Ungoliant: An Optimized Pipeline for the Generation of a Very Large-Scale Multilingual Web Corpus

Julien Abadji¹ Pedro Javier Ortiz Suárez¹,² Laurent Romary¹ Benoît Sagot¹

¹Inria, Paris, France ²Sorbonne Université, Paris, France
{julien.abadji, pedro.ortiz, benoit.sagot, laurent.romary}@inria.fr

Abstract

Since the introduction of large language models in Natural Language Processing, large raw corpora have played a crucial role in Computational Linguistics. However, most of these large raw corpora are either available only for English or not available to the general public due to copyright issues. Nevertheless, there are some examples of freely available multilingual corpora for training Deep Learning NLP models, such as the OSCAR and Paracrawl corpora. However, they have quality issues, especially for low-resource languages. Moreover, recreating or updating these corpora is very complex. In this work, we try to reproduce and improve the goclassy pipeline used to create the OSCAR corpus. We propose a new pipeline that is faster, modular, parametrizable, and well-documented. We use it to create a corpus similar to OSCAR but larger and based on recent data. Also, unlike OSCAR, the metadata information is at the document level. We release our pipeline under an open source license and publish the corpus under a research-only license.

1 Introduction

With the increasing interest in language modeling in recent years in Natural Language Processing (NLP) (Rogers et al., 2020), particularly concerning contextualized word representations¹ (Peters et al., 2018; Devlin et al., 2019), there has also been an explosion in interest for large raw corpora, as some of these latest models require almost 1 TiB of raw text for pre-training (Raffel et al., 2020; Brown et al., 2020).

While most of these language models were initially trained in English (Devlin et al., 2019; Yang et al., 2019; Clark et al., 2020; Zaheer et al., 2020; Xiong et al., 2021) and consequently most of the large corpora used to pre-train them were in English, there has been a recent push to produce larger high-quality corpora for other languages, namely those of Grave et al. (2018), CCNet (Wenzek et al., 2020), Multilingual C4 (mC4) (Xue et al., 2020) and OSCAR (Ortiz Suárez et al., 2019, 2020) for pre-training language models, as well as, Paracrawl (Esplà et al., 2019; Bañón et al., 2020), CCaligned (El-Kishky et al., 2020) and WikiMatrix (Schwenk et al., 2021) which are parallel corpora for training Machine Translation (MT) models. Of these, only OSCAR, Paracrawl, CCaligned and WikiMatrix are freely available and easily downloadable.

In this paper we propose a new multilingual corpus for language modeling, and for that we take inspiration in the OSCAR corpus and its pipeline goclassy² (Ortiz Suárez et al., 2019, 2020), but we propose a new pipeline Ungoliant³ that is faster, modular, parametrizable and well-documented. We then use it to produce a new corpus similar to OSCAR, yet larger, based on recent data containing mentions of last years' events such as the COVID-19 pandemic, the 2020–2021 United States racial unrest, the Australian wildfires, the Beirut explosion and Brexit among others. Moreover, contrarily to OSCAR, our corpus retains metadata information at the document level. We release our pipeline under an Apache 2.0 open source license and we publish the corpus under a research-only use license following the licensing schemes proposed by OSCAR (Ortiz Suárez et al., 2019, 2020) and Paracrawl (Esplà et al., 2019; Bañón et al., 2020).

¹https://github.com/oscar-corpus/goclassy
²https://github.com/oscar-corpus/ungoliant

© 2021 The Author(s)

This work is licensed under a Creative Commons Attribution 4.0 International License.

DOI: https://doi.org/10.14618/ids-pub-10468
Limitations of the OSCAR Corpus and its Generation Pipeline

2.1 OSCAR

OSCAR is a multilingual corpus derived from CommonCrawl\(^4\), a project that provides web crawl data for everyone on a periodic manner, usually each month. CommonCrawl provides data in several formats, from raw HTML source code to pure text. OSCAR was generated from the pure text data version (WET files) of the November 2018 crawl, distributed in the form of 56,000 shards, that were then filtered and classified by language (Ortiz Suárez et al., 2019, 2020). OSCAR is available through several means, and has been used in numerous projects (Ortiz Suárez et al., 2019). OSCAR's generation pipeline also suffers from numerous issues, which we plan to address simultaneously with the release of a new, more powerful, stable, and higher quality pipeline.

Simply put, OSCAR is composed of single language files that contain textual data (\texttt{t.a.txt} for the Tamil language, for example). However, due to the often huge sizes of these files, and subsequently the impracticality of storage and distribution, OSCAR files are split and compressed in equally sized parts.

