applied sciences

MDPI

Article

ESCORPIUS-M: A Massive Multilingual Crawling Corpus with a Focus on Spanish

Asier Gutiérrez-Fandiño [1], David Pérez-Fernández [2], Jordi Armengol-Estapé [3], David Griol [4,*], Ksenia Kharitonova [4] and Zoraida Callejas [4,5]

[1] LHF Labs, 48007 Bilbao, Spain; asier@lhf.ai
[2] Department of Mathematics, Universidad Autónoma de Madrid, 28049 Madrid, Spain; david.perez@inv.uam.es
[3] School of Informatics, University of Edinburgh, Edinburgh EH8 9YL, UK
[4] Department of Software Engineering, University of Granada, 18071 Granada, Spain; ksenia@ugr.es (K.K.); zoraida@ugr.es (Z.C.)
[5] Research Centre for Information and Communications Technologies (CITIC-UGR), 18071 Granada, Spain
* Correspondence: dgriol@ugr.es

Abstract: In recent years, transformer-based models have played a significant role in advancing language modeling for natural language processing. However, they require substantial amounts of data and there is a shortage of high-quality non-English corpora. Some recent initiatives have introduced multilingual datasets obtained through web crawling. However, there are notable limitations in the results for some languages, including Spanish. These datasets are either smaller compared to other languages or suffer from lower quality due to insufficient cleaning and deduplication. In this paper, we present ESCORPIUS-M, a multilingual corpus extracted from around 1 petabyte of Common Crawl data. It is the most extensive corpus for some languages with such a level of high-quality content extraction, cleanliness, and deduplication. Our data curation process involves an efficient cleaning pipeline and various deduplication methods that maintain the integrity of document and paragraph boundaries. We also ensure compliance with EU regulations by retaining both the source web page URL and the WARC shared origin URL.

Keywords: corpus; dataset; massive; multilingual; crawling; common crawl; Spanish; NLP

check for updates

Citation: Gutiérrez-Fandiño, A.; Pérez-Fernández, D.; Armengol-Estapé, J.; Griol, D.; Kharitonova, K.; Callejas, Z. ESCORPIUS-M: A Massive Multilingual Crawling Corpus with a Focus on Spanish. *Appl. Sci.* **2023**, *13*, 12155. https://doi.org/10.3390/app132212155

Academic Editor: Douglas O'Shaughnessy

Received: 16 September 2023
Revised: 28 October 2023
Accepted: 6 November 2023
Published: 8 November 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Deep Learning (DL) models in Natural Language Processing (NLP) have achieved unseen performance and thus have largely replaced conventional machine learning approaches across multiple applications, such as machine translation, natural language understanding, or natural language generation. One of the main tasks in these areas is language modeling, for which the current state-of-the-art is to use models pre-trained on a data-rich tasks or languages that are subsequently fine-tuned to the target language or task at hand, so that stakeholders do not have to perform expensive pre-training themselves.

To tackle these challenges, in this paper we present the ESCORPIUS-M multilingual corpus with the following properties:

- It is cleaner than state-of-the-art corpora and deduplicated.
- It maintains both document and paragraph boundaries allowing language models to deal with textual data in the same way as humans do, thus unlocking the capabilities of Natural Language Generation to understand paragraph representation.
- The data downloaded maintains the traceability of the origin of each document. This level of traceability makes it possible to apply the right of withdrawal of individual website owners or individual persons whose data are cited on websites and are protected by GDPR. It also allows for systematically excluding blacklisted websites.

- It is a high-quality multilingual corpus that excels in content cleaning and deduplication. Depending on the language, e.g., in Spanish, it is the largest web corpus of this quality available for the development of large language models.

This paper is an extension of our prior work with the Spanish massive corpus ESCORPIUS [1]. In this paper, we present the ESCORPIUS-M multilingual corpus that comprises 34 languages different from English. This study aims to provide the scientific community with an extensive, multilingual corpus for training large language models. Our corpus stands out for its superior cleanliness and reduced duplication when compared to similar-sized corpora. Furthermore, we present the innovative pipeline employed to ensure its high quality.

The rest of the paper is organized as follows. Section 2 presents the state-of-the-art providing the context for ESCORPIUS-M, which is described in detail in Section 3. Section 4 describes the process followed to download, clean, and deduplicate the data and Section 5 details the technical environment employed to accomplish these tasks. Finally, Section 6 provides the conclusions drawn from our research and outlines avenues for future work.

2. State-of-the-Art

Foundational models refer to large-scale language models that serve as the basis or foundation for various downstream applications and tasks [2–4]. They have become a fundamental building block for a wide range of AI applications covering natural language understanding (text classification including sentiment analysis, spam detection and topic categorization, named entity recognition, language translation, etc.) [5–8], text generation (content creation, code generation for programming languages, etc.) [9,10], question answering, conversational AI [11,12], language summarization [13,14], content recommendation and moderation [15,16], search engines, web pages and documents ranking [17], and data extraction and knowledge graph creation [18,19].

Application domains include key areas such as education (e.g., assistants, facilitators, virtual tutors and assessment aids [20,21]), healthcare (medical education, medical histories summarization, patient data analysis, conversational agents for patients, preliminary diagnosis, or clinical skills assessment [22]), legal and compliance (legal research, document generation, legal information providing, legal analysis [23]), e-government (assisted citizen service, interaction with public, FAQs, gathering feedback, accessibility to information, personalized support to groups, data analysis, summaries, automated decision-making, improving cybersecurity [24]), marketing and financial analysis (risk analysis, generating reports, interactive data analysis, generating summaries and financial news briefs, automated customer interaction [25–27]), and accessibility (text-based communication, multilingual support, 24/7 access, customization and personalization, integration with assistive technologies [28]), among others.

