LoRaLay: A Multilingual and Multimodal Dataset for Long Range and Layout-Aware Summarization

Laura Nguyen^{1,3}

Thomas Scialom^{2*} Benjamin Piwowarski³ Jacopo Staiano^{4*}

¹reciTAL, Paris, France ²Meta AI, Paris, France ³Sorbonne Université, CNRS, ISIR, F-75005 Paris, France

4University of Trento Italy

⁴University of Trento, Italy

laura@recital.ai tscialom@fb.com benjamin.piwowarski@cnrs.fr jacopo.staiano@unitn.it

Abstract

Text Summarization is a popular task and an active area of research for the Natural Language Processing community. It requires accounting for long input texts, a characteristic which poses computational challenges for neural models. Moreover, real-world documents come in a variety of complex, visually-rich, layouts. This information is of great relevance, whether to highlight salient content or to encode long-range interactions between textual passages. Yet, all publicly available summarization datasets only provide plain text content. To facilitate research on how to exploit visual/layout information to better capture longrange dependencies in summarization models, we present *LoRaLay*, a collection of datasets for long-range summarization with accompanying visual/layout information. We extend existing and popular English datasets (arXiv and PubMed) with visual/layout information and propose four novel datasets – consistently built from scholar resources - covering French, Spanish, Portuguese, and Korean languages. Further, we propose new baselines merging layout-aware and long-range models - two orthogonal approaches - and obtain state-of-theart results, showing the importance of combining both lines of research.

1 Introduction

Deep learning techniques have enabled remarkable progress in Natural Language Processing (NLP) in recent years (Devlin et al., 2018; Raffel et al., 2019; Brown et al., 2020). However, the majority of models, benchmarks, and tasks have been designed for unimodal approaches, i.e. focusing exclusively on a single source of information, namely plain text. While it can be argued that for specific NLP tasks, such as textual entailment or machine translation, plain text is all that is needed, there exist several tasks for which disregarding the visual appearance of text is clearly sub-optimal: in a real-world context (business documentation, scientific articles, etc.), text does not naturally come as a sequence of characters, but is rather displayed in a bi-dimensional space containing rich visual information. The layout of e.g. this very paper provides valuable semantics to the reader: in which section are we right now? At the blink of an eye, this information is readily accessible via the salient section title (formatted differently and placed to highlight its role) preceding these words. Just to emphasize this point, *imagine having to scroll this content in plain text to access such information*.

In the last couple of years, the research community has shown a growing interest in addressing these limitations. Several approaches have been proposed to deal with visually-rich documents and integrate layout information into language models, with direct applications to Document Understanding tasks. Joint multi-modal pretraining (Xu et al., 2021; Powalski et al., 2021; Appalaraju et al., 2021) has been key to reach state-of-the-art performance on several benchmarks (Jaume et al., 2019; Graliński et al., 2020; Mathew et al., 2021). Nonetheless, a remaining limitation is that these (transformer-based) approaches are not suitable for processing long documents, the quadratic complexity of self-attention constraining their use to short sequences. Such models are hence unable to encode global context (e.g. long-range dependencies among text blocks).

Focusing on compressing the most relevant information from long texts to short summaries, the Text Summarization task naturally lends itself to benefit from such global context. Notice that, in practice, the limitations linked to sequence length are also amplified by the lack of visual/layout information in the existing datasets. Therefore, in this work, we aim at spurring further research on how to incorporate multimodal information to better capture long-range dependencies.

Our contributions can be summarized as follows:

^{*}Work partially done while at reciTAL.

- We extend two popular datasets for long-range summarization, arXiv and PubMed (Cohan et al., 2018), by including visual and layout information – thus allowing direct comparison with previous works;
- We release 4 additional layout-aware summarization datasets (128K documents), covering French, Spanish, Portuguese, and Korean languages;
- We provide baselines including adapted architectures for multi-modal long-range summarization, and report results showing that (1) performance is far from being optimal; and (2) layout provides valuable information.

All the datasets are available on HuggingFace.¹

2 Related Work

2.1 Layout/Visually-rich Datasets

Document Understanding covers problems that involve reading and interpreting visually-rich documents (in contrast to plain texts), requiring comprehending the conveyed multimodal information. Hence, several tasks with a central layout aspect have been proposed by the document understanding community. Key Information Extraction tasks consist in extracting the values of a given set of keys, e.g., the total amount in a receipt or the date in a form. In such tasks, documents have a layout structure that is crucial for their interpretation. Notable datasets include FUNSD (Jaume et al., 2019) for form understanding in scanned documents, and SROIE (Huang et al., 2019), as well as CORD (Park et al., 2019), for information extraction from receipts. Graliński et al. (2020) elicit progress on deeper and more complex Key Information Extraction by introducing the Kleister datasets, a collection of business documents with varying lengths, released as PDF files. However, the documents in Kleister often contain single-column layouts, which are simpler than the various multi-column layouts considered in LoRaLay. Document VQA is another popular document understanding task that requires processing multimodal information (e.g., text, layout, font style, images) conveyed by a document to be able to answer questions about a

visually rich document (e.g., What is the date given at the top left of the form?, Whose picture is given in this figure?). The DocVQA dataset (Mathew et al., 2021) and InfographicsVQA (Mathew et al., 2022) are commonly-used VQA datasets that respectively provide industry documents and infographic images, encouraging research on understanding documents with complex interplay of text, layout and graphical elements. Finally, to foster research on visually-rich document understanding, Borchmann et al. (2021) introduce the Document Understanding Evaluation (DUE) benchmark, a unified benchmark for end-to-end document understanding, created by combining several datasets. DUE includes several available and transformed datasets for VQA, Key Information Extraction and Machine Reading Comprehension tasks.

2.2 Existing Summarization Datasets

Several large-scale summarization datasets have been proposed to boost research on text summarization systems. Hermann et al. (2015) proposed the CNN/DailyMail dataset, a collection of English articles extracted from the CNN and The Daily Mail portals. Each news article is associated with multi-sentence highlights which serve as reference summaries. Scialom et al. (2020) bridge the gap between English and non-English resources for text summarization by introducing MLSum, a largescale multilingual summarization corpus providing news articles written in French, German, Spanish, Turkish and Russian. Going toward more challenging scenarios involving significantly longer documents, the arXiv and PubMed datasets (Cohan et al., 2018) consist of scientific articles collected from academic repositories, wherein the paper abstracts are used as summaries. To encourage a shift towards building more abstractive summarization models with global content understanding, Sharma et al. (2019) introduce BIGPATENT, a large-scale dataset made of U.S. patent filings. Here, invention descriptions serve as reference summaries.

The vast majority of summarization datasets only deal with plain text documents. As opposed to other Document Understanding tasks (e.g., form understanding, visual QA) in which the placement of text on the page and/or visual components are the main source of information needed to find the desired data (Borchmann et al., 2021), text plays a predominant role in document summarization. However, guidelines for summarizing texts – espe-

I https://hf.co/datasets/nglaura/arxivlay-summarization, https://hf.co/datasets/nglaura/pubmedlay-summarization, https://hf.co/datasets/nglaura/hal-summarization, https://hf.co/datasets/nglaura/scielo-summarization, https://hf.co/datasets/nglaura/koreascience-summarization

cially long ones – often recommend roughly previewing them to break them down into their major sections (Toprak and Almacioğlu, 2009; Luo et al., 2019). This suggests that NLP systems might leverage multimodal information in documents. Miculicich and Han (2022) propose a two-stage method which detects text segments and incorporates this information in an extractive summarization model. Cao and Wang (2022) collect a new dataset for long and structure-aware document summarization, consisting of 21k documents written in English and extracted from WikiProject Biography.

