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

**Anonymous ACL submission** 

#### Abstract

Text Summarization is a popular task and an 002 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 summa-011 rization datasets only provide plain text con-012 013 tent. 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 017 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 022 built from scholar resources - covering French, Spanish, Portuguese, and Korean languages. Further, we propose new baselines merging layout-aware and long-range models - two or-026 thogonal approaches - and obtain state-of-theart results, showing the importance of combining both lines of research.

#### 1 Introduction

027

034

038

040

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.

042

043

044

045

046

047

051

052

054

059

060

061

062

063

064

065

066

067

068

069

070

071

072

073

074

075

076

077

079

081

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:

- 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 models, including adapted architectures for multi-modal longrange summarization, and report their results, showing that (1) performance is far from being optimal; and (2) layout might be a valuable information.

#### 2 Related Work

090

094

#### 2.1 Layout/Visually-rich Datasets

Document Understanding covers problems that in-100 volve reading and interpreting visually-rich docu-101 ments (in contrast to plain texts), requiring com-102 prehending the conveyed multimodal information. 103 Hence, several tasks with a central layout aspect 104 have been proposed by the document understand-105 ing community. Key Information Extraction tasks 106 consist in extracting the values of a given set of 107 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. No-110 table datasets include FUNSD (Jaume et al., 2019) 111 for form understanding in scanned documents, and 112 SROIE (Huang et al., 2019), as well as CORD 113 (Park et al., 2019), for information extraction from 114 receipts. Graliński et al. (2020) elicit progress on 115 deeper and more complex Key Information Extrac-116 tion by introducing the Kleister datasets, a collec-117 tion of business documents with varying lengths, 118 released as PDF files. However, the documents 119 in Kleister often contain single-column layouts, 120 which are simpler than the various multi-column 121 layouts considered in LoRaLay. Document VQA 122 is another popular document understanding task 123 that requires processing multimodal information 124 (e.g., text, layout, font style, images) conveyed by 125 a document to be able to answer questions about a 126 visually rich document (e.g., What is the date given 127 at the top left of the form?, Whose picture is given 128 in this figure?). The DocVQA dataset (Mathew 129 et al., 2021) and InfographicsVQA (Mathew et al., 130

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. 131

132

133

134

135

136

137

138

139

140

141

142

143

144

145

146

147

148

149

150

151

152

153

154

155

156

157

158

159

160

161

162

163

164

165

166

167

168

169

170

171

172

173

174

175

176

177

178

179

180

To the best of our knowledge, document summarization has not yet been considered by the document understanding community, despite being one of the most popular NLP tasks among researchers.

#### 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.

All the summarization datasets proposed so far 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-

266

267

268

269

227

cially long ones – often recommend roughly pre-181 viewing them to break them down into their ma-182 jor sections (Toprak and Almacioğlu, 2009; Luo et al., 2019). This suggests that NLP systems might leverage multimodal information in documents. Although not all documents are explicitly 186 organized into clearly defined sections, the great 187 majority contains layout and visual clues (e.g., a physical organization into paragraphs, bigger headings/subheadings) which help structure their textual 190 contents and facilitate reading. Thus, we argue that layout is crucial to summarize long documents, and 192 propose a corpus of long documents with layout 193 information. Furthermore, to address the need for 194 multilingual training data (Chi et al., 2020), we in-195 clude not only English documents, but also French, Spanish, Portuguese and Korean ones. 197

## **3** Datasets Construction

199

201

202

205

209

210

211

212

214

215

216

217

218

219

222

225

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.<sup>1</sup>

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)<sup>2</sup>, 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)<sup>3</sup>, 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)<sup>4</sup>, 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.<sup>5</sup> In total, about 39% (35%) of the original documents in arXiv (PubMed) were lost.

**arXiv-Lay** The original arXiv dataset was constructed by converting the LATEX files to plain text. To be consistent with the other datasets – for which LATEX 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.<sup>6</sup>

**PubMed-Lay** For PubMed, we use the PMC OAI Service<sup>7</sup> to retrieve both abstracts and PDF files.

**HAL** We use the HAL API<sup>8</sup> to download research papers written in French. To avoid excessively long (e.g. theses) or short (e.g. posters)

<sup>&</sup>lt;sup>1</sup>We make the corpus-construction code publicly available at https: /anonymous.4open.science/r/loralay-datasets-03FF.