These models are initially pre-trained on vast and diverse text corpora to learn large language patterns, grammar, syntax, and world knowledge. This pre-training phase helps the model acquire a broad understanding of language. After pre-training, these models can be fine-tuned on specific datasets or tasks to make them more specialized. Fine-tuning adapts the model's knowledge to perform specific tasks such as text generation, translation, sentiment analysis, and more.

Foundational models are designed to generalize well across a wide range of natural language understanding and generation tasks. These models are often very large, with billions or even trillions of parameters, which allows them to capture intricate language nuances and patterns. However, their size also requires significant computational resources for training and inference. Transfer learning from these models has become a standard practice in many NLP applications.

Therefore, clean and well-curated text data is crucial for LLMs' training, and their performance hinges on the quality and volume of this data. Modern models are trained on hundreds of billions of tokens, and each developer adopts distinct training methods including their own recipes for creating the datasets [29–31]. As commercial interests in

LLMs grow, there is a trend towards proprietary datasets [32], underscoring the importance of accessible, clean data.

Despite the positive results reported in the literature for English pre-trained models, many voices highlight the disparity in the availability and quality of models and data for other languages. As noted in ref. [33], most NLP research is conducted in English, followed by Mandarin Chinese, and at a great distance to other languages, including Spanish despite its large number of speakers around the world. This situation has negative repercussions on the development of fair NLP technology for all [34], e.g., disparate access to clinical NLP for speakers of different languages [35].

To address this issue, language models are being trained on monolingual corpora in different languages. For example Multilingual BERT (https://github.com/google-research/bert/blob/master/multilingual.md, accessed on 15 September 2023), which has been recently outperformed by mT5 [36], a massively multilingual model trained with mC4, a dataset of text in 101 languages. However, the quality of mC4 is not as good for non-English languages (according to one of the authors: https://github.com/allenai/allennlp/discussions/5265#discussioncomment-2596110, accessed on 15 September 2023), which is making some researchers produce clean excerpts of mC4 corresponding to their language of interest, see e.g., the clean Italian mC4 [37]. European researchers recently collaborated to train the open-source multilingual model, Bloom [38], and in this endeavor, they created the multilingual dataset ROOTS [39] to address the limitations of previous multilingual corpora.

As warned in ref. [40], low-quality data can have pernicious effects, not only in the quality of the results produced, but also because it may lead to the false idea that languages different from English are represented well enough with sufficient high-quality resources.

There also exist parallel corpora, which exploit text sources that are produced in several languages (e.g., translated legal documents or subtitles for the same videos in different languages). The texts are then aligned and processed to produce parallel sentences in the different languages. For instance, the CCaligned corpus contains web-document pairs in 137 languages obtained by identifying URLs that are translations of the same page [41]. However, a recent evaluation of the main available corpora has shown that in some cases, up to two-thirds of the audited samples were misaligned [40]. Also, the reliance on sources that are produced in several languages makes it difficult to find more spontaneous texts in the dataset.

3. esCorpius-m at a Glance

A total of 39,502 compressed WARC (Web Archive) files were processed which represent two months of Common Crawl (see Section 4.3 for more details). The compressed information occupied about 180 TB and the size of the processed decompressed information is estimated to be more than 0.8 PB. Prior to content deduplication, the downloaded corpus was composed of 549,887,283,621 words. The deduplicated and cleaned corpus size is 2,710,379,754,463 (2.5 TB), with 645,772,362 paragraphs and 242,248,582,193 tokens.

esCorpius-m supports 34 languages, with Spanish being the most dominant, covering nearly 20% of the entire corpus (see Figure 1).

While languages like French, German, Spanish, and Portuguese make up a significant portion, it also features low-resource languages like Basque, Galician, Occitan, and Maltese. Table 1 shows the figures per language. The Data Availability Section contains the link to the resource.

As argued in ref. [42], corpora for language modeling should not only be compared in terms of size, but also in the quality and traceability of their data. Table 2 shows a comparison of esCorpius-m with the main state-of-the-art multilingual datasets. For our comparison, we focused on massive crawled datasets consisting of hundreds of billions of tokens, similar in size to our corpus with 242 billion tokens. The only deviation from this criterion was the ParaCrawl corpus (we include ParaCrawl as a reference, but regarding the size, note that the comparison is apples-to-oranges because unlike the other datasets

shown in the table, Paracrawl is a parallel dataset, focusing on machine translation, rather than language modeling or similar tasks), given its exceptional processing quality. The comparison has been performed not only in terms of size, but also encompassing quality factors related to language identification, parsing, cleaning and deduplication that will be further explained in the following section. For multilanguage datasets, the numbers reported in the table correspond to non-English samples only.

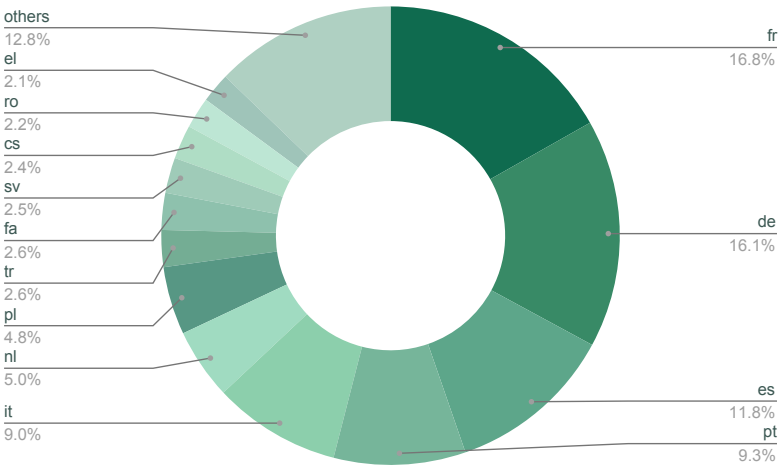

Figure 1. Distribution of the languages in ESCORPIUS-M by the amount of words.