Although not all documents are explicitly organized into clearly defined sections, the great majority contains layout and visual clues (e.g., a physical organization into paragraphs, bigger headings/subheadings) which help structure their textual contents and facilitate reading. Thus, we argue that layout is crucial to summarize long documents. We propose a corpus of more than 345K long documents with layout information. Furthermore, to address the need for multilingual training data (Chi et al., 2020), we include not only English documents, but also French, Spanish, Portuguese and Korean ones.

3 Datasets Construction

Inspired by the way the arXiv and PubMed datasets were built (Cohan et al., 2018), we construct our corpus from research papers, with abstracts as ground-truth summaries. As the PDF format allows simultaneous access to textual, visual and layout information, we collect PDF files to construct our datasets, and provide their URLs.²

For each language, we select a repository that contains a high number of academic articles (in the order of hundreds of thousands) and provides easy access to abstracts. More precisely, we chose the following repositories:

- Archives Ouverte HAL (French),³ an open archive of scholarly documents from all academic fields. As HAL is primarily directed towards French academics, a great proportion of articles are written in French;
- SciELO (Spanish and Portuguese),⁴ an open access database of academic articles published in journal collections from Latin America,

Iberian Peninsula and South Africa, and covering a broad range of topics (e.g. agricultural sciences, engineering, health sciences, letters and arts). Languages include English, Spanish, and Portuguese.

• KoreaScience (Korean),⁵ an open archive of Korean scholarly publications in the fields of natural sciences, life sciences, engineering, and humanities and social sciences. Articles are written in English or Korean.

Further, we provide enhanced versions of the arXiv and PubMed datasets, respectively denoted as arXiv-Lay and PubMed-Lay, for which layout information is provided.

3.1 Collecting the Data

Extended Datasets The arXiv and PubMed datasets (Cohan et al., 2018) contain long scientific research papers extracted from the arXiv and PubMed repositories. We augment them by providing their PDFs, allowing access to layout and visual information. As the abstracts contained in the original datasets are all lowercased, we do not reuse them, but rather extract the raw abstracts using the corresponding APIs.

Note that we were unable to retrieve all the original documents. For the most part, we failed to retrieve the corresponding abstracts, as they did not necessarily match the ones contained in the PDF files (due to *e.g.* PDF-parsing errors). We also found that some PDF files were unavailable, while others were corrupted or scanned documents.⁶ In total, about 39% (35%) of the original documents in arXiv (PubMed) were lost.

arXiv-Lay The original arXiv dataset (Cohan et al., 2018) was constructed by converting the LAT_EX files to plain text. To be consistent with the other datasets – for which LAT_EX files are not available – we instead use the PDF files to extract both text and layout elements. For each document contained in the original dataset, we fetch (when possible) the corresponding PDF file using Google Cloud Storage buckets. As opposed to the original procedure, we do not remove tables nor discard sections that follow the conclusion. We retrieve the corresponding abstracts from a metadata file provided by Kaggle.⁷

 $^{^2}We$ make the corpus-construction code publicly available at $\tt https://github.com/recitalAI/loralay-datasets.$

³https://hal.archives-ouvertes.fr/

⁴https://www.scielo.org/

⁵http://www.koreascience.or.kr

⁶For more details on this, see Section A.1 in the Appendix.

⁷https://www.kaggle.com/Cornell-University/arxiv

PubMed-Lay For PubMed, we use the PMC OAI Service⁸ to retrieve abstracts and PDF files.

HAL We use the HAL API⁹ to download research papers written in French. To avoid excessively long (e.g. theses) or short (e.g. posters) documents, extraction is restricted to journal and conference papers.

SciELO Using Scrapy,¹⁰ we crawl the following SciELO collections: Ecuador, Colombia, Paraguay, Uruguay, Bolivia, Peru, Portugal, Spain and Brazil. We download documents written either in Spanish or Portuguese, according to the metadata, obtaining two distinct datasets: SciELO-ES (Spanish) and SciELO-PT (Portuguese).

KoreaScience Similarly, we scrape the Korea-Science website to extract research papers. We limit search results to documents whose publishers' names contain the word *Korean*. This rule was designed after sampling documents in the repository, and is the simplest way to get a good proportion of papers written in Korean.¹¹ Further, search is restricted to papers published between 2012 and 2021, as recent publications are more likely to have digital-born, searchable PDFs. Finally, we download the PDF files of documents that contain an abstract written in Korean.

3.2 Data Pre-processing

For each corpus, we use the 95th percentile of the page distribution as an upper bound to filter out documents with too many pages, while the 5th (1st for HAL and SciELO) percentile of the summary length distribution is used as a minimum threshold to remove documents whose abstracts are too short. As our baselines do not consider visual information, we only extract text and layout from the PDF files. Layout is incorporated by providing the spatial position of each word in a document page image, represented by its bounding box (x_0, y_0, x_1, y_1) , where (x_0, y_0) and (x_1, y_1) respectively denote the coordinates of the top-left and bottom-right corners. Using the PDF rendering library Poppler¹², text and word bounding boxes are extracted from each PDF, and the sequence order is recovered based on heuristics around the document layout (e.g., tables, columns). Abstracts are then

9 https://api.archives-ouvertes.fr/docs/search 10,... removed by searching for exact matches; when no exact match is found, we use fuzzysearch¹³ and regex¹⁴ to find near matches.¹⁵ For the non-English datasets, documents might contain several abstracts, written in different languages. To avoid information leakage, we retrieve the abstract of each document in every language available – according to the API for HAL or the websites for SciELO and KoreaScience – and remove them using the same strategy as for the main language. In the case an abstract cannot be found, we discard the document to prevent any unforeseen leakage. The dataset construction process is illustrated in Section A in the Appendix.

3.3 Datasets Statistics

The statistics of our proposed datasets, along with those computed on existing summarization datasets of long documents (Cohan et al., 2018; Sharma et al., 2019) are reported in Table 1. We see that document lengths are comparable or greater than for the arXiv, PubMed and BigPatent datasets.

For arXiv-Lay and PubMed-Lay, we retain the original train/validation/splits and try to reconstruct them as faithfully to the originals as possible. For the new datasets, we order documents based on their publication dates and provide splits following a chronological ordering. For HAL and Korea-Science, we retain 3% of the articles as validation data, 3% as test, and the remaining as training data. To match the number of validation/test documents in HAL and KoreaScience, we split the data into 90% for training, 5% for validation and 5% for test, for both SciELO datasets.

4 Experiments

4.1 Models

For reproducibility purposes, we make the models' implementation, along with the fine-tuning and evaluation scripts, publicly available.¹⁶

We do not explore the use of visual information in long document summarization, as the focus is on evaluating baseline performance using state-of-theart summarization models augmented with layout information. While visual features might provide a better understanding of structures such as tables and figures, we do not expect substantial gains with

⁸ https://www.ncbi.nlm.nih.gov/pmc/tools/oai/

¹⁰ https://scrapy.org/

¹¹For further details, see Section A.2 in the Appendix.

¹²https://poppler.freedesktop.org/

¹³https://pypi.org/project/fuzzysearch/

¹⁴ https://pypi.org/project/regex/

¹⁵We use a maximum Levenshtein distance of 20 with fuzzysearch, and a maximum number of errors of 3 with regex.

¹⁶https://github.com/recitalAI/loralay-modeling

Dataset	# Docs	Mean Article Length	Mean Summary Length
arXiv (Cohan et al., 2018)	215,913	3,016	203
PubMed (Cohan et al., 2018)	133,215	4,938	220
BigPatent (Sharma et al., 2019)	1,341,362	3,572	117
arXiv-Lay	130,919	7,084	125
PubMed-Lay	86,668	4,038	144
HAL	46,148	4,543	134
SciELO-ES	23,170	4,977	172
SciELO-PT	21,563	6,853	162
KoreaScience	37,498	3,192	95

Table 1: Datasets statistics. Article and summary lengths are computed in words. For KoreaScience, words are obtained via white-space tokenization. Difference between arXiv and arXiv-Lay is due to the fact that we retain the whole document, while Cohan et al. (2018) truncate it after the conclusion.

respect to layout-aware models. Indeed, the information provided in figures (i.e., information that cannot be captured by layout or text) are commonly described in the caption or related paragraphs.