<sup>&</sup>lt;sup>2</sup>https://hal.archives-ouvertes.fr/

https://www.scielo.org/
4....

<sup>&</sup>lt;sup>4</sup>http://www.koreascience.or.kr

<sup>&</sup>lt;sup>5</sup>For more details on this, see Section A.1 in the Appendix.

<sup>&</sup>lt;sup>6</sup>https://www.kaggle.com/Cornell-University/arxiv

<sup>//</sup>https://www.ncbi.nlm.nih.gov/pmc/tools/oai/

<sup>8</sup> https://api.archives-ouvertes.fr/docs/search

documents, extraction is restricted to journal andconference papers.

272SciELOUsing Scrapy,9 we crawl the following273SciELO collections: Ecuador, Colombia, Paraguay,274Uruguay, Bolivia, Peru, Portugal, Spain and Brazil.275We download the PDF files of articles whose ab-276stract and contents are written either in Spanish or277Portuguese. The document language is extracted278from the corresponding metadata. Articles are279then grouped by language, leading to two separate280datasets: SciELO-ES (Spanish) and SciELO-PT281(Portuguese).

282KoreaScienceSimilarly, we scrape the Korea-283Science website to extract research papers. We284limit search results to documents whose publishers'285names contain the word *Korean*. This rule was de-286signed after sampling documents in the repository,287and is the simplest way to get a good proportion288of papers written in Korean.<sup>10</sup> Further, search is289restricted to papers published between 2012 and2902021, as recent publications are more likely to have291digital-born, searchable PDFs. Finally, we down-292load the PDF files of documents that contain an293abstract written in Korean.

#### 3.2 Data Pre-processing

295

299

301

303

305

307

311

312

313

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 Poppler<sup>11</sup>, 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 removed by searching for exact matches; in the case no exact match is found, we look for near

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.

matches using fuzzysearch<sup>12</sup> and regex.<sup>1314</sup> 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.

315

316

317

318

319

320

321

322

323

324

325

326

327

329

330

331

332

333

334

335

336

337

340

341

342

343

344

345

## 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.

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

<sup>9</sup>https://scrapy.org/

<sup>&</sup>lt;sup>10</sup>For further details, see Section A.2 in the Appendix.

<sup>&</sup>lt;sup>11</sup>https://poppler.freedesktop.org/

<sup>12</sup> https://pypi.org/project/fuzzysearch/

 $<sup>^{14}\</sup>mbox{We}$  use a maximum Levenshtein distance of 20 with fuzzy search, and a maximum number of errors of 3 with regex.

Datasat	Iı	Instances			Length	Output	Output Length	
Dataset	Train	Dev	Test	Median	90 <i>%</i> -ile	Median	90%-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.

#### 4 Experiments

### 4.1 Models

346

347

351

354

355

361

367

370

372

374

376

379

For reproducibility purposes, we make the models implementation, along with the fine-tuning and evaluation scripts, publicly available.<sup>15</sup>

**Text-only models with standard input size** Following Zaheer et al. (2020), 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 gap-sentences 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),<sup>16</sup> 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 co-

loralay-modeling-1870/

ordinate 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. 381

383

384

387

389

390

391

392

393

394

395

396

397

398

399

400

401

402

403

404

405

406

407

408

409

410

411

412

413

414

415

416

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 longrange Longformer-Encoder-Decoder (LED) (Beltagy et al., 2020), and iii) the layout-aware, longrange LED+Layout, which we implement similarly to the previous layout-aware models.

<sup>&</sup>lt;sup>15</sup>https://anonymous.4open.science/r/

<sup>&</sup>lt;sup>16</sup>For the sake of clarity, we refer to MBART-50 as MBART.