Table 1. Data per language.

Lang	000 Words	% Words	Par.	% Par.	URLs	% URLs	Size, GB
Afrikaans	114,386	0.05%	248,588	0.04%	245,361	0.04%	0.66
Arabic	4,542,747	1.88%	11,502,118	1.78%	11,403,653	1.79%	49.38
Bengali	594,746	0.25%	1,772,168	0.27%	1,759,258	0.28%	10.44
Catalan	1,680,554	0.69%	4,284,590	0.66%	4,227,541	0.66%	10.95
Czech	5,764,676	2.38%	16,978,203	2.63%	16,667,666	2.62%	44.12
Danish	3,155,961	1.30%	7,551,657	1.17%	7,410,853	1.17%	21.15
German	38,899,553	16.06%	109,705,164	16.99%	107,381,398	16.88%	297.11
Greek	5,151,481	2.13%	12,042,049	1.86%	11,923,822	1.87%	61.87
Spanish	28,618,504	11.81%	57,808,392	8.95%	57,327,087	9.01%	195.7
Basque	116,918	0.05%	420,725	0.07%	414,677	0.07%	0.99
Persian	6,171,391	2.55%	12,631,282	1.96%	12,475,782	1.96%	155.13
Finnish	2,793,619	1.15%	7,945,694	1.23%	7,805,202	1.23%	25.68
French	40,739,063	16.82%	94,895,680	14.69%	92,999,095	14.62%	298.24
Galician	223,711	0.09%	603,920	0.09%	595,481	0.09%	1.5
Hindi	1,326,133	0.55%	2,735,525	0.42%	2,714,560	0.43%	16.93
Croatian	1,222,123	0.50%	4,067,720	0.63%	4,026,201	0.63%	8.93
Italian	21,879,938	9.03%	50,844,721	7.87%	50,094,971	7.88%	149.39
Japanese	913,584	0.38%	23,856,552	3.69%	23,593,379	3.71%	111.23
Korean	1,711,686	0.71%	3,810,567	0.59%	3,782,029	0.59%	18.45
Maltese	24,340	0.01%	51,934	0.01%	51,457	0.01%	0.2
Dutch	12,015,439	4.96%	32,497,299	5.03%	31,779,475	5.00%	80.77
Norwegian	2,361,749	0.97%	5,725,188	0.89%	5,646,580	0.89%	15.59
Occitan	7371	0.00%	20,631	0.00%	20,323	0.00%	0.05
Punjabi	54,496	0.02%	104,737	0.02%	103,879	0.02%	0.7
Polish	11,731,521	4.84%	33,017,904	5.11%	32,436,496	5.10%	92.34
Portuguese	22,577,860	9.32%	50,055,128	7.75%	49,535,350	7.79%	149.34
Romanian	5,351,974	2.21%	11,995,288	1.86%	11,851,961	1.86%	36.29
Slovenian	903,213	0.37%	2,296,724	0.36%	2,262,941	0.36%	6.36
Serbian	445,014	0.18%	907,411	0.14%	894,739	0.14%	5.2
Swedish	6,135,501	2.53%	14,777,204	2.29%	14,528,724	2.28%	48.93
Turkish	6,393,599	2.64%	21,232,882	3.29%	21,032,030	3.31%	56.65
Ukrainian	3,483,220	1.44%	9,066,207	1.40%	8,958,138	1.41%	46.51
Urdu	335,909	0.14%	568,139	0.09%	563,963	0.09%	2.81
Chinese	4,806,600	1.98%	39,750,371	6.16%	39,533,316	6.22%	504.62
Total	**242,248,582**	**100.00%**	**645,772,362**	**100.00%**	**636,047,388**	**100.00%**	**2524.24**

Table 2. Comparison of the main state-of-the-art multilingual corpora (excluding English excerpts).

	OSCAR 22.01 [43]	**mC4 [36]**	**CC-100 [44]**	**ROOTS [39]**	**ParaCrawl v9 [45]**	**esCorpius-m (Ours)**
Size	5.2 TB	3.5 TB	2.0 TB	1.2 TB	24 GB	2.5 TB
Docs	560M	-	-	521M	-	1427M
Words	381B	3567B	284B	189B	8B	242B
WARC/WET	WET	WET	WET	WET	WARC	WARC
Lang. identification	fastText	CLD3	fastText	fastText	CLD2	CLD2 + fastText
Content identification	WET heuristic	WET heuristic	WET heuristic	WET heuristic	Sentence Alignment	JustText (modified)
Elements	Document	Document	Document	Document	Sentence	Document and paragraph
Parsing quality	Medium	Low	Medium	High	High	High
Cleaning quality	Low	No cleaning	Low	High	High	High
Deduplication	No	No	No	SimHash + SuffixArray	Bicleaner	dLHF
If parallel	-	-	-	-	✓	-
Traceability	URL	URL	URL	URL	URL	URL + WARC
Licence	CC BY 4.0	OCD-BY 1.0	Common Crawl	CC-BY-SA-4.0	CC0	CC-BY-NC-ND 4.0

The enhancements incorporated in esCorpius-m compared to other similar, large, crawled corpora can be summarized as follows:

1. Contrary to most other corpora—with the exception of ParaCrawl [45]—our content is extracted directly from WARC files. This method is notably more reliable than sourcing from the frequently error-prone WET files.
2. Language identification is a two-step process: first, using the less computationally-intensive cld2, followed by fastText for paragraph candidates. The latter is recognized for its superior quality (As can be demonstrated in https://modelpredict.com/language-identification-survey, accessed on 15 September 2023). Other corpora only employ a segment of this language identification pipeline, missing out on the combined strengths of the entire process.
3. The process of main content identification utilizes a version of JustText, a modified version of high-quality tool (as demonstrated here: https://trafilatura.readthedocs.io/en/latest/evaluation.html, accessed on 15 September 2023), while the other corpora rely solely on the less detailed set of WET heuristics.
4. Our deduplication process implemented is both robust and precise. Compared to ROOTS [38] we implement the deduplication not only at the document but also at the paragraph level, carrying out both exact and soft deduplication.
5. Our system provides enhanced traceability, going beyond just URL tracking found in other corpora. It includes tracing by WARC segment location, allowing users to seamlessly trace textual segments back to their origins in Common Crawl and identify any undesired content.