Text-only models with standard input size We use Pegasus (Zhang et al., 2020) as a text-only baseline for arXiv-Lay and PubMed-Lay. Pegasus is an encoder-decoder model pre-trained using gapsentences generation, making it a state-of-the-art model for abstractive summarization. For the non-English datasets, we rely on a finetuned MBART as our baseline. MBART (Liu et al., 2020) is a multilingual sequence-to-sequence model pretrained on large-scale monolingual corpora in many languages using the BART objective (Lewis et al., 2019). We use its extension, MBART-50 (Tang et al., 2020),¹⁷ which is created from the original MBART by extending its embeddings layers and pre-training it on a total of 50 languages. Both Pegasus and MBART are limited to a maximum sequence length of 1,024 tokens, which is well below the median length of each dataset.

Layout-aware models with standard input size We introduce layout-aware extensions of Pegasus and MBART, respectively denoted as Pegasus+Layout and MBART+Layout. Following LayoutLM (Xu et al., 2020), which is state-ofthe-art on several document understanding tasks (Jaume et al., 2019; Huang et al., 2019; Harley et al., 2015), each token bounding box coordinates (x_0, y_0, x_1, y_1) is normalized into an integer in the range [0, 1000]. Spatial positions are encoded using four embedding tables, namely two for the coordinate axes (x and y), and the other two for the bounding box size (width and height). The layout representation of a token is formed by summing the resulting embedding representations The final representation of a token is then obtained through point-wise summation of its textual, 1D-positional and layout embeddings.

Long-range, text-only models To process longer sequences, we leverage BigBird (Zaheer et al., 2020), a sparse-attention based Transformer which reduces the quadratic dependency to a linear one. For arXiv-Lay and PubMed-Lay, we initialize Big-Bird from Pegasus (Zaheer et al., 2020) and for the non-English datasets, we use the weights of MBART. The resulting models are referred to as BigBird-Pegasus and BigBird-MBART. For both models, BigBird sparse attention is used only in the encoder. Both models can handle up to 4,096 inputs tokens, which is greater than the median length in PubMed-Lay, HAL and KoreaScience.

Long-range, layout-aware models We also include layout information in long-range text-only models. Similarly to layout-aware models with standard input size, we integrate layout information into our long-range models by encoding each token's spatial position in the page. The resulting models are denoted as BigBird-Pegasus+Layout and BigBird-MBART+Layout.

Additional State-of-the-Art Baselines We further consider additional state-of-the-art baselines for summarization: i) the text-only T5 (Raffel et al., 2019) with standard input size, ii) the long-range Longformer-Encoder-Decoder (LED) (Beltagy et al., 2020), and iii) the layout-aware, long-range LED+Layout, which we implement similarly to the previous layout-aware models.

4.2 Implementation Details

We initialize our Pegasus-based and MBART-based models with, respectively, the google/pegasus-large and facebook/mbart-large-50 checkpoints shared through the Hugging Face Model Hub. As for T5 and LED, we use the weights from t5-base and allenai/led-base-16384, respectively.¹⁸

Following Zhang et al. (2020) and Zaheer et al. (2020), we fine-tune our models up to 74k (100k) steps on arXiv-Lay (PubMed-Lay). On HAL, the total number of steps is set to 100k, while it is de-

¹⁷For the sake of clarity, we refer to MBART-50 as MBART.

¹⁸The large versions of T5 and LED did not fit into GPU due to their size.

Dataset	Iı	Instances			Length	Output	Output Length	
Dataset	Train	Dev	Test	Median	90%-ile	Median	90 <i>%</i> -ile	
arXiv (Cohan et al., 2018)	203,037	6,436	6,440	6,151	14,405	171	352	
PubMed (Cohan et al., 2018)	119,924	6,633	6,658	2,715	6,101	212	318	
arXiv-Lay	122,189	4,374	4,356	6,225	12,541	150	249	
PubMed-Lay	78,234	4,084	4,350	3,761	7,109	182	296	
HAL	43,379	1,384	1,385	4,074	8,761	179	351	
SciELO-ES	20,853	1,158	1,159	4,859	8,519	226	382	
SciELO-PT	19,407	1,078	1,078	6,090	9,655	239	374	
KoreaScience	35,248	1,125	1,125	2,916	5,094	219	340	

Table 2: Datasets splits and statistics. Input and output lengths are computed in tokens, obtained using Pegasus and MBART-50's tokenizers for the English and non-English datasets, respectively.

creased to 50k for the other non-English datasets.¹⁹ For each model, we select the checkpoint with the best validation loss. For Pegasus and MBART models, inputs are truncated at 1,024 tokens. For BigBird-Pegasus models, we follow Zaheer et al. (2020) and set the maximum input length at 3,072 tokens. As the median input length is much greater in almost every non-English dataset, we increase the maximum input length to 4,096 tokens for BigBird-MBART models. Output length is restricted to 256 tokens for all models, which is enough to fully capture at least 50% of the summaries in each dataset.

For evaluation, we use beam search and report a single run for each model and dataset. Following Zhang et al. (2020); Zaheer et al. (2020), we set the number of beams to 8 for Pegasus-based models, and 5 for BigBird-Pegasus-based models. For the non-English datasets, we set it to 5 for all models, for fair comparison. For all experiments, we use a length penalty of 0.8. For more implementation details, see Section B.1 in the Appendix.

5 Results and Discussion

5.1 General Results

In Table 3, we report the ROUGE-L scores obtained on arXiv and PubMed datasets (reported by Zaheer et al. (2020)), as well as on the corresponding layout-augmented counterparts we release. ²⁰ On arXiv-Lay and PubMed-Lay, we observe that, while the addition of layout to Pegasus does not improve the ROUGE-L scores, there are gains in integrating layout information into BigBird-Pegasus. To assess whether these gains are significant, we perform significance analysis at the 0.05 level using bootstrap, and estimate a ROUGE-L threshold that predicts when improvements are significant. ROUGE-L improvements between each pair of models are reported in Table 11 in the appendix. On arXiv-Lay, we compute a threshold of 1.48 ROUGE-L, showing that BigBird-Pegasus+Layout significantly outperforms all Pegasus-based models. In particular, we find a 1.56 ROUGE-L improvement between BigBird-Pegasus and its layoutaugmented counterpart, demonstrating that the addition of layout to long-range modeling significantly improves summarization. On PubMed-Lay, we compute a threshold of 1.77. Hence, the 0.96 ROUGE-L improvement from BigBird-Pegasus to its layout-augmented counterpart is not significant. However, the variance in font sizes in PubMed-Lay is much smaller compared to arXiv-Lay (see Table 12 in the appendix), reflecting an overall more simplistic layout. Therefore, we argue that layout integration has a lesser impact in PubMed-Lay, which can explain the non-significance of results. In addition, we find that BigBird-Pegasus significantly outperforms Pegasus and Pegasus+Layout only when augmented with layout, with an improvement of, respectively, 2.3 and 2.2 points. This demonstrates the importance of combining layout and long-range modeling.

While T5 and LED obtain competitive results, we find that the gain in adding layout to LED is minor. However, the models we consider have all been pre-trained only on plain text. As a result, the layout representations are learnt from scratch during fine-tuning. Similarly to us, Borchmann et al. (2021) show that their layout-augmented T5 does not necessarily improve the scores, and that performance is significantly enhanced only when the model has been pre-trained on layout-rich data.