#### 4.2 Implementation Details

417

418

419

420

421

422

423

424

425

426

427

428

429

430

431

432

433

434

435

436

437

438

439

440

441

442

443

444

445

446

447

448

449

450

451

452

453

454

455

456

457

458

459

460

461

462

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.<sup>17</sup>

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 decreased to 50k for the other non-English datasets.<sup>18</sup> 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. <sup>19</sup> 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 significant gains in integrating layout information into BigBird-Pegasus. This shows that layout information is important when processing long documents, matching our motivations in creating this dataset. Ground-truth The exact formulation of multi-configuration densityfunctional theory (DFT) is discussed in this work. As an alternative to range-separated methods, where electron correlation effects are split in the coordinate space, the combination of Configuration Interaction methods with orbital occupation functionals is explored at the formal level through the separation of correlation effects in the orbital space. When applied to model Hamiltonians, this approach leads to an exact Site-Occupation Embedding Theory (SOET). An adiabatic connection expression is derived for the complementary bath functional and a comparison with Density Matrix Embedding Theory (DMET) is made. Illustrative results are given for the simple two site Hubbard model. SOET is then applied to a quantum chemical Hamiltonian, thus leading to an exact Complete Active Space Site-Occupation Functional Theory (CASSOFT) where active electrons are correlated explicitly within the CAS and the remaining contributions to the correlation energy are described with an orbital occupation functional The computational implementation of SOET and CASSOFT as well as the development of approximate functionals are left for future work

Pegasus We propose a multi-configurational formulation of density functional theory (DFT) based on the separation of static and dynamical correlation effects in the orbital space. We show that this approach allows for a rigorous description of the ground state of strongly correlated systems at the formal level. In particular, we show that the orbital occupation can be used as a basic variable rather than the electron density in order to avoid the notorious double counting problem. Pegasus+Layout We propose a multi-configurational extension of densityfunctional theory (DFT) by separating correlation effects in the coordinate space. The resulting range-separated DFT methods combine rigorously long-range wavefunction theory with short-range DFT. Even though range separation allows for a multi-configurational description of the electron density, it cannot completely isolate static correlation from dynamical correlation simply because the former is usually not a purely long-range effect, even in a dissociated molecule. This point will be discussed further in the following BigBird-Pegasus The exact formulation of multi-configurational density functional theory (DFT) is revisited. A range-separated DFT approach is proposed where correlation effects in the orbital space are rigorously sepaated from those in the coordinate space. This approach is then applied to the H molecule. In particular, it is shown that an orbital-occupation functional is sufficient to describe the ground-state energy of a dissociated molecule. BigBird-Pegasus+Lavout We revisit the exact formulation of multi configurational density-functional theory (DFT) based on the separation of correlation effects in the orbital space. In contrast to range-separated DFT the orbitals occupation will be used as basic variable rather than the electron density. The theory is derived for a particular choice of orbitals, namely those obtained by diagonalizing the non-interacting Hamiltonian (kinetic and nuclelar potential energy operators only are considered). Illustrative results are then given for the simple two-site Hubbard model. Finally, an exact Site-Occupation Embedding Theory is derived and compared with the Density Matrix Embedding Theory of Knizia and Chan.

Figure 1: Ground-truth summary from arXiv-Lay and corresponding summaries generated by each Pegasusbased model. We manually look for concepts covered in the ground-truth that appear in the prediction, and highlight them. Best viewed in color.

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 datasets contain less training data due to the inability to process all original documents. Secondly,

<sup>&</sup>lt;sup>17</sup>The large versions of T5 and LED did not fit into GPU due to their size.
<sup>18</sup>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.

<sup>&</sup>lt;sup>19</sup>For 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	<b>41.15</b>	<b>42.05</b>

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.

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.

479

480

481

482

483 484

485

486

487

488

489

490

491

492

493

494

495

496

497

498

499

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 library<sup>20</sup> 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 we hope our work will boost research on better long-range, multimodal and multilingual models.

500

501

502

503

504

505

506

507

508

509

510

511

512

513

514

515

516

517

518

519

520

521

522

523

524

525

526

<sup>&</sup>lt;sup>20</sup>https://github.com/GregBowyer/cld2-cffi



Figure 2: 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.

Overall, results show a clear benefit of integrating layout information into long document summarization. To visualize this, we provide, in Figure 1, summaries generated by each Pegasus-based model for a document in arXiv-Lay, along with the corresponding ground-truth summary. The summary generated by BigBird-Pegasus+Layout covers more elements in the ground-truth summary than the other models: we note that it provides a more informative summary than BigBird-Pegasus, although both models are fed with the same context, which emphasizes the importance of multimodality in capturing long-term dependencies.