These improvements are explained in greater detail in the following sections.

4. Data Download and Cleaning Process

To generate esCorpius-m, we have processed WARC files from Common Crawl. We have ideated and implemented a novel cleaning pipeline that allows for obtaining high-quality data, which we have applied to the WARC files in order to obtain a clean dataset.

4.1. Common Crawl Repository

Common Crawl is a web archive that contains petabytes (see https://commoncrawl.github.io/cc-crawl-statistics/plots/crawlsize (accessed on 15 September 2023) to explore current crawling sizes) of data collected since 2008. The repository contains raw web page data in the WARC file format, request/response metadata files in WAT format, and text data contents extracted from WARC and stored in WET files.

Some corpora such as mC4 use WET files to generate the corpus content. However, this has several disadvantages. The main problem with WET is that the process followed

to remove HTML tags and extract the text is error-prone. After HTML tags are removed, frequently the text is either divided into unconnected pieces or merged with other unrelated textual information. For example, the sentence "We offer fast transportation" could be divided into {"We offer", "fast", and "transportation"}, thus loosing the integrity of the original text.

Additionally, the text extractions from WET files are HTML-content agnostic in the sense it is not checked whether the content belongs to less relevant parts of the web (e.g., a menu or a footer).

Hence, for our work, we decided to use the original WARC files as some transformations made for creating the WETs are irreversible or, at least, very costly to repair. A quick look at the same example in WARC and WET format shows this difference (same example in WET: https://gist.githubusercontent.com/Smerity/e750f0ef0ab9aa366558/raw/313b675a01ee1d1f05829439165a9eb991571547/bbc.wet (accessed on 15 September 2023) and WARC: https://gist.githubusercontent.com/Smerity/e750f0ef0ab9aa366558/raw/313b675a01ee1d1f05829439165a9eb991571547/bbc.warc (accessed on 15 September 2023) formats).

The reason why the related work utilizes WET [29,46] is that in WET format there is a close relation between the file size and the amount of textual data it contains: nearly 100% of the contents of WET files is pure text, thus optimizing the resources invested in processing the files. On the contrary, the textual data and WARC file size ratio is low, and WARC files require parsing to obtain the text, which entails increased computing times and resources.

4.2. WARC Files and Archiving Standard

Common Crawl distributes contents in different folders (prefixes in Amazon S3 terminology) at a rate of one folder per month since August 2014 (prior to 2014, several months were stored in the same folder). For each month, a growing number of WARC files are generated, currently reaching more than 72,000 (according to https://commoncrawl.org/the-data/get-started/, accessed on 15 September 2023).

The size of the WARC files is variable: they are published compressed in Gzip compression format to save space in the repository with a size that, since January 2015, ranges from 0.9 GB to 1.1 GB. The compression ratios observed in the files are between 4:1 and 5:1, so the final size of these files, once decompressed, ranges from 3.9 GB to 5.2 GB.

WARC is an extension of the ARC file format (ARC) used to store "web crawls" as sequences of content blocks. The WARC (Web ARChive) file format concatenates multiple resource records (data objects), consisting of simple text headers and data blocks into a long file. The WARC format is a preservation format defined by the International Internet Preservation Consortium (IIPC). The WARC format offers a standardized way to manage billions of web-collected resources (see the following web page for more information on this standard file format: https://www.iso.org/obp/ui/#iso:std:iso:28500:ed-2:v1:en, accessed on 15 September 2023).

4.3. Common Crawl Subcorpus Selection and Cleaning

A Common Crawl Subcorpus of WARC files from the period 2015–2022 has been selected to guarantee stability in the file format and content encoding. The compressed information occupied about 180 TB and the size of the processed decompressed information is estimated to be more than 0.8 PB. Each WARC file is usually divided on 100 segment files. A total of 39,502 segments of compressed WARC files were processed, so we have computed an equivalent of two months on Common Crawl.

For each one of the CPU cores employed in the generation of esCorpius-m, the following protocol was performed:

1. Download a WARC from Common Crawl.
2. Open a Gzip file reader.
3. While reading the Gzip file, partially parse the WARC format.
4. Parse the webpage and fix the encoding to UTF-8.

5. Obtain the language used in the document (see Section 4.4). Proceed if the language is not English.
6. Extract the text that is correct.
7. Store the record in the format described in Section 4.6.

In order to avoid obtaining too much duplicated content and/or content which is very similar (e.g., COVID-19 content) we randomized the WARC URL order that is fed to the cleaning cluster. We also performed a deduplication process over the data obtained that is described in Section 4.7.

4.4. Language Detection Process

After extracting the text from Common Crawl, it was necessary to discard English text as we were creating the non-English corpora. Language detection is a very important part of the pipeline to produce accurate results. However, many times the language identification methods used are not robust and text in languages different from the target one appear in the datasets. For example, in the case of Spanish, it can be easily confused with other romance languages.