Further, we observe, for both Pegasus and BigBird-Pegasus, a drop in performance w.r.t. the scores obtained on the original datasets. This can be explained by two factors. First, our extended

 $^{^{19}}$ We tested different values for the number of steps (10k, 25k, 50k, 100k) and chose the one that gave the best validation scores for MBART.

²⁰For detailed results, please refer to Section C.1 in the Appendix.

Model	# Params	arXiv/ arXiv-Lay	PubMed/ PubMed-Lay
Pegasus (Zhang et al., 2020)	568M	38.83	41.34
BigBird-Pegasus (Zaheer et al., 2020)	576M	41.77	42.33
T5 (Raffel et al., 2019)	223M	37.90	39.23
LED (Beltagy et al., 2020)	161M	40.74	41.54
LED+Layout	165M	40.96	41.83
Pegasus	568M	39.07	39.75
Pegasus+Layout	572M	39.25	39.85
BigBird-Pegasus	576M	39.59	41.09
BigBird-Pegasus+Layout	581M	41.15	42.05

Table 3: ROUGE-L scores on arXiv-Lay and PubMed-Lay. Reported results obtained by Pegasus and BigBird-Pegasus on the original arXiv and PubMed are reported with a gray background. The best results obtained on arXiv-Lay and PubMed-Lay are denoted in bold.

Model	# Params	HAL (fr)	SciELO-ES (es)	SciELO-PT (pt)	KoreaScience (ko)
MBART	610M	42.00	36.55	36.42	16.94
MBART+Layout	615M	41.67	37.47	34.37	14.98
BigBird-MBART	617M	45.04	37.76	39.63	18.55
BigBird-MBART+Layout	621M	45.20	40.71	40.51	19.95

Table 4: ROUGE-L scores on the non-English datasets. The best results for each dataset are reported in bold.

Dataset	Train	Validation	Test
HAL (fr)	90.72	90.54	85.84
SciELO-ES (es)	84.86	84.28	84.90
SciELO-PT (pt)	90.95	90.58	91.96
KoreaScience (ko)	73.53	70.26	68.78

Table 5: Percent confidence obtained for the main language, for each dataset split.

datasets contain less training data due to the inability to process all original documents. Secondly, the settings are different: while the original arXiv and PubMed datasets contain clear discourse information (e.g., each section is delimited by markers) obtained from LATEX files, documents in our extended versions are built by parsing raw PDF files. Therefore, the task is more challenging for text-only baselines, as they have no access to the discourse structure of documents, which further underlines the importance of taking the structural information, brought by visual cues, into account.

Table 4 presents the ROUGE-L scores reported on the non-English datasets. On HAL, we note that BigBird-MBART does not benefit from layout. After investigation, we hypothesize that this is due to the larger presence of single-column and simple layouts, which makes layout integration less needed. On both SciELO datasets, we notice that combining layout with long-range modeling brings substantial improvements over MBART. Further, we find that the plain-text BigBird models do not improve over the layout-aware Pegasus and MBART on arXiv-Lay and SciELO-ES, demonstrating that simply capturing more context does not always suffice. Regarding performance on KoreaScience, we can see a significant drop in performance for every model w.r.t the other non-English datasets. At first glance, we notice a high amount of English segments (e.g., tables, figure captions, scientific concepts) in documents in KoreaScience. To investigate this, we use the cld2 library²¹ to detect the language in each non-English document. We consider the percent confidence of the top-1 matching language as an indicator of the presence of the main language (i.e., French, Spanish, Portuguese or Korean) in a document, and average the results to obtain a score for the whole dataset. Table 5 reports the average percent confidence obtained on each split, for each dataset. We find that the percentage of text written in the main language in KoreaScience (i.e., Korean) is smaller than in other datasets. As the MBART-based models expect only one language in a document (the information is encoded using a special token), we claim the strong presence of non-Korean segments in KoreaScience causes them to suffer from interference problems. Therefore, we highlight that KoreaScience is a more challenging dataset, and

²¹ https://github.com/GregBowyer/cld2-cffi

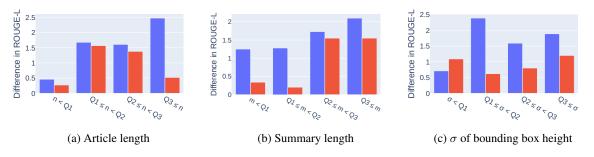


Figure 1: Benefit of using layout on arXiv-Lay (blue) and PubMed-Lay (red), defined as the difference in ROUGE-L scores between BigBird-Pegasus+Layout and BigBird-Pegasus. For each dataset, quartiles are calculated from the distributions of article lengths (a), summary lengths (b) and variance in the height of the bounding boxes (c). ROUGE-L scores are then computed per quartile range, and averaged over each range.

we hope our work will boost research on better long-range, multimodal *and* multilingual models.

Overall, results show a clear benefit of integrating layout information for long document summarization.

5.2 Human Evaluation

Metric	BigBird	BigBird+Layout
Precision %	35.15 (0.81)	37.51 (0.70)
Recall %	28.07 (0.73)	33.59 (0.86)
Coherence	3.80 (0.38)	3.75 (0.62)
Fluency	4.48 (0.03)	4.34 (0.16)
Overlap %	8.77 (0.24)	7.49 (0.36)
Flow %	30.75 (0.68)	33.02 (0.71)

Table 6: Average human judgement scores obtained by comparing gold-truth abstracts and summaries generated by BigBird and BigBird+Layout from 50 documents sampled from arXiv-Lay and HAL. Inter-rater agreement is computed using Krippendorff's alpha coefficient, and enclosed between parentheses.

To gain more insight into the effect of document layout for summarizing long textual content, we conduct a human evaluation of summaries generated by BigBird-Pegasus/BigBird-MBART and their layout-aware counterparts. We choose the BigBird-based models over the LED ones, as the gain in augmenting BigBird with layout is much more apparent. We evenly sample 50 documents from arXiv-Lay and HAL test sets, filtering documents by their topics (computer science) to match the judgment capabilities of the three human annotators. We design an evaluation interface (see Section C.2 in the appendix). For each sentence s_i in the generated summary, we ask the annotators to highlight the relevant tokens in s_i , along with the equivalent parts in the ground-truth abstract (denoted h_i). Further, we ask them to rate the summary in terms of coherence and fluency, on a scale of 0 to 5, following the DUC quality guidelines (Dang, 2005). Finally, annotators are asked to penalize summaries with hallucinated facts. The highlighting process allows us to compute precision and recall as the percentage of highlighted information in the generated summary and the ground-truth abstract, respectively. Moreover, we can compute an overlap ratio as the percentage of highlighted information that appears several times in the generated summary. Lastly, we calculate a flow percentage that evaluates how well the order of the groundtruth information is preserved by computing the percentage of times where the highlighted text h_i in the gold summary for one generated sentence s_i follows the highlighted text h_{i-1} for the previous sentence s_{i-1} (i.e. where any token from h_i occurs after a token in h_{i-1}). Table 6 reports the scores for each metric and model, averaged over all 50 documents, along with inter-rater agreements, computed using Krippendorff's alpha coefficient. We find that adding layout to the models significantly improves precision and recall, results in less overlap (repetition), and is more in line with the ground truth order. Further, annotators did not encounter any hallucinated fact in the 50 generated summaries. To conclude, reported results show that human annotators strongly agree that adding layout generates better summaries, further validating our claim that layout provides vital information for summarization tasks.