OSCAR comes in four different versions, each suited differently for different tasks, and allows less limited ways of sharing the corpus more widely. These versions are either unshuffled or shuffled (that is, for each language, lines have been shuffled, destroying records integrity), and non-deduplicated or deduplicated (since duplicate lines account for more than half of the total data\(^5\) generated by the pipeline). For the unshuffled versions, each language file contains paragraphs that come from the same record, and each paragraph is separated by a newline.

OSCAR is inherently linked to its generation pipeline, and as such its quality partly depends on the pipeline's quality. While OSCAR is considered to be one of the cleanest multilingual corpora available (Caswell et al., 2020, 2021), several problems have been described, and the state of the publicly available code raises questions about maintenance and maintenability of the pipeline itself.

Apart from the fact that its content dates back to 2018, the current OSCAR corpus suffers from quality issues discussed in (Caswell et al., 2020, 2021), including:

- **Language label mismatches and inconsistencies**, which occurs earlier in the pipeline and would be fixable downstream,
- **Representation washing** as defined by Caswell et al. (2021), whereby low resource languages, while present in the corpus, are of a significantly lower quality than higher resource languages without any quality metric available publicly.

The most recent Common Crawl dump contains 64,000 shards. Each shard is composed of numerous records, and each record holds textual content along with metadata. While CommonCrawl shards hold document-level metadata that could be useful downstream, they were discarded and do not appear in OSCAR, whereas other corpora generated from the same source include them, e.g. CCNet (Wenzek et al., 2020). This limits OSCAR users to the textual content only, whereas metadata could have been distributed along with the corpus itself.

2.2 goclassy

OSCAR was built using goclassy, a high-performance asynchronous pipeline written in Go (Ortiz Suárez et al., 2019). However, it suffers from several caveats that makes the re-generation and update of the corpus relatively complex in practice.

While goclassy’s source code is easily readable thanks to the choice of an uncluttered language and a pragmatic approach, the lack of structure in both the source and the project itself makes goclassy difficult to extend and maintain.

The pipeline is not functional out-of-the-box, as the user has to provide the compressed shards from CommonCrawl, manually install fasttext (Joulin et al., 2016, 2017) and create specific directories by themselves, since only partial instructions are given in the supplied README file.

goclassy also makes heavy use of I/O, as data is saved and loaded repeatedly between steps; as an example, the identification step stores language identification data and individual sentences in two files, before generating the final files (one per language). Despite these limitations, goclassy’s performance is good due to Go’s emphasis on easy and efficient parallelization and inherent speed. The pipeline uses clever handling of file descriptors, limiting I/O calls cost in some parts.

\(^4\)\url{https://commoncrawl.org}

\(^5\)OSCAR-orig: 6.3TB, OSCAR-dedup: 3.2TB
3 Building a new OSCAR-like corpus

We introduce Ungoliant, a new corpus generation pipeline that, like goclassy, creates a large-scale multilingual text corpus from a CommonCrawl dump. Contrarily to goclassy, Ungoliant is fully modular, better structured, and highly parametrizable; thereby allowing comparisons between several parallelization strategies. A specific effort was put in testing and documentation. Parts of Ungoliant are heavily inspired by goclassy, although it is implemented in Rust rather than in Go, which is sometimes faster.\(^6\)

Additionally, we use Ungoliant to generate a new corpus from a recent Common Crawl dump. The new corpus includes metadata information while retaining backward compatibility with the OSCAR corpus.

3.1 Ungoliant

3.1.1 Rationale and scope

While Ungoliant is heavily inspired by goclassy, it provides a better set of tools to download, process, filter and aggregate textual and contextual data from CommonCrawl. These operations can be sequential, parallel or both, depending on contexts and performance requirements.

We provide both batch and streaming processing, so that the whole pipeline could be run either online, with every step running on streams of data, or offline, with every step running on tangible files, or a mix of both, using already downloaded CommonCrawl dumps but streaming the rest of the process. Moreover, we embed numerous filtering and deduplication utilities directly inside Ungoliant, making these features available for pipeline composition and post-processing.

Ungoliant features a loosely defined pipeline interface, on which we re-implement goclassy’s one, while improving performance by threading more aggressively and avoiding I/O where it is not necessary: While goclassy uses intermediate files for tags and sentences, we try to keep everything in memory in order to avoid losing time loading or writing files. The Rust language provides constructs that helps us build complex abstractions and pipelines while limiting proactive file I/O or computing, since nearly all the reimplemented pipeline is built around lazy evaluation. File I/O is only used when loading shards, and when writing sentences in language files.