To avoid such mistakes and produce more robust results, we carried out language identification in two steps. First, a quick filtering based on the Compact Language Detector 2 (CLD2) tool (https://github.com/CLD2Owners/cld2, accessed on 15 September 2023) was performed. Secondly, we used the fastText tool (https://github.com/facebookresearch/fastText, accessed on 15 September 2023) [47] which requires larger computational resources in order to verify the language identification made by CLD2. In the creation of the corpus, the criterion of quality prevailed over that of quantity. We have sacrificed corpus length and processing time in exchange for greater certainty that the language of the text is not English.

4.5. Main Content Identification

Extracting information from web pages is a challenging task given the vast variety of visual formats and communication platforms (forums, blogs, etc.). For the generation of an extensive corpus, we sought to extract the core contents, eliminating headers, footers, tables, including potential titles, and (optionally) comments. This task is also known as boilerplate removal, main content identification, or web-page cleaning.

For the corpus generation, we used a derivative of JustText (https://nlp.fi.muni.cz/projects/justext/, accessed on 15 September 2023) due to its performance/quality ratio (https://trafilatura.readthedocs.io/en/latest/evaluation.html, accessed on 15 September 2023). The choice of this method for main content identification has disadvantages such as higher computational cost and a significant reduction in the amount of content downloaded. However, its main strength lies in the quality of the paragraphs obtained.

4.6. Output Storage Format

The output format of this process and the format in which the corpus is distributed is JSONL. For each line of the corpus, there is a separate JSON document representing a separate paragraph of the corpus with the following fields:

- `id`: unique document identifiers UUIDv4 (the complete specification of UUIDv4 can be found in https://www.ietf.org/rfc/rfc4122.txt, accessed on 15 September 2023) over the whole corpus.
- `text`: textual content of the paragraph.
- `url_warc`: this is the identifier of the WARC file from which the web page from which the text has been extracted following the Common Crawl segments nomenclature ("`s3://commoncrawl/crawl-data/CC-MAIN-<YYYY>-<MM>/segments/<id>/warc/CC-MAIN-<id>.warc.gz`", where YYYY is the 4 digits year and MM the WARC archive month).
- `url`: URL address from which the text has been extracted.

A sample corpus JSON register is shown below. Note that it is a paragraph from the article that currently does not exist although it is archived on Common Crawl.

```
{
"id":"8280bafd-5984-4a5e-8436-af56a474d9cd",
"text":"<textual content>",
"url_warc":"s3://commoncrawl/crawl-data/CC-MAIN-2019-04/segments/
1547583730728.68/warc/
CC-MAIN-20190120184253-20190120210253-00091.warc.gz",
"url":"http://alertatierra.com/continente/61-noticias/volcanes/
2135-erupcion-de-turrialba-y-rincon-de-la-vieja-en-costa-rica"
}
```


Consequently, esCorpius-m is fully traceable as including *url_warc* facilitates the retrieval of URLs from Common Crawl, ensuring the reproducibility of the study. Additionally, this traceability simplifies source-based dataset filtering.

4.7. Deduplication

As described in ref. [48], it is crucial to generate datasets for training language models that are free of duplicity. This makes it possible to train models that do not memorize sentences due to their high degree of duplicity, which artificially results in fewer training steps and higher accuracy. Using deduplication, the overlap between training, validation and test sets is reduced, improving the training procedure as well as the confidence on the model proficiency.

We do not perform any content filtering process based on URLs as we want to avoid any kind of censorship. We expect users to perform this filtering if they are interested in it (we encourage users to understand the license terms before doing sot). Additionally, depending on the purpose of the model, the corpus can be filtered out to show only the results from specific URLs.

Our novel deduplicacion process, which we have named dLHF, is comprised of two main steps. First of all, we deduplicate complete contents of the corpus via exact matching at the document level. Subsequently, we perform the same deduplication at the paragraph level. In order to deal with paragraphs, the text is normalized and noisy information is removed. Even though the process may appear naïve, the greatest part of the deduplication happens in this step.

After that, we perform a more complex deduplication based on Local Sensitive Hashing. We adapted state-of-the-art code to work in parallel, use fewer computational resources and, more importantly, to avoid unstructuring the document. This last feature is highly relevant as we have found that some of the deduplication software available performs operations that break the document integrity. Actually, the fact that the OSCAR corpus suffers this issue, leads us to think that it was not subject to deduplication whatsoever (see the last point of the section "Changes" at the following URL for more details: https://oscar-corpus.com/post/oscar-v22-01/, accessed on 15 September 2023).

In related work performed in ROOTS [38], deduplication is implemented at the document level, with the use of hashing (SimHash) for soft deduplication. Due to the bag-of-words characteristics of SimHash, an extra step was needed to handle longer documents as they tend to appear similar. In contrast, ParaCrawl implements deduplication at the sentence level. Rather than considering entire documents, they look at individual sentences within those documents to identify duplicates, making it a non-comparable approach [45].

5. Technical Environment Used for Corpus Generation

The selected massively parallel processing infrastructure relies on open source solutions for data analytics, Apache Hadoop and Apache Spark framework, integrated in the Amazon Elastic Map Reduce (EMR) service of Amazon Web Services (AWS). In particular, the following application stack has been used:

- `Hadoop File System (HDFS)`, a distributed and scalable file system, was used as an auxiliary repository in the tasks executed in EMR, which is built with part of the cluster nodes (Core Nodes).

- `Amazon Simple Storage Service (S3)`, in particular EMR FS, which is a Hadoop reimplementation of AWS object storage.
- `Apache Spark`, this open source engine was used for the parallelisation of data cleaning tasks. The parallel WARCs processor is based on PySpark.
- `Apache YARN` was used as resource manager used for scheduling and monitoring the execution of tasks in the EMR cluster.
- `Ganglia` has been used for monitoring the cluster status. Cluster monitoring is important in the early stages, for node load visualization and selecting the number and type of nodes used in the cluster.