### 5.2 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, since longer texts are intuitively easier to understand when layout is provided, 2) summary length, as longer summaries are likely to cover more salient information, and 3) variance in the height of bounding boxes, which reflects the variance in font sizes, 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. For each dataset, we compute quartiles from the distributions of article lengths, ground-truth summary lengths, and variance in the height of bounding boxes.<sup>21</sup> Based on the aforementioned factors, the scores obtained by each model are then grouped by quartile range, and averaged over each range, see Figure 2. On arXiv-Lay, we find that layout

563

564

565

566

567

568

569

570

571

572

573

574

575

576

577

578

579

580

581

582

583

584

585

586

587

588

590

591

592

593

594

595

### 6 Limitations and Risks

The proposed corpus is limited to a single domain, that of scientific literature; such limitation arguably extends also to the diversity of documents in terms of visual appearance. In terms of risks, we acknowledge the presence of Personally Identifiable Information such as author names and affiliations; nonetheless, such information are voluntarily made public by the authors themselves and thus the proposed corpora do not bring additional downsides.

#### 7 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.

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.

<sup>528</sup> 529 530 531 534 535 536 537 541 542 543 544 545 549 550 553 554 555 556 557 558

 $<sup>^{21}</sup>$ The quartiles are provided in Section C.2 in the Appendix.

#### References

596

599

605

606

607

610

611

612 613

614

615

621

626

627

629

631

635

641

643

- 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.
- 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*.
- 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.
- 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.

651

652

653

654

655

656

657

658

659

660

661

662

663

664

665

666

667

670

671

672

673

674

675

676

677

678

679

680

681

682

683

684

685

686

687

688

689

690

691

692

693

694

695

696

698

699

700

703

- 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.
- 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 705 Lee, Sharan Narang, Michael Matena, Yanqi Zhou, 706 Wei Li, and Peter J Liu. 2019. Exploring the limits 708 of transfer learning with a unified text-to-text transformer. arXiv preprint arXiv:1910.10683. 710 Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, 711 Benjamin Piwowarski, and Jacopo Staiano. 2020. Ml-712 sum: The multilingual summarization corpus. arXiv 713 preprint arXiv:2004.14900. 714 Eva Sharma, Chen Li, and Lu Wang. 2019. Bigpatent: 715 A large-scale dataset for abstractive and coherent summarization. arXiv preprint arXiv:1906.03741. 716 717 Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In International Conference on Machine Learning, 719 pages 4596-4604. PMLR. 721 Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and An-722 gela Fan. 2020. Multilingual translation with exten-723 sible multilingual pretraining and finetuning. arXiv 724 preprint arXiv:2008.00401. 725 Elif Toprak and Gamze Almacioğlu. 2009. Three read-726 ing phases and their applications in the teaching of 727 728 english as a foreign language in reading classes with young learners. Journal of language and Linguistic 729 730 Studies, 5(1). Thomas Wolf, Lysandre Debut, Victor Sanh, Julien 731 Chaumond, Clement Delangue, Anthony Moi, Pier-732 ric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, 733 734 et al. 2019. Huggingface's transformers: State-ofthe-art natural language processing. arXiv preprint 735 arXiv:1910.03771. Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu 737 Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha 738 Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. 740 2021. LayoutLMv2: Multi-modal pre-training for 741 visually-rich document understanding. In Proceed-742

743

744

745

746

747

748

749 750

751

754

756

- ings 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.

759

760

761

762

## 767

#### 768

770

774

775

776

778

779

780

781

783

## Multimodal Dataset for Long Range and Layout-Aware Summarization – Appendix

LoRaLay: A Multilingual and

## A Datasets Construction



Figure 3: Dataset Construction Process.

## A.1 Extended Datasets – Lost Documents

Figure 4 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 4: Distribution of failure types in arXiv-Lay (top) and PubMed-Lay (bottom).

784

785

786

787

788

789

790

791

792

793

794

795

796

797

## 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 5 shows that the distribution obtained using our rule remains roughly the same as the original.



Figure 5: 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 6.

798

# 799

802

803

804

819

800

### A.4 Datasets Statistics

The distribution of research areas in HAL is provided in Figure 7. Such distributions are not available for the other datasets, as we did not have access to topic information during extraction.



Figure 7: Distribution of research areas in HAL.

#### **Experiments** B

#### Implementation Details **B.1**

Models were implemented in Python using Py-805 Torch (Paszke et al., 2017) and Hugging Face (Wolf et al., 2019) librairies. In all experiments, we use 807 Adafactor (Shazeer and Stern, 2018), a stochastic 808 optimization method based on Adam (Kingma and Ba, 2014) that reduces memory usage while retain-810 ing the empirical benefits of adaptivity. We set 811 a learning rate warmup over the first 10% steps -812 except on arXiv-Lay where it is set to 10k consis-813 tently with Zaheer et al. (2020), and use a square 814 root decay of the learning rate. All our experiments 815 have been run on four Nvidia V100 with 32GB 816 each. 817

#### С **Results** 818

#### C.1 Detailed Results

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

Table 7: 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

Table 9: ROUGE scores on KoreaScience. The best results are reported in bold.

