphose et les mésaventures de un représentant de commerce qui se réveille un matin transformé en un un « monstrueux insecte » . À partir de cette situation absurde , l' auteur présente une critique sociale , aux multiples lectures possibles , en mêlant thématiques économiques et sociétales et questionnements sur l' individu , le déclassement , la dépendance , la solidarité familiale , la solitude et la mort .

Summary (ES): La historia trata sobre Gregorio Samsa , cuya repentina transformación en un enorme insecto dificulta cada vez más la comunicación de su entorno social con él , hasta que es considerado intolerable por su familia y finalmente perece .

Summary (IT): All' inizio del racconto , il protagonista Gregor Samsa si risveglia una mattina ritrovandosi trasformato in un enorme insetto. La causa di tale mutazione non viene mai rivelata . Tutto il seguito del racconto narra dei tentativi compiuti dal giovane Gregor per cercar di regolare - per quanto possibile - la propria vita a questa sua nuova particolarissima condizione mai vista prima , soprattutto nei riguardi dei genitori e della sorella e con il suo datore di lavoro.

Figure 5: An excerpt of a book text along with multilingual summaries from Echo-XSum.

Iliad

The Iliad (;"Iliad". Random House Webster's Unabridged Dictionary. , ; sometimes referred to as the Song of Ilion or Song of Ilium) is an ancient Greek epic poem in dactylic hexameter, traditionally attributed to Homer. Usually considered to have been written down circa the 8th century BC, the Iliad is among the oldest extant works of Western literature, along with the Odyssey, another epic poem attributed to Homer, which tells of Odysseus's experiences after the events of the Iliad.Vidal-Naquet, Pierre. Le monde d'Homère (The World of Homer), Perrin (2000), p. 19 In the modern vulgate (the standard accepted version), the Iliad contains 15,693 lines, divided into 24 books; it is written in Homeric Greek, a literary amalgam of Ionic Greek and other dialects. It is usually grouped in the Epic Cycle. Set during the Trojan War, the ten-year siege of the city of Troy (Ilium) by a coalition of Mycenaean Greek states (Achaeans), it tells of the battles and events during the weeks of a quarrel between King Agamemnon and the warrior Achilles. Although the story covers only a few weeks in the final year of the war, the Iliad mentions or alludes to many of the Greek legends about the siege; the earlier events, such as the gathering of warriors for the siege, the cause of the war, and related concerns, tend to appear near the beginning. Then the epic narrative takes up events prophesied for the future, such as Achilles's imminent death and the fall of Troy, although the narrative ends before these events take place. However, as these events are prefigured and alluded to more and more vividly, when it reaches an end, the poem has told a more or less complete tale of the Trojan War.

Figure 6: Echo-XSum annotation process consists of highlighting plot-specific pieces of text from the lead section of the Wikipedia page.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
7, right after Conclusion.

☑ A2. Did you discuss any potential risks of your work?
7

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

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

B ☑ Did you use or create scientific artifacts?

1,3,4,5.

☑ B1. Did you cite the creators of artifacts you used?
1,2,3,4,5

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
3,Limitations

☒ 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)?
All artifacts have been used according to their original purpose.

☐ 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.?
1,2,3,4,5,Limitations

☑ 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.
2,3,4

C ☑ Did you run computational experiments?

4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
4

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?
We do not perform hyperparameter tuning

☒ 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?
Experiments are computational expensive, so we were able to afford just one run per configuration.

☑ 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.)?
ROUGE:4

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

3,5

☒ 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.?
We provide a short description of the guidelines and pointers to existing guidelines.

☑ 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)?
We report information about the students in Section 3. The expert annotators prefer not to disclose their information.

☐ 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?
Not applicable. Left blank.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Our research group does not have an ethics review board.

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
The annotators prefer not to disclose their information.