5.3 Case Studies

To have a better understanding of the previous results, we focus on uncovering the cases in which layout is most helpful. To this end, we identify features that relate to the necessity of having layout: 1) article length, as longer texts are intuitively easier to understand with layout, 2) summary length, as longer summaries are likely to cover more salient information, and 3) variance in font sizes (using the height of the bounding boxes), and, as such, the complexity of the layout. The benefit of using layout is measured as the difference in ROUGE-L scores between BigBird-Pegasus+Layout and its purely textual counterpart, on arXiv-Lay and PubMed-Lay. We compute quartiles from the distributions of article lengths, ground-truth summary lengths, and variance in the height of bounding boxes.²² Based on the aforementioned factors, the scores obtained by each model are then grouped by quartile range, and averaged over each range, see Figure 1. On arXiv-Lay, we find that layout brings most improvement when dealing with the 25% longest documents and summaries, while, for both datasets, layout is least beneficial for the shortest documents and summaries. These results corroborate our claim that layout can bring important information about long-range context. Concerning the third factor, we see, on PubMed-Lay, that layout is most helpful for documents that have the widest ranges of font sizes, showcasing the advantage of using layout to capture salient information.

6 Conclusion

We have presented LoRaLay, a set of large-scale datasets for long-range and layout-aware text summarization. LoRaLay provides the research community with 4 novel multimodal corpora covering French, Spanish, Portuguese, and Korean languages, built from scientific articles. Furthermore, it includes additional layout and visual information for existing long-range summarization datasets (arXiv and PubMed). We provide adapted architectures merging layout-aware and long-range models, and show the importance of layout information in capturing long-range dependencies.

7 Limitations

The proposed corpus is limited to a single domain, that of scientific literature. Such limitation arguably extends to the layout diversity of documents. In terms of risks, we acknowledge the presence of Personally Identifiable Information such as author names and affiliations; nonetheless, such information is already voluntarily made public by the authors themselves.

8 Acknowledgements

We thank the reviewers for their insightful comments. This work is supported by the Association Nationale de la Recherche et de la Technologie (ANRT) under CIFRE grant N2020/0916. It was partially performed using HPC resources from GENCI-IDRIS (Grant 2021-AD011011841).

References

- Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R Manmatha. 2021. Docformer: End-to-end transformer for document understanding. *arXiv preprint arXiv:2106.11539*.
- Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv* preprint arXiv:2004.05150.
- Łukasz Borchmann, Michał Pietruszka, Tomasz Stanislawek, Dawid Jurkiewicz, Michał Turski, Karolina Szyndler, and Filip Graliński. 2021. Due: End-to-end document understanding benchmark. In *Thirty-fifth* Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2).
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Shuyang Cao and Lu Wang. 2022. Hibrids: Attention with hierarchical biases for structure-aware long document summarization. *arXiv preprint arXiv:2203.10741*.
- Zewen Chi, Li Dong, Furu Wei, Wenhui Wang, Xian-Ling Mao, and Heyan Huang. 2020. Cross-lingual natural language generation via pre-training. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7570–7577.
- 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. *arXiv preprint arXiv:1804.05685*.
- Hoa Trang Dang. 2005. Overview of duc 2005. In *Proceedings of the document understanding conference*, volume 2005, pages 1–12.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

²²The quartiles are provided in Appendix C.3.

- Filip Graliński, Tomasz Stanisławek, Anna Wróblewska, Dawid Lipiński, Agnieszka Kaliska, Paulina Rosalska, Bartosz Topolski, and Przemysław Biecek. 2020. Kleister: A novel task for information extraction involving long documents with complex layout. arXiv preprint arXiv:2003.02356.
- Adam W Harley, Alex Ufkes, and Konstantinos G Derpanis. 2015. Evaluation of deep convolutional nets for document image classification and retrieval. In 2015 13th International Conference on Document Analysis and Recognition (ICDAR), pages 991–995. IEEE.
- Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28.
- Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and CV Jawahar. 2019. Icdar2019 competition on scanned receipt ocr and information extraction. In 2019 International Conference on Document Analysis and Recognition (ICDAR), pages 1516–1520. IEEE.
- Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. 2019. Funsd: A dataset for form understanding in noisy scanned documents. In 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW), volume 2, pages 1–6. IEEE.
- Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.
- Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pretraining for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.
- Ling Luo, Xiang Ao, Yan Song, Feiyang Pan, Min Yang, and Qing He. 2019. Reading like her: Human reading inspired extractive summarization. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3033–3043.
- Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. 2022. Infographicvqa. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 1697–1706.

- Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2200–2209.
- Lesly Miculicich and Benjamin Han. 2022. Document summarization with text segmentation.
- Seunghyun Park, Seung Shin, Bado Lee, Junyeop Lee, Jaeheung Surh, Minjoon Seo, and Hwalsuk Lee. 2019. Cord: a consolidated receipt dataset for post-ocr parsing. In *Workshop on Document Intelligence at NeurIPS 2019*.
- Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch.
- Rafał Powalski, Łukasz Borchmann, Dawid Jurkiewicz, Tomasz Dwojak, Michał Pietruszka, and Gabriela Pałka. 2021. Going full-tilt boogie on document understanding with text-image-layout transformer. *arXiv preprint arXiv:2102.09550*.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.
- Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. Mlsum: The multilingual summarization corpus. *arXiv preprint arXiv:2004.14900*.
- Eva Sharma, Chen Li, and Lu Wang. 2019. Bigpatent: A large-scale dataset for abstractive and coherent summarization. *arXiv preprint arXiv:1906.03741*.
- Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR.
- Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.
- Elif Toprak and Gamze Almacioğlu. 2009. Three reading phases and their applications in the teaching of english as a foreign language in reading classes with young learners. *Journal of language and Linguistic Studies*, 5(1).
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-ofthe-art natural language processing. *arXiv preprint arXiv:1910.03771*.

- Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. 2021. LayoutLMv2: Multi-modal pre-training for visually-rich document understanding. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 2579–2591, Online. Association for Computational Linguistics.
- Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. Layoutlm: Pre-training of text and layout for document image understanding. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pages 1192–1200.
- Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. Advances in Neural Information Processing Systems, 33:17283–17297.
- Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

LoRaLay: A Multilingual and Multimodal Dataset for *Long Range* and *Layout-Aware Summarization –* Appendix

A Datasets Construction

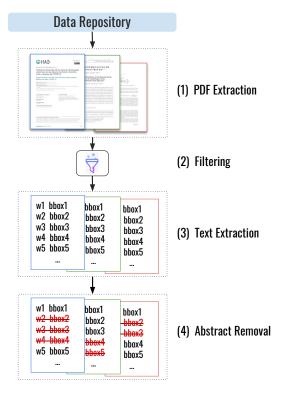


Figure 2: Dataset Construction Process.

A.1 Extended Datasets – Lost Documents

Figure 3 provides details on the amount of original documents lost in the process of augmenting arXiv and PubMed with layout/visual information. We observe four types of failures, and provide numbers for each type:

- The link to the document's PDF file is not provided (*Unavailable PDF*);
- The PDF file is corrupted (i.e., cannot be opened) (*Corrupted PDF*);
- The document is not digital-born, making it impossible to parse it with PDF parsing tools (*Scanned PDF*);
- The document's abstract cannot be found in the PDF (*Irretrievable Abstract*).

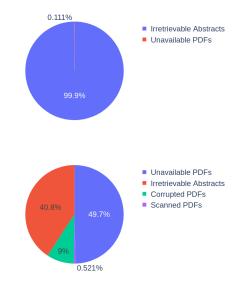


Figure 3: Distribution of failure types in arXiv-Lay (top) and PubMed-Lay (bottom).

A.2 KoreaScience – Extraction Rule

Korean documents in KoreaScience are extracted by restricting search results to documents containing the word "Korean" in the publisher's name. We show that this rule does not bias the sample towards a specific research area. We compute the distribution of topics covered by all publishers, and compare it to the distribution of topics covered by publishers whose name contains the word *Korean*. Figure 4 shows that the distribution obtained using our rule remains roughly the same as the original.