#### C.2 Analysis of the Impact of Layout

Table 10 lists the quartiles computed from the dis-821 tributions of article lengths, summary lengths, and 822 variation in the height of bounding boxes, for arXiv-823 Lay and PubMed-Lay. 824

Model	arXi	v / arXiv	-Lay	PubM	PubMed / PubMed-Lay		
	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 6: 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.

	SciELO-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
<b>BigBird-MBART</b>	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 8: ROUGE scores on the SciELO datasets. The best results are reported in bold.

Distribution		Q1		Q2		Q3	
Distribution	arXiv-Lay	PubMed-Lay	arXiv-Lay	PubMed-Lay	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	
$\sigma$ of bounding box height	3.37	1.34	3.98	1.73	4.70	2.28	

Table 10: Quartiles calculated from the distributions of article lengths, summary lengths, and variation in the height of bounding boxes, for arXiv-Lay and PubMed-Lay.

		MINIEVIEW ACTICLE
	MITP-2004-111	Anthocyanins and Human Health: An In Vitro
Experimental Review of Pho	ton Structure Func-	investigative Approach
tion Data		Department of National Resonance of National Sciences, College of Apricalized Consumer and Invicemental Sciences, University of Hintory, Usakana, 18. 2020, USA
Max-Planck-Institut für Physik (Werner-Heisenberg-Inst chen, Germany, E-mail: Richard.Nielusfimpp.mpg.de*	inst), Föhringer Ring 6, D-80805 Min-	Businind 2 April 2000, sociand 10 May 2000, assigned 12 May 2004
DOE uil he assigned		Anthropamin pippments and accounted flavorable have resonantized ability to provide a special of human diseases, hours been matchrough difficult to study with regard to human hashs. Induces particular diseases to disease their phytochem to presentario hadingial affects, this constrbitions from individual supportant are difficult a daviphers. The complex, multicare
The present knowledge of the structure of the photon is by measurements of photon structure functions at $\sigma^*\sigma^-$	prevented based on venilet obtained sublider, Remails are prevented back to be a set of the set of	med divisitant of composition in a formalism from the angle and the angle angle and the angle and th
are also compared to recent parameterisations of the has Property of fature photon structure function measures Unsue Collider one colling	hardware constrained, where it is defined by the second s	metricine research because they are quickly, only indicat, lack inferferences found in whele featr, can be incided to periodic a and periodic accountations, and are amounted to biolabeling on that metabolic fair can be investigated after ingestion.
1 Introduction		ANTHOCYANING AND BIOMEDICINAL PROPERTIES upgreat that other mechanisms of action are also rep shifts for observed hashib benefits [2, 3, 4, 5]. Ashiber
The measurements of photon structure functions have measurements use performed by the PLUTO Collaborat	v long tradition since the first of such ion in 1981. The investigations concern	Anthocyanits are members of the favorooid group of physichenicals, a procp productionate in the new may provide protection from DNA clearage, entro- wines first-workshow cancer observed concerned and consider.
the QED structure of the photon as well as the hadronic function $F_2^{-1}(x, Q^2)$ the main areas of interest are the evolution with the momentum scale $Q^2$ , which is need	structure. For the hadronic structure behavior at for values of x and the red to COL to be invariants. The	The flavenoide, penhape the most important single group disease symptomic), enzyme inhibition, boosting proc of phanolics in fields, comprise a group of our 4000 tion of cystakase (base regulating immune respon C., accentic phase composed with multiple substrations and control trade-tradementations. Berr