For the EMR cluster, we used three types of nodes: one Master Node, which is the orchestrator of the cluster and on which YARN resource manager runs; three Core Nodes, which support the local HDFS file system; and a large collection of Task Nodes, which are nodes dedicated exclusively to the execution of tasks. All of these nodes are supported by virtual compute instances (similar to the concept of a virtual machine in a AWS data center environment) thanks to the Amazon Elastic Compute Cloud (EC2) service.

Amazon EMR supports different types of EC2 instances depending on the characteristics of the task to be executed. For the corpus generation process, general purpose instances have been considered, although instances with GPU and advanced networking capabilities are also available. In particular, we used the following EC2 instance types:

- `M5 instances` (https://aws.amazon.com/ec2/instance-types/m5/, accessed on 15 September 2023): general-purpose instances powered by Intel Xeon Platinum 8000 series processors up to 3.1 GHz. They have a network bandwidth ranging from 10 to 25 GBPS depending on the selected size.
- `R5 instances` (https://aws.amazon.com/ec2/instance-types/r5/, accessed on 15 September 2023): memory-optimized instances, with the same type of processors as the M5 instances but with a vCPU:RAM ratio of 1:8, allowing more memory-demanding tasks to be executed.

Textual Content Deduplication Infrastructure

Due to the structure of the problem and the large memory requirements, the paradigm of the Hadoop cluster was not a good fit for the deduplication process.

For our deduplication tools, we used a potent machine with an Intel Xeon Platinum 8176 2.10 GHz processor (112 processor threads) and 1.5 TB of RAM. The importance of using such amount of RAM lies in the possibility to deduplicate the corpus corresponding to each language in one go (we could deduplicate the whole corpus efficiently in less than 3 h). Usually, corpora processed in infrastructures with smaller capacities are split into smaller chunks that are deduplicated independently, then merged, and split again in a iterative process, which is usually stopped before complete deduplication is achieved.

6. Conclusions and Future Work

This paper presents esCORPIUS-M, a massive cleaned multilingual web-crawling corpus, which has been produced by means of a novel and effective approach to produce high-quality deduplicated corpora from Common Crawl. The collected data fully respects the document and paragraph boundaries in order for language models to be more accurate. The document source URLs allow full traceability of the data which permits creating domain-specific language models, indexing data, complying with EU regulations (such as the right to be forgotten).

As for future work, we suggest increasing the period for extracting the corpus as, according to our calculations, there is the chance to still extract a corpus 200 times bigger. As this corpus is only a crawling-based corpus, we propose to extend it and create a multilingual corpus similar to The Pile [49].

We have shared the corpus in HuggingFace with the hope that its potential users can contribute to the advancement of NLP technologies in non-English languages by performing different analyses of the corpus (e.g., topic modeling) and creating new language

models and embeddings for the community. In addition, it would be also interesting to extend the crawling to other languages such as Basque using the same methodology, and analyzing the usefulness of this corpus in downstream applications.

Author Contributions: Conceptualization, A.G.-F., D.P.-F., J.A.-E., D.G., K.K. and Z.C.; methodology, A.G.-F., D.P.-F. and J.A.-E.; software, A.G.-F. and D.P.-F.; resources, A.G.-F., D.P.-F., J.A.-E. and D.G.; data curation, A.G.-F., D.P.-F. and J.A.-E.; writing—original draft preparation, A.G.-F., D.P.-F., D.G., K.K. and Z.C.; writing—review and editing, A.G.-F., D.P.-F., D.G., K.K. and Z.C.; project administration, D.G. and Z.C. All authors have read and agreed to the published version of the manuscript.

Funding: This publication is part of the project "CONVERSA: Effective and efficient resources and models for transformative conversational AI in Spanish and co-official languages" (TED2021-132470B-I00) funded by MCIN/AEI/10.13039/501100011033 and by the European Union "NextGenerationEU/PRTR".

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: esCorpius-m has been shared under CC-BY-NC-ND 4.0 (https://creativecommons.org/licenses/by-nc-nd/4.0/, accessed on 15 September 2023), license in HuggingFace: https://huggingface.co/datasets/LHF/escorpius-m, accessed on 15 September 2023). This license allows reusers to copy and distribute the material in any medium or format in unadapted form only, for noncommercial purposes only, and only so long as attribution is given to the creator.

Acknowledgments: We want to thank Amazon Web Services Spain, specially, to José (Pepe) López Rodríguez and Alberto González Dueñas for their help on setting up the cluster and managing the communication with the Common Crawl AWS team. This work could not have been possible without their invaluable help.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers
DL	Deep Learning
GDPR	General Data Protection Regulation
LLM	Large Language Model
NLP	Natural Language Processing
WARC	Web ARChive