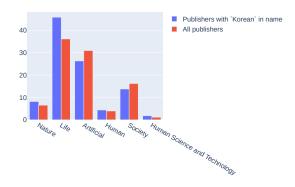


Figure 4: Distribution of topics covered by all publishers (red) vs distribution of topics covered by publishers whose name contains the word *Korean* (blue).

A.3 Samples

We provide samples of documents from each dataset in Figure 5.

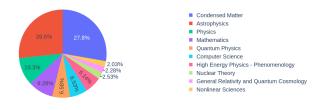
A.4 Datasets Statistics

C Results

Detailed Results

C.1

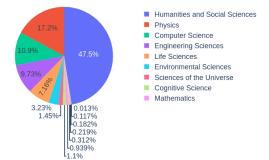
The distribution of research areas in arXiv-Lay and HAL are provided in Figures 6 and 7, respectively. Such distributions are not available for the other datasets, as we did not have access to topic information during extraction.



Model	R-1	R-2	R-L
MBART	47.05	22.23	42.00
MBART+Layout	46.65	21.96	41.67
BigBird-MBART	49.85	25.71	45.04
BigBird-MBART+Layout	49.99	25.20	45.20

Figure 6: Distribution of research areas in arXiv-Lay.

Table 8: ROUGE scores on HAL. Best results are reported in bold.



Model	R-1	R-2	R-L
MBART	17.33	7.70	16.94
MBART+Layout	15.43	6.69	14.98
BigBird-MBART	18.96	8.01	18.55
BigBird-MBART+Layout	20.36	9.49	19.95

Figure 7: Distribution of research areas in HAL.

B Experiments

B.1 Implementation Details

Models were implemented in Python using Py-Torch (Paszke et al., 2017) and Hugging Face (Wolf et al., 2019) librairies. In all experiments, we use Adafactor (Shazeer and Stern, 2018), a stochastic optimization method based on Adam (Kingma and Ba, 2014) that reduces memory usage while retaining the empirical benefits of adaptivity. We set a learning rate warmup over the first 10% steps – except on arXiv-Lay where it is set to 10k consistently with Zaheer et al. (2020), and use a square root decay of the learning rate. All our experiments have been run on four Nvidia V100 with 32GB each.

Table 10: ROUGE scores on KoreaScience. The bestresults are reported in bold.

C.2 Human Evaluation

Using the Streamlit²³ framework, we design and develop an interface to aid human evaluation of summarization models.²⁴

²³https://streamlit.io/

²⁴The code is publicly available at https://anonymous.4open.science/r/loralay-eval-interface-C20D.

Model	arXi	v / arXiv	-Lay	PubMe	PubMed / PubMed-Lay		
Widder	R-1	R-2	R-L	R-1	R-2	R-L	
Pegasus (Zhang et al., 2020)	44.21	16.95	38.83	45.97	20.15	41.34	
BigBird-Pegasus (Zaheer et al., 2020)	46.63	19.02	41.77	46.32	20.65	42.33	
T5 (Raffel et al., 2019)	42.79	15.98	37.90	42.88	17.58	39.23	
LED (Beltagy et al., 2020)	45.41	18.14	40.74	45.28	19.86	41.54	
LED+Layout	45.51	18.55	40.96	45.41	19.74	41.83	
MBART	37.64	13.29	33.49	41.19	16.04	37.47	
Pegasus	43.81	17.27	39.07	43.52	17.96	39.75	
Pegasus+Layout	44.10	17.01	39.25	43.59	18.24	39.85	
BigBird-Pegasus	44.43	17.74	39.59	44.80	19.32	41.09	
BigBird-Pegasus+Layout	46.02	18.95	41.15	45.69	20.38	42.05	

Table 7: ROUGE scores on arXiv-Lay and PubMed-Lay. Reported results obtained by Pegasus and BigBird-Pegasus
on the original arXiv and PubMed are reported with a gray background. The best results obtained on arXiv-Lay and
PubMed-Lay are denoted in bold.

	S	ciELO-E	ES	S	ciELO-P	Т
Model	R-1	R-2	R-L	R-1	R-2	R-L
MBART	41.04	15.65	36.55	41.18	15.53	36.42
MBART+Layout	42.27	15.73	37.47	39.45	14.17	34.37
BigBird-MBART	42.64	16.60	37.76	44.85	18.70	39.63
BigBird-MBART+Layout	45.64	19.33	40.71	45.47	20.40	40.51

Table 9: ROUGE scores on the SciELO datasets. The best results are reported in bold.

649

Next

LoRaLay Evaluation Interface

Evaluation guidelines ~ Document 0806.3537 (1/50) Statistical Learning of Arbitrary Computable Classifiers Link to full document Model A Model B Ground-truth abstract Statistical learning theory chiefly studies restricted hypothesis classes, particularly those with finite Vapnik-Cherononekis (VC) dimension. The fundamental quantity of interest is the sample complexity, the number of samples required to learn to a specified level of accuracy, here we consider learning own the set of all computable leaking interforms. Since the VC-dimension is infinite and a priori (uniform) bounds on the number of samples to have learned, we first show that learning in this setting is indeed possible, and develop a learning algorithm. We first show that however, that bounding sample complexity independently of the distribution is impossible. Notably, his immossibility, earlier clark of the environment and the environment of the environment of the commutable and however, that bounding sample complexity independently of the distribution is impossible. Notably, this immossibility is entried vice to the environment and the commutable and the commutable complexity independently of the distribution is impossible. this impossibility is entirely due to the requirement that the **learning** algorithm be computable, and not due to the **statistical** nature of the problem. You selected the following sentence generated by Model B. Highlight the parts in the sentence that can be found in the ground-truth abstract. Conventional statistical learning theory attempts to bound the number of samples needed to learn to a specified level of accuracy for each of the above models (e.g. neural networks, support vector machines). Next sentence Summary generated by Model B Sentence Precision (%) Conventional statistical learning theory attempts to bound the number of samples needed to learn to a specified level of accuracy for each of the above models (e.g. neural networks, support vector machines). However, if we allow ourselves to change the model, then the VC-dimension of the overall learning algorithm is not finite, and much of statistical learning theory does not directly apply. 16.67 In contrast, we prove that distribution-independent bounds do not exist 72.22 altogether for computable learning algorithms in our setting. Our results imply that computable learning algorithms in the universal 0.0 setting must "waste samples" in the sense of requiring more samples than is necessary for statistical reasons alone. Pecall (9 30.15 Coherence 0 0 1 0 2 0 3 0 4 0 5 Fluency 0 0 0 1 0 2 0 3 0 4 0 5 I am unable to evaluate this document.

C.3 Analysis of the Impact of Layout

Table 12 lists the quartiles computed from the distributions of article lengths, summary lengths, and variation in the height of bounding boxes, for arXiv-Lay and PubMed-Lay.

Figure 8: LoRaLay evaluation interface.

	LED	LED+Layout	Pegasus	Pegasus+Layout	BigBird-Pegasus	BigBird-Pegasus+Layout
Т5	2.84 / 2.31	3.06 / 2.60	1.17 / 0.52	1.35 / 0.62	1.69 / 1.86	3.25 / 2.82
LED	-	0.22 / 0.29	1.67 / 1.79	1.49 / 1.69	1.15 / 0.45	0.41/0.51
LED+Layout	-	_	1.89 / 2.08	1.71 / 1.98	1.38 / 0.74	0.19/0.22
Pegasus	-	_	_	0.34 / 0.10	0.52 / 1.34	2.08 / 2.30
Pegasus+Layout	-	_	_	-	0.34 / 1.24	1.90 / 2.20
BigBird-Pegasus	-	-	-	-	-	1.56 / 0.96

Table 11: Absolute ROUGE-L score differences between each pair of models, on arXiv-Lay/PubMed-Lay.