experimental information is dominated by the results for This review is based on earlier work [1, 2] and as an e- forces, together with a comparison of the experiment	an the four LEP experiments. transion protides a number of updated al data with new normativizations of	parames (we're nat anda goe'hiel (bodempfinder ktrut)). ing capillary permoshility and fogility, and memb The primary players in this, group includes the as- the capitale (or c, canidat, perimoshile, primarile), the attention of the is- the capitale (or c, canidat, perimoshile, primarile).
$F_2^{ii}(x, Q^2)$ that became available since then. Only result are discussed here. The structure of virtual photons as effective structure functions are detailed in U.	on the structure of quasi-real photons d the conseponding measurements of	flavonoli (pairardin, kasinghirik), flavonos (handin, vidual anthecpanio molicedu also has a baaring on apignin), flavanose (myricarin, naningin, hosportni, dogram to shich anthecpanio exort their bioaction y naningingin), flavan-bok (acatasi, spitanchia, galanceita).
2 Structure function measuremen	ats	echic), and, although sometimes choided separately, the indianeous (guidania, databasis). Physichemicals in this means (17). The anthough although effective of the second of the class are forecaster information as indicateneous databases.
The photon can fluctuate into a fermion-anti-fermion st bers of the photon and within the limitations set by the F	ate consistent with the quantum num- tionalisty manytainty principle. These	their multificient index in human health measurements, theory interview approace in terms of bisorrable and antheorymiss in food an typically inspired as com- and binarchy poundar (7, 13), 14, 15, 16]. Original ponents of complex unknows of fluencial components - was assumed that only approace could near the ci-
Bactuations are inverse, i.e. have the magnet lifetimes, I ality. If such a fluctuation of the photon is probed, the feature, measurements of photon structure functions a	at high energytic photons of how varius- handon severals its structure. Using this s obtained from the differential encou-	that y mana is estimated twin young is in y can be in the cardinal section be interesting of the cardinal section because the section of the cardinal section because the cardinal section becard section becard section be
- section of the deep-inetisetic electron-photon scattering "favited talk presented at the Platind® Conference in Rankes 'In this paper, the term electron recompanies positions through the third paper, the term electron recompanies positions through the term electron recompanies positions are positioned in the paper.	process statuting in Fegure 1. In this y on May 12, 2008. (and)	the output of the index of the
PBOTONO	1	The validad interneting and antibilities of particular of game shows more services to appear and the system is presented as the service of the system of the service of the service of the service of the service the system of the service of the service human theorynemic targets, her, in fact, research density of the service of the servi
		© 2016 Mandani Publishing Corporation
(a) arXiv	7-Lay	(b) PubMed-Lay
Les représentations des enseignants de 2 travers le prisme des élèves en j	EP sur la relation école/famille à grande réussite scolaire	
ubilé dans la revue Cahier EBD 2017 Cahier N° 2F Envelles	Parents, Ecole	WIAD D
en vers le site <u>Education &amp; Devenir</u>		DCtc http://doi.org/10.22585/houpdomic=564.148
n vers le site <u>Les cahiers pidagogiques</u>		Tendencias temporales de los patrones de búsqueo sobre Servicios de Atención de Salud a Domicilio
s families sont des partenaires essentiels de l'école. Pourt mme problématique. Quelles représentations les enseig	ant, la relation école/tamble est vouvent décrite nants cet de cette relation et de l'influence du	antes y después del COVID-19