References

1. Gutiérrez-Fandiño, A.; Pérez-Fernández, D.; Armengol-Estapé, J.; Griol, D.; Callejas, Z. esCorpius: A Massive Spanish Crawling Corpus. In Proceedings of the IberSPEECH 2022 Conference, Granada, Spain, 14–16 November 2022; pp. 126–130. [CrossRef]
2. Bommasani, R.; Hudson, D.A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M.S.; Bohg, J.; Bosselut, A.; Brunskill, E.; et al. On the Opportunities and Risks of Foundation Models. *arXiv* **2022**, arXiv:2108.07258.
3. Khan, W.; Daud, A.; Khan, K.; Muhammad, S.; Haq, R. Exploring the frontiers of deep learning and natural language processing: A comprehensive overview of key challenges and emerging trends. *Nat. Lang. Process. J.* **2023**, *4*, 100026. [CrossRef]
4. OECD. *AI Language Models: Technological, Socio-Economic and Policy Considerations*; OECD Publishing: Paris, France, 2023; pp. 20–28. [CrossRef]
5. Rafiepour, M.; Sartakhti, J.S. CTRAN: CNN-Transformer-based network for natural language understanding. *Eng. Appl. Artif. Intell.* **2023**, *126*, 107013. [CrossRef]
6. Li, B.; Weng, Y.; Xia, F.; Deng, H. Towards better Chinese-centric neural machine translation for low-resource languages. *Comput. Speech Lang.* **2023**, *84*, 101566. [CrossRef]
7. Li, R.; Liu, C.; Jiang, D. Efficient dynamic feature adaptation for cross language sentiment analysis with biased adversarial training. *Knowl.-Based Syst.* **2023**, *279*, 110957. [CrossRef]
8. Park, J.; Cho, S. Incorporation of company-related factual knowledge into pre-trained language models for stock-related spam tweet filtering. *Expert Syst. Appl.* **2023**, *234*, 121021. [CrossRef]
9. López Espejel, J.; Yahaya Alassan, M.S.; Chouham, E.M.; Dahhane, W.; Ettifouri, E.H. A comprehensive review of State-of-The-Art methods for Java code generation from Natural Language Text. *Nat. Lang. Process. J.* **2023**, *3*, 100013. [CrossRef]

10. Goswamy, T.; Singh, I.; Barkati, A.; Modi, A. Adapting a Language Model for Controlled Affective Text Generation. In Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain, 8–13 December 2020; pp. 2787–2801.
11. Abro, W.A.; Aicher, A.; Rach, N.; Ultes, S.; Minker, W.; Qi, G. Natural language understanding for argumentative dialogue systems in the opinion building domain. *Knowl.-Based Syst.* **2022**, *242*, 108318. [CrossRef]
12. McTear, M. *Conversational AI. Dialogue Systems, Conversational Agents, and Chatbots*; Morgan and Claypool Publishers: San Rafael, CA, USA, 2020. [CrossRef]
13. Abdelfattah Saleh, A.; Weigang, L. TxLASM: A novel language agnostic summarization model for text documents. *Expert Syst. Appl.* **2024**, *237*, 121433. [CrossRef]
14. Xie, Q.; Bishop, J.A.; Tiwari, P.; Ananiadou, S. Pre-trained language models with domain knowledge for biomedical extractive summarization. *Knowl.-Based Syst.* **2022**, *252*, 109460. [CrossRef]
15. Bansal, S.; Gowda, K.; Kumar, N. Multilingual personalized hashtag recommendation for low resource Indic languages using graph-based deep neural network. *Expert Syst. Appl.* **2024**, *236*, 121188. [CrossRef]
16. Franco, M.; Gaggi, O.; Palazzi, C.E. Analyzing the Use of Large Language Models for Content Moderation with ChatGPT Examples. In Proceedings of the 3rd International Workshop on Open Challenges in Online Social Networks (OASIS'23), Rome, Italy, 4–8 September 2023. [CrossRef]
17. Habernal, I.; Konopík, M. SWSNL: Semantic Web Search Using Natural Language. *Expert Syst. Appl.* **2013**, *40*, 3649–3664. [CrossRef]
18. Hao, S.; Tan, B.; Tang, K.; Ni, B.; Shao, X.; Zhang, H.; Xing, E.; Hu, Z. BertNet: Harvesting Knowledge Graphs with Arbitrary Relations from Pretrained Language Models. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2023, Toronto, ON, Canada, 9–14 July 2023; pp. 5000–5015.
19. Wang, C.; Liu, X.; Song, D. Language Models are Open Knowledge Graphs. *arXiv* **2020**, arXiv:2010.11967.
20. Kasneci, E.; Sessler, K.; Küchemann, S.; Bannert, M.; Dementieva, D.; Fischer, F.; Gasser, U.; Groh, G.; Günnemann, S.; Hüllermeier, E.; et al. ChatGPT for good? On opportunities and challenges of large language models for education. *Learn. Individ. Differ.* **2023**, *103*, 102274. [CrossRef]
21. Arnau-González, P.; Arevalillo-Herráez, M.; Luise, R.A.D.; Arnau, D. A methodological approach to enable natural language interaction in an Intelligent Tutoring System. *Comput. Speech Lang.* **2023**, *81*, 101516. [CrossRef]
22. Xiao, D.; Meyers, P.; Upperman, J.S.; Robinson, J.R. Revolutionizing Healthcare with ChatGPT: An Early Exploration of an AI Language Model's Impact on Medicine at Large and its Role in Pediatric Surgery. *J. Pediatr. Surg.* **2023**, *58*, 2410–2415. [CrossRef] [PubMed]
23. Sukanya, G.; Priyadarshini, J. Modified Hierarchical-Attention Network model for legal judgment predictions. *Data Knowl. Eng.* **2023**, *147*, 102203. [CrossRef]
24. Peña, A.; Morales, A.; Fierrez, J.; Serna, I.; Ortega-Garcia, J.; Puente, Í.; Córdova, J.; Córdova, G. Leveraging Large Language Models for Topic Classification in the Domain of Public Affairs. In Proceedings of the Document Analysis and Recognition Conference—ICDAR 2023 Workshops, San Jose, CA, USA, 21–26 August 2023; pp. 20–33. [CrossRef]
25. Jansen, B.J.; Gyo Jung, S.; Salminen, J. Employing large language models in survey research. *Nat. Lang. Process. J.* **2023**, *4*, 100020. [CrossRef]
26. Suzuki, M.; Sakaji, H.; Hirano, M.; Izumi, K. Constructing and analyzing domain-specific language model for financial text mining. *Inf. Process. Manag.* **2023**, *60*, 103194. [CrossRef]
27. Liu, S.; Peng, C.; Wang, C.; Chen, X.; Song, S. icsBERTs: Optimizing Pre-trained Language Models in Intelligent Customer Service. In Proceedings of the International Neural Network Society Workshop on Deep Learning Innovations and Applications (INNS DLIA'23), Gold Coast, Australia, 23 June 2023; Volume 222, pp. 127–136. [CrossRef]
28. Kaddour, J.; Harris, J.; Mozes, M.; Bradley, H.; Raileanu, R.; McHardy, R. Challenges and Applications of Large Language Models. *arXiv* **2023**, arXiv:2307.10169.
29. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models are Few-Shot Learners. In *Proceedings of the Advances in Neural Information Processing Systems, Virtual, 6–12 December 2020*; Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H., Eds.; Curran Associates, Inc.: New York, NY, USA, 2020; Volume 33, pp. 1877–1901.
30. Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv* **2023**, arXiv:2307.09288.
31. Rae, J.W.; Borgeaud, S.; Cai, T.; Millican, K.; Hoffmann, J.; Song, F.; Aslanides, J.; Henderson, S.; Ring, R.; Young, S.; et al. Scaling Language Models: Methods, Analysis & Insights from Training Gopher. *arXiv* **2022**, arXiv:2112.11446.
32. OpenAI. GPT-4 Technical Report. *arXiv* **2023**, arXiv:2303.08774.
33. Otter, D.W.; Medina, J.R.; Kalita, J.K. A Survey of the Usages of Deep Learning for Natural Language Processing. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *32*, 604–624. [CrossRef] [PubMed]
34. Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; Galstyan, A. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv. (CSUR)* **2021**, *54*, 115. [CrossRef]
35. Wu, S.; Roberts, K.; Datta, S.; Du, J.; Ji, Z.; Si, Y.; Soni, S.; Wang, Q.; Wei, Q.; Xiang, Y.; et al. Deep learning in clinical natural language processing: a methodical review. *J. Am. Med. Inform. Assoc.* **2020**, *27*, 457–470. [CrossRef] [PubMed]