Distribution	Q1 arXiv-Lay PubMed-Lay		Q2 arXiv-Lay PubMed-Lay		Q3 arXiv-Lay PubMed-Lay	
Article Length	6,226	3,513	9,142	5,557	13,190	8,036
Summary Length	119	130	159	182	202	247
σ of bounding box height	3.37	1.34	3.98	1.73	4.70	2.28

Table 12: Quartiles calculated from the distributions of article lengths, summary lengths, and variation in the height of bounding boxes, for arXiv-Lay and PubMed-Lay.

	Journal of Biomedicine and Biomedicine (2006) 306-311 - PE, ST1127-0304-0502X - http://f/h.biod.aud.co 3025/15274114 Activities			
MPF-000-131	Anthocyanins and Human Health: An In Vitro Investigative Approach			
Experimental Review of Photon Structure Func- tion Data	Mary Ann Lilas			
Richard Neise Mar Fauk-barttar für Frysk (Weiner-Hörenberg Institut), Filninger Hing 6, D-8005 Min- chen, Germang, E. vank, Bichard Moiorflaup-app, di [*]	Department of Natural Browners & Ferrinmental Summer, Gelorg of Appindence Concerns and Aniversential Science, Elisionity of Hilbinis, Ushan, R. 40005, USA Basinish Zaquii 2006, annial 1411 Mary 2008			
den, Gemany, E-mail: Echard-Noisethupp.mpg.de* DDR: uil le anigord	Advector in present on a discussion of the sector of the s			
The present handedge of the structure of the photon is presented hand on a results obtained by measurements of photons from times from times $d = v^{-1}$ colliders. From the seprestical both	w) poststati en anglese ments, ten de constitución nem intervisión compositori en architecto a metapore, simulatore en al compositori en architecto en al compositori en architecto en al compositori en architecto en al compositore en al composi			
The point hand-large of the strengt of large hands in pointed hand and a strength column by restrictions in the strength restriction of large hands in pointed hand and the strength restriction of large hands in the strength restriction of large hand large hand are also response to most pointed intradictions of the hadron in our strength restrength r	Data are summaring produktably produced to suggestedy constrained plant and achieves and the agent adversing plan health and multiplication means the summer of the angle splant and the three theoremses found in which forkers as the stillable of provide app and profile accommittation, and are anomable to biolubology on that methodic fate can be investigated after impedien.			
1 Introduction	ANTHOCKNING AND BIOMEEXCINAL PROPERTIES suggests that other mechanisms of action are also responsed and the second hashin benefits [2, 3, 6, 5], stathocynal			
• Another interface and a strength indication is a large fielding data both field of each field in the strength of the stre	distribution and intervention of the intervention groups of interactional activity interactivity interactiv			
function $I_i^{\alpha}(x, Q^{\alpha})$ the main sense of interact on the behavior at the values of x and the evolution with the measurum scale Q^{α} , which is predicted by QCD to be input built. The experimental interaction is dominismly by the results thrum the four LBP experiments.	of phanolec in foods, comparie a gauge of our 4000 C _{ell} anomatic plant ampendia with miniple-molectanics parame (non-rail-anlagarchic/closdcomplinder.html). Ing capillary permissibility and trajility, and mouber			
This review is based on outfor work [1, 2] and an an extransic protoin a number of epident fequence, together with a comparison of the experimental data with the user parameterizations of $P_2^{-1}(a, Q^2)$ that because available sizes these. Only somethics at the structure of epident photons $P_2^{-1}(a, Q^2)$ that because $P_2^{-1}(a, Q^2)$ is the structure of epident photons of	The primary players in this group include the ar- menophaning (5, 1, 6, 8). The chemical arcreams (p- these) primarily, the primarily, the facenois (querratis, karepters), farenses (kareas, - data archive), and you can be a barriege at minority (farenses (kareas), barriers).			
	nariopositi, flavari-loik (zatachia, opizanchia, galizara- ortise [11, 12] and the encarazortianzion relational echicit, and , athracig sometimes classified separately. the isolances (opizaria, databai), Physical methods in this mental [1]. The arthacyaria likensus induces come are			
2 Structure function measurements The photon on florator into a formise and emission that consider with the quantum num- less of the shaken and within the final coses of the Heisenberg meantainty spinsible. These	chas are frequently referred to as biofaceoside due to their multifactional due is in human handor, however, approach to there are bioardiabilit and antheopenia is find are stylicitly ingested as concerning (7), 13, 14, 15, 16, 10; (righting) poments of contrastruct dimensional contrastructs.			
The plane is an induced and a balance and the discussion of monotone WM for experiment mat- ing of the plane and start is a balance as if the discussion of monotone with the plane is a start is a s	Daily instate is estimated from 500mg to 1 g, but can be neural golf an individual in constanting flownoid experi- ments (golps used accuracity in the state of t			
section of the deep-inclustic electron-photon scattering ¹ process sketched in Figure 1. In this shorth this provide the Photonff Candowne is Hamberg on May 12 (200). ⁺ Sin this paper, the two decision are majorized photon photon.	weight [11]. The coincil anthecyanias are the most recognized, yields members of the biofeneous dyspectencicals. The first net and the second dyspectencicals are second as a first net first net and the second dyspectencicals. The second dyspect is first net first net and the second dyspectencicals. The second dyspect is first net first net and the second dyspect net and the second dyspect is first net first net and the second dyspect net and the second dyspect net first net and the second dyspect net and the second dyspect net and the second dyspect net first net and the second dyspect net and the second			
Paloritone 1	<text><text><text><text><text></text></text></text></text></text>			
(a) arXiv-Lay	(b) PubMed-Lay			
(1)				
Les représentations des enregionais de ZPP aur la reletais des dévalements travers la prisme des élèves en grande résustes scolaire Ablé dans hour cleur des 25.207 (cleur VI) finisk, mit, mich				
en vers le site <u>(duration & Devenir</u>	DDt: http://doi.org/10.22585/houpdowic=54.1.48			
en vers la ute <u>Les cohien pildageficus.</u>	Tendencias temporales de los patrones de búsqued sobre Servicios de Atención de Salud a Domicilio			
In length and the partnerse security of Fall Res Nortex L hardner scheduler desting desting and an experimental scheduler and the scheduler desting desting of the scheduler desting desting when the fall result of the scheduler desting and the scheduler desting desting desting desting when the fall result of the scheduler desting des	antes y después del COVID-19			
robussaum des (colm (PI) en zone d' douariso prioritaire (IZP) pais des ventreles avec des d'entres eux. Le risme des élèves en grande réunités sociaire (GAS) dans les 12P a étés chois pour étudier la différence de exerption des enseignaisment en forcitor des la rismita des élévieux les politiciai des parents.	Temporal trends in Home Care Services search patterns before and after COVID-19			