Through benchmarking we found that the best parallelization strategy is to use rayon\(^7\), a work-stealing (Blumofe and Leiserson, 1999) parallel and concurrent library enabling massive parallelization. We parallelize on shard-, record- and sentence-level processing.

To evaluate Ungoliant performance, we run both goclassy and Ungoliant’s implementation on 1, 10, 25 and 100 Common Crawl shards both on a middle-range laptop computer (i5-7200u, 8GB RAM, NVMe SSD) and a HPC node (Xeon 5218 (64 Threads), 180GB RAM). Results are shown in Table 1.

<table>
<thead>
<tr>
<th>Platform</th>
<th>#shards</th>
<th>goclassy</th>
<th>Ungoliant</th>
<th>Approx. speedup</th>
</tr>
</thead>
<tbody>
<tr>
<td>Desktop</td>
<td>1</td>
<td>30s</td>
<td>13s</td>
<td>×2.3</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>3m6s</td>
<td>2m12s</td>
<td>×1.3</td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>9m10s</td>
<td>5m47s</td>
<td>×1.5</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>40s</td>
<td>6s</td>
<td>×6.6</td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>2m40s</td>
<td>1m6s</td>
<td>×2.4</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>7m59s</td>
<td>4m14s</td>
<td>×1.8</td>
</tr>
</tbody>
</table>

Table 1: Comparison of approximate generation times depending on platform and number of shards.

In other words, switching from the original OSCAR corpus and the newly generated one should be a drop-in operation.

\(^6\)https://benchmarksgame-team.pages.debian.net/benchmarksgame/fastest/rust-go.html

\(^7\)https://github.com/rayon-rs/rayon
3.2.1 Metadata extraction and linking

Our choice of keeping the corpus backward compatible with the original OSCAR introduces changes in the way the corpus is generated, namely regarding metadata: a record’s body is composed of sentences that aren’t guaranteed to be of the same language. Since OSCAR merges sentences from multiple records into a single file, special attention has to be paid to the metadata dispatch too.

Approaches to tackle this problem range from (1) storing all metadata in a single location to (2) having language-specific metadata files that contain the metadata for each line in the language file.

Both (1) and (2) have their strengths and weaknesses, namely:

1. Having all metadata at the same place may facilitate wide queries about whole metadata, but at a cost of a very large size (which harms both accessibility and performance).

2. Getting the metadata for a given line is fast since line numbers are synchronized, but there is repeated information and a potentially important increase in size.

We choose a hybrid approach which keeps metadata local to each language, while trying to limit the information repetition by keeping an entry by group of chunks rather than by line, where a chunk is a series of contiguous sentences that share the same language from the same document.

An overview of the pipeline can be seen in Figure 1, with a more precise view on record processing and metadata extraction in Figure 2.

Metadata are distributed via JSON-encoded files holding an ordered list of metadata entries, along with offsets \( (o) \) and paragraph lengths \( (l) \), enabling any user to get the content of a said metadata by querying for lines \( (o, o + l] \) in the content file.

This approach still has drawbacks, in particular when looking for the corresponding metadata of a given sentence/paragraph, where one has to perform a search on the metadata file, or when working with multilingual documents. Another drawback is the resulting cost of potentially merging back numerous language parts: Since metadata query is offset-based, merging back metadata files implies updating those offsets.

Having paragraphs and metadata linked by offsets in a highly parallelized pipeline implies to take special care at the offset level. The solution is to use shard-scoped offsets (starting from 0 for each language), and to keep global offsets protected by a mutex guard. This way, when a given shard is done processing and is ready to be written on disk, we convert shard-scoped offsets to global-scoped ones, update the global-scoped ones and then write text and metadata on disk.

We compare running times for the reimplementation of the goclassy pipeline, and our new pipeline adding metadata extraction, using both desktop and HPC contexts. The results are reported in Table 2.

<table>
<thead>
<tr>
<th>Platform</th>
<th>#shards</th>
<th>OSCAR With Metadata</th>
<th>Speedup</th>
</tr>
</thead>
<tbody>
<tr>
<td>Desktop</td>
<td>1</td>
<td>13s 12s</td>
<td>×1.1</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>2m12s 1m55s</td>
<td>×1.1</td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>5m47s 4m50s</td>
<td>×1.2</td>
</tr>
<tr>
<td>HPC</td>
<td>1</td>
<td>6s 7s</td>
<td>×0.9</td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>1m6s 1m12s</td>
<td>×0.9</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>4m14s 4m36s</td>
<td>×0.9</td>
</tr>
</tbody>
</table>

Table 2: Comparison of approximate generation times with and without metadata generation.