36. Xue, L.; Constant, N.; Roberts, A.; Kale, M.; Al-Rfou, R.; Siddhant, A.; Barua, A.; Raffel, C. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Virtual, 6–11 June 2021; pp. 483–498. [CrossRef]
37. Sarti, G.; Nissim, M. IT5: Large-scale Text-to-text Pretraining for Italian Language Understanding and Generation. *arXiv* **2022**, arXiv:2203.03759.
38. Scao, T.L.; Fan, A.; Akiki, C.; Pavlick, E.; Ilić, S.; Hesslow, D.; Castagné, R.; Luccioni, A.S.; Yvon, A.; Gallé, M.; et al. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. *arXiv* **2023**, arXiv:cs.CL/2211.05100.
39. Laurençon, H.; Saulnier, L.; Wang, T.; Akiki, C.; del Moral, A.V.; Le Scao, T.; Von Werra, L.; Mou, C.; González Ponferrada, E.; Nguyen, H.; et al. The BigScience ROOTS Corpus: A 1.6TB Composite Multilingual Dataset. *arXiv* **2023**, arXiv:2303.03915.
40. Kreutzer, J.; Caswell, I.; Wang, L.; Wahab, A.; van Esch, D.; Ulzii-Orshikh, N.; Tapo, A.; Subramani, N.; Sokolov, A.; Sikasote, C.; et al. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *Trans. Assoc. Comput. Linguist.* **2022**, *10*, 50–72. [CrossRef]
41. El-Kishky, A.; Chaudhary, V.; Guzmán, F.; Koehn, P. CCAligned: A Massive Collection of Cross-Lingual Web-Document Pairs. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, Online, 16–20 November 2020; pp. 5960–5969. [CrossRef]
42. Bender, E.M.; Gebru, T.; McMillan-Major, A.; Shmitchell, S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In Proceedings of the ACM Conference on Fairness, Accountability, and Transparency. Association for Computing Machinery, FAccT '21, Virtual, 3–10 March 2021; pp. 610–623. [CrossRef]
43. Abadji, J.; Ortiz Suarez, P.; Romary, L.; Sagot, B. Towards a Cleaner Document-Oriented Multilingual Crawled Corpus. In Proceedings of the 13th Language Resources and Evaluation Conference, Marseille, France, 20–25 June 2022; European Language Resources Association: Paris, France, 2022; pp. 4344–4355.
44. Wenzek, G.; Lachaux, M.A.; Conneau, A.; Chaudhary, V.; Guzmán, F.; Joulin, A.; Grave, E. CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data. In Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France, 11–16 May 2020; pp. 4003–4012.
45. Bañón, M.; Chen, P.; Haddow, B.; Heafield, K.; Hoang, H.; Esplà-Gomis, M.; Forcada, M.L.; Kamran, A.; Kirefu, F.; Koehn, P.; et al. ParaCrawl: Web-Scale Acquisition of Parallel Corpora. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Virtual, 5–10 July 2020; pp. 4555–4567. [CrossRef]
46. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* **2020**, *21*, 1–67.
47. Bojanowski, P.; Grave, E.; Joulin, A.; Mikolov, T. Enriching Word Vectors with Subword Information. *Trans. Assoc. Comput. Linguist.* **2017**, *5*, 135–146. [CrossRef]
48. Lee, K.; Ippolito, D.; Nystrom, A.; Zhang, C.; Eck, D.; Callison-Burch, C.; Carlini, N. Deduplicating Training Data Makes Language Models Better. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Dublin, Ireland, 22–27 May 2022.
49. Gao, L.; Biderman, S.; Black, S.; Golding, L.; Hoppe, T.; Foster, C.; Phang, J.; He, H.; Thite, A.; Nabeshima, N.; et al. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. *arXiv* **2020**, arXiv:2101.00027.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.