	Rubis Palono-Ukarnel @ 0000-0002-1850-4337 Jela Sanchaz-Tomel @ 0000-0001-0541-8737			
keb duh pristerations – enseignents – coéducation – grande visuate scolaire - éducation prioritaire	Julia Sanchara Somer d' († 2003) 2001–2014, 2777 Benjanni Petros Channes († 2007) 2001–2013 2013 1. Universidad Migraf Hendrider, Departmente de Schul Phâtea Hendrie de la Cinecia, Sant Janes d'Alasant, Alasant, Eguida			
Canolana MCRICE – ADIG – ANU Leanning Associativity annu 30	Equilation to a second second and the second s			
NextBATE with controls and an experiment of Relation, and Annual	Correspondencia/Correspondence Phonoloción/Vending Rulas Inducedorase Esta tudes engen la machada engen la funcionada, plantas adiguesta ano Creativo de altero construito de antero construito de antero construito de antero construito de antero			
the list effects got absolutent et executed. On the internative data is littlefeature (Larranie, 2015) une nilgonaution des conduies des éléves longul is fant face aux apprentisages en fonction de l'attande de un premet. Les states efficiés (encourges une sellados possible decalignation), car la familie est	polena addigmat ann Cantribucione de autoria/Aufbar contributione Bacticido/Bactived Table to same has contribuido por spoil en la exitación 20 00 2021 de esta telapor			
nnidérie comme un partensire de l'école avec uns place importante dans la sostanté de l'éléve (Houssaye, 2011, Que pensent les enseignents de ces déclassion 7 quelles cont les repelsentations des enseignants comment l'Influence de la métries availaires un la réaction ser de la réaction de la series de la contraction de	Acaptado/Accepted 12.10.2021			
ctor étude se propose, dans une première enquête, d'interroper par questionnaire 1750 enseignants. Acché elémentaire, tocéns en 2007, autour de leur quotides dans les classes et, dans une devolème	Coefficito de Interesen/Comparing interesed Los catores no presentes carrillato de intereses.			
der halfung angen, dass zur gereiten angelich, Certerunge zur gesternennen 130 mergenetis der fahlung angelichen einer einen einer eine einer einer				
our children Transformation and engineering and an and and	CÓBIO CITAR ESTE TRABAJO HOW TO CITE THES RAFER			
rtre les parents des CGIS et les auton.	Palomobiliname R, Sanchez-Tormo J, Palomobiname B. Tendenciae temporales de los patrones de biospeda e Servicios de Atención de Salud a Danicilia ontes y despute del COVID-19. Haup Danic. 2021;5(d):187/45.			
1	Heap Domic. 2021.5(4):187.196			
(c) HAL	(d) SciELO-ES			
m ha (Jan kad, s.1), 4/313, 201 hgyllandwar (Si 104/a-2019 et al. 3.201 katil)				
	50% 1225 ~ 1742) 제품: - 신하/6/27 1/122015년 hos 2011 17 2.015 원리(미미미미미미미미미미미미미미미미미미미미미미미미미미미미미미미미미미미미			
UMA LUZ SOBRE AS				
Reparting me properties MOCAMBIQUE NO	수도권 제조업 창업 활동의 공간적 분포 변화			
ktolini coectular. MOÇAMBIQUE NO istretisgilandate OITOCENTOS: A	- 공간 마르코프 체인의 응용 -*			
Alitz Lews Alitz Litrica DE JOÃO LUIZ AIROZA	송장현** • 안순변*** • 임업****			
DE JOAO LUIZ AIROZA (1827-1828)*	Changes in Spatial Distribution of Manufacturing Startup Activities in the Capital Region, Korea:			
	A Spatial Markov Chain Approach*			
Contato @Gilberto da Silva Guizelin''	Changhyan Seerg* - Sounbeam Ahn** - Up Lim***			
Ra Bali Huranon, 423, que 300 BECODORO- Landon - Huran	國際監督 문 연구는 2000년부터 2009년부터 2019년세지를 문식의 시간적 법취료 설립하여 제조업 열점 확실이 공간하고 이해한 범죄를 보여했는지를 탑성적으로 보산하고, 항후 열업 활성적 문도 위한 변화를 배하하는 것을 목적으로 또			
	다. 문식을 취해 2000년부터 2018년카지의 [전문사업해조사」마이그프레이터 제조업 사업에 자료를 활용하였 한국산업업구관의 1875ASS 문제세계에서 제시하는 40대 제조업 기준에 따라 제조업을 내외 세우 산업군으로			
Apóra a suinatura da Converção de Hiláz com a Gel-Teretanha, pela qual o go- verno de D. Peloto I concreden, em traca do recombricimento británicos, coltir o terefora transmitiativa de doravera na esta deversita en acuita de HING este en estador da	문한 후, 수도권 행동구역 유명을 수준에서 공간자가들한 문식 및 공간 데티프프 체인 문식을 수행하였다. 문식 과에 데티핀, 그리가슬란입군 및 중고리가슬란입군의 장업 활동은 시간이 조즘에 따라 경기도 남부를 중심으로			
tranco transmuttino de abricanse para o Impérido a partir de 1030, foram cristala representaçãos consultares abraditaras na Adras Portugenas com a seguitará fina- lidadas de prestegra a atacição de negretavis bossidations nass de leguit- dadas de prestegra e atacição de negretavis bossidations nass de leguit-	응지고 있는 것으로 나타났으며, 유가제가 높았다. 이 유가제가 높았다. 이 유가제가 높았다. 같은 것 위해가 높았다. 같은 것을 해놓았다. 지금 하는 수소는 전 의해으로 든 산 지고 있는 것으로 나타났다. 것이에서 해외 제외에서 하지 위해 가장에 가장하는 것이 가능하는 것으로 다양해 특히 하는 것이 다. 것 말 같은 아니 이 또 방법 하는 것이 가능 것이 가능하는 것이 가능하는 것이 가능하는 것이 가능하는 것으로 가장하는 것이 가능하는 것으로 가장하는 것이 가능하는 것으로 가능하는 것이 가능하는 것으로			
Remain The main and the former than 6 and the main as 6 a fibre starting play spin of gen- trees of the fibre starting play and the starting play spin of the starting play spin spin of the starting play and the starting play and the starting play spin of play and the starting play and the starting play and the starting play and the starting play and the starting play and the starting play and play and the starting play and the starting play and the starting play and play and the starting play and the starting play and the starting play and play and the starting play and the starting play and the starting play and play and the starting play and the starting play and the starting play and play and the starting play and the starting play and the starting play and play and the starting play and the starting play and the starting play and play and the starting play and the starting play and the starting play and play and the starting play and the starting play and the starting play and play and the starting play and the starting play and the starting play and play and the starting play and the starting play and the starting play and the starting play and play and the starting play	때, 항접 활동이 철망하여 발망하는 가지 및 그와 전점하고 있는 가격의 결주 등 등은 관정 등 적당가 가능한 가능 나타나 금정 방법 공간 표가가 관객하는 것으로 확당되었다. 문 연구는 영구의 불음의 주요 원란이 당근 부모으 업 활동이 문도 제법 변화품 들여적으로 논성적으로에 장점 속 빛 입기의 불음과 관련한 가격 철명하여 시사용			
	제공하고가 하였다.			
Palavras-chave Relações internacionais - Relações Itzatil-Moçambique - Missão consular - Tráfico everavos - África Oriental	40000 제조한, 같던, 함께적공간가요 초선, 공간아이오그드 바인			
	Abstract. This study also so sphere how manufacturing start up activities from 2000 to 2018 how changed spatial and to positive changes in distribution parameter of fature start-up activities. For the analysis, the or cannot an iterativities and the positive change in the distribution of the start and the s			
* Todar ao obras e todos os documentos sufilizados na pesquisa e na elaboração do artigo são citados nas notas e na bibliográfia.	nicrodus from 2000 to 2018 tour coul, and the manufacturing industry was chariffed into frast decided industri * 이 논문은 2019년 대한국국 교육부터 한국국구수에서 인원시험(무수 공부구수에서 인원시험(무수 공부수) 사람은 약수 여행은 단구 2010년 - 201			
**Indian at attack taking a discussion statistica in propins or an observation statistica in discussion statistica in discussion. Construction of the statistica in the sta	* 시 고등은 300% 행동(1 소용자) 등 40% 등 10% 50% 50% 50% 50% 50% 50% 50% 50% 50% 5			
die processo st ² 2018/07796-1, Fundação de Anguesa Propins do Estado do São Pado (13/9537). 1				
(e) SciELO-PT	(f) KoreaScience			