illeu familial sur la réussite de leurs élèves ? Nous avons ofesseurs des écoles (PE) en zone d'éducation prioritaire (	réalisé une enquête nationale auprès de 1790 2019 jouis des entretiens avec dis d'entre eux. Le en 2716 dés des interésions des définiences de	Temporal trends in Home Care Services search patterns
sme des ellèves en grande néusite scolaire (EGRS) dans l rceptions des enseignants en fonction de la néusite de l'e Neves. Ils souhaitent davantage d'implication de la neré	s zuv a rets cronsu pour éludier la différence de Neve. Les PX décrivent le profil idéal des parents des families et vourisent mettre en place une	before and after COVID-19
elle colducation qu'ils jugent indispensable à la réussite d	es áláves.	Rubán Palomo-Ulharael @ 0000-0001-1850-4337 Julia Sanaha-Sanaré @ 0000-0001-0541-8737
<b>ICS CIEN</b> présentations – enseignants – coéducation – grande réus	site scolaire - éducation prioritaire	Benjamin Palama Ulasansi ⊕ 0000-0003-3002-3553 1. Universidad Miguel Hersteinder, Departments de Salud Paldena e Hannia de la Ciencia, Sont Joan d'Alacant, Alacan Equila.
	Caroline HACHE – ADEF – AMU (saroline hachelf universa (s)	<ol> <li>Canto International Venal de Interdipación en Nariación (CVPb), Alicante, España.</li> <li>Universitat Miguel Hieradorda d'Els, Eshes, España.</li> </ol>
traduction		Annual 144 A State And
rsqu'ils étudient la proportion d'élèves de milieu popul doubler, Ould Ferhat et Terrail (2005) indique qu'un désir t la la désentation	ire ayant obtenu le baccalauriut gledral sans ort de la part des parents peut faire la différence menne dense la dischaire d'article anno 2001 une	Consegundencia/Consegundence Promotocia/Newling Ruber Rolamodinava Eas subage no ha exclude reguna Transcoador. polaron additional com
we we velves qui reussisent et ceux qui échouent. On n tégorisation des conduites des élèves lorsqu'ils font face urs parents. Les textes officiels' encouragent veux valu	numere wann is maar skille (Loffeller, ALLS) Wite aux apprentissages en fonction de l'attitude de ion positive decolfemilie, cas la familie est	Contribuciones de outorio/Aufor contributiones     Recibilo/Received     Totala la causare han contributa por qual en la mataria     20 02 2021 de este Nataria
nidérée comme un partenaire de l'école avec une place la 01). Que pensent les enseignants de ces déclarations 7 C	rportante dans la scolasté de l'élève (Houssaye, guelles sont les représentations des enseignants	Aneptody/Accepted
ncemant l'influence des familles populaires sur la réussite stre étude se propose, dans une première enquête, d'	scolaire de leur enfant ? interroger par questionnaire 1790 enseignants	Conflicto de Intervese,/Competing Intervet
école élémentaire, toutes en ZEP <sup>1</sup> , autour de leur quo quête, de réaliser des entretiens avec dix d'entre eux.	idlen dans les classes et, dans une deuxième Le sujet des parents d'élèves a pris une place	Dat calcier no presentar conflicto de interese.
sportante dans les entretiens de tous les enseignants, con 68) qui ont plus parlé « des porents que des élèves ».	me osus interrogils par Molsan et Simon (1967,	
choix du prisme des élèves en grande réussite scolaire (0) sur étudier l'avis des enseignants sur un profil particulie	RS) et donc ici, des parents de ceux-ci, a été pris , celui des familles dont les élèves révaissent	
s appelle les « néussites paradoxales » car ils néussissent ussite scolaire. Cela a permis aux enseignants de s'exprin	dans un milieu qualité de défavorable pour la ner sur la différence ou l'absence de différence	CÓMO CITAR ESTE TRABAJO   HOW TO CITE THIS PAPER
tre les parents des EGRS et les autres.		Palamodiname R, Sanchez-Torma J, Palamodiname B. Tendencias temporales de los patores de biarpeda Servicios de Alencian de Salud a Domicilio antes y despute del CCVID-19. Hosp Domic: 2021;5(4):187/95.
	1	
		Hosp Damic: 2001;5(4):187-196
(c) H	AL .	(d) SciELO-ES
neu biat (Site Paula), n.178, ad0218, 2019 http://da.doi.org/10.1160/neu-2236-9141-6-2019.146021	0.00	
		voim         Justima di nei himatori ingenera controle disposibiliti NULZI. Ne2202111           Mate-Unit-Unit-Unit-Unit-Unit-Unit-Unit-Unit
DJSSE Action	UMA LUZ SUBRE AS RELACÕES REASIL-	
perspectiva: historias speectades.	MOÇAMBIQUE NO	수도권 제조업 창업 활동의 공간적 분포 변화 _ 고가 메리ㅋ프 페이어 오 0 *
interdisciplinaridade e asses	OITOCENTOS: A	6 12 "I====- "III = 6 6 6 - A330". OAMT. OOMT.
sujeitus historices	MISSAU CUNSULAK DE IQÃO IUIZ AIROZA	1997 - 1 1997 - 19
and minimum and a margineers behavior.	(1827-1828)	Changes in Spatial Distribution of Manufacturing Startup
		A Spatial Markov Chain Approach*
		Changhyan Song* - Soonbeem Ahn** - Up Lim***
Eus Belo Horizonte, 433, apro. 603 B602D060 – Londrino – Paroné – Broail	Universidade de Sao Paulo	電壓痛管 큰 연구는 2009년까지 2019년까지를 문식의 사건적 범위로 상황하여 제조실 장실 활동이 공간적으
gubeln.gu@gnail.con	3do Poslo - 5do Poslo - Brasil	아이라한 영화를 보여했는가족을 탑성적으로도 분성하고, 항후 집업 활동이 흔드로 개인 영화를 예측하는 것을 두적으로 다. 문식을 위해 2000년부터 2018년까지의 '전국사업해포스'과 아이고프레이터 제조업 사업에 자료를 활용하였
Resumo	IIX con a Gri-Bottarba rela cual o co-	반강산업업구원의 NTAANS 문화에서에서 제시하는 40대 제조업 가운데 여러 제조업을 시켜와 세우 산업군으로 문한 후, 수도관 별별구약 충전을 수준에서 공간가 알란 분석 및 공간 데도프로 체인 문식을 수행하였다. 문식
verno de D. Pedro I concordou, em tr tráfico transatilatico de africanos para	en do recombecimento británcia, coster o o traplrio a partir de 1850, foram criadas	과에 해도면, 고위가슴산업은 및 중고위가슴산업문지 불업 활동은 시간이 조류에 하라 생기도 난하를 등심으며 중대고 있는 것으로 나타났으며, 중이하기술산업문 및 개위가슴산업문 필업 활동한 적도권 회약으로 1
representações consulares brasileiras i lidade de proteger a atuação de negrei dade do comércio de escravos sob a bi	a Artnea vertaguesa com a explicita Ina- ros brazileiros nos últimos anos de legali- inderia imperial. Nexte sentido, o presente	되고 있는 것으로 나타났다. 2000년부터 2015년까지의 추세를 연장하여 2015년까지의 관 또 변화를 예측하기 때 장업 활동이 활동이 발생하는 저희 및 그라 인접하고 있는 저역의 질문 방부 전체 실수의 가능성이 높은 것으
artigo investiga a atuação de João Luiz entre 1827 e 1828, na defesa do circuit Porte tarito o territo atual a	Airoza, cónsul do Brasil em Moçambique, o negro entre o Brasil e a África Oriental. iocircus come fuent de rativado a decumen-	나타나 공물방법 공간 효과가 흔해하는 것으로 확인되었다. 본 연구는 일자의 불율의 주요 원분이 되는 제조일 업 활동의 흔드 패턴 변화를 통해적으로 분석합으로서 작성합으로 분석합으로서 장성 유실 및 입자리 실출과 관련한 개억 전체에서 시나
tação consular produzida por Airoza e Negócios Estrangeiros.	dirigida à antiga Secretaria de Estado dos	제공하고부 하였다.
Palavras-chave Relações internacionais - Relações I	rasil-Moçambique - Missão consular -	<b>후에(6)</b> 제조업, 같임, 함께하공간하죠 문석, 공간하지. 도제 예약
Tráfico de escravos - África Oriental.		Abstract. This multi aims to replete how manufacturing error up activities from 2000 to 2018 how changed spati and to predict chances in distribution statement of finame state to activities. For the worked with Chance on FordAblam
* Todas as obras e todos os documentos =+10+=0	ss na pesquisa e na elaboração do artigo são	microdure from 2000 to 2018 www.used, and the manufacturing inductry was cheffed into four detailed induce
citados nas notas e na bibliografia. ** Doutor em História pela faculdade de Géncias I Estadual Padata bi?- de Menorie Pile.	farmanas e Sociais de Franca, da Universidade Nacionatorandos em Universida de la construcción de	* 이 논문은 300명에 대한단국 교육부터 적구선구에서 학교 전부가 제작하여 관련한구세가입니었다. 지원을 받아 주택된 연구성(3008~2 30.05.5.000309090, 전 연구는 3000년 연구석(1414 주가학을(제작하여 수수는문율을 수 일반 연구선)
totanuat raunota julio de Mesquita Ilho (Unesp). Filosofía, Letras e Ciências Barnanas, da Univers doc processo nº 2018/10798–1, Fundação de Amp	vorvansammere van notifikalit plut Backeland de dade de Salo Paulo DHELPULSEN Bolskan pår- aro å Prospiss do Estado de São Paulo (IAPESP).	** 经查询期上 国际图 上公司的本 全部分型的等于不可见于mail-changersensingPrenoticaled *** 经管理制度 电转用 化合物分子分子 化一面描述 SectoremailsHowersensingPrenoticaled ************************************
1		n - Andrew - Alderland w. Der Grund - Alder - Alder - Alder - Alder Alder Andre - Alder - Alde
(e) SciEI	JO-PT	(f) KoreaScience