<table>
<thead>
<tr>
<th>Version</th>
<th>Source Textual (dedup)</th>
<th>Metadata</th>
<th>Total (increase)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2018</td>
<td>7.42TB</td>
<td>N/A</td>
<td>6.3TB</td>
</tr>
<tr>
<td>2021</td>
<td>8.06TB</td>
<td>7.2TB (3.3TB)</td>
<td>1.2TB 8.4TB (+33%)</td>
</tr>
</tbody>
</table>

Table 3: Comparison of CommonCrawl and OSCAR sizes between 2018 and 2021 versions. Compressed (CommonCrawl) sources are from November 2018 and February 2021. Total is Textual + Metadata without deduplication.

3.3 Characteristics of our new backward compatible OSCAR-like corpus

We evaluate the newly generated corpus, assessing its ability to reflect events that occurred after the publication of OSCAR 2018 and detail the metadata format and potential use.

3.3.1 Comparison with OSCAR

While it is expected that our new corpus has a larger file size than OSCAR since CommonCrawl itself grew from 7.42TB to 8.06TB, metadata quickly adds up and take for nearly 15% of the whole uncompressed data.

The size augmentation is not the same for each language, and while the whole corpus is bigger...
Figure 1: Scheme of the Ungoliant pipeline. The red square represents CommonCrawl content hosting, where the compressed shards are fetched. The Process Shard steps hold shard processing, paragraph creation and merging (see Figure 2), and are internally parallelized.

- **CommonCrawl** compressed shard.
- **file**: Uncompressed shard, containing records.
- **/code**: Record Metadata
- **/tag**: Language identification
- **¶**: Paragraph, composed of sentences identified as **/tag**

Figure 2: Record processing with metadata extraction. Headers are kept aside while sentences are identified and grouped into same-language bins. Headers are then cloned for each bin, and are sequentially stamped with an offset that is recorded for the whole operation, and written to disk into text and metadata files by language.
now, some languages are smaller than they were before.

Results show that already largely represented languages gain more and more data (like the English language, which constitutes more than a third of the original OSCAR), except for the Russian language which loses approximately 100Gb of textual content. These results are summarized in Figure 3.

However, in a context where the number of languages is very high (higher than 150) and of varying sizes, evolution can’t be analyzed via a mere size evaluation. By computing, for each language, the relative size difference between the 2018 and 2021 releases of OSCAR, less resourced languages do appear, hinting at a better representation of some of them. These results can be found in Figure 4.

Numerous languages have been omitted from Figure 4, either:

- because they were present in the original OSCAR and are now absent (Central Bikol and Cantonese)
- because they were absent in the original OSCAR and are now present (Manx, Rusyn, Scots and West Flemish)

Precautions have to be taken when using these corpora and further work has to be done to correctly assess the quality of low-to-mid resource languages in order to better reflect the quality of each corpus to the OSCAR users. Some languages exhibited either a particularly low number of sentences or a very low quality, and as such couldn’t be usable, while still accounting for a language in the total language count of the original OSCAR.

3.3.2 Metadata

Metadata provides new contextual data that is useful to evaluate the corpus and draw metrics.

The total size of metadata is 1.2TB, ranging from 4Kb to 500Gb, depending on the number of lines. Relative size varies from 100% to 20%, diminishing with the textual data size, which is expected.

Metadata are provided in single files for now, but split versions of both textual and contextual data will be released soon after the release of the corpus, enabling easy access.

Our choice of keeping metadata aside from the main content adds some complexity when working with both textual and contextual data:

- When trying to get the metadata of a given sentence, one has to get the line number $k$, then sequentially (or use a search algorithm since offsets are sorted) look for the record (with offset $o$ and length $l$), where $k \in [o, o+l]$.
- Looking for lines corresponding to a particular metadata entry is easier: one has to read the textual file, skipping until the $o$-th line, then read $l$ lines.

3.3.3 Presence of events

Using a sample of an English part of our corpus, we perform a simple search of terms in order to assess and compare the presence of pre- and post-2018 events and persons in both corpora. Terms and frequency are grouped in Table 4.

Our corpus keeps around the same number of occurrences for pre-2018 events or public figures such as Barack Obama, while increasing the occurrence of people linked to more recent events (Joe Biden).
We include search terms linked to post-2018 events in French and Arabic which are smaller corpora (resp. 200 and 80 GB), and in Burmese, a mid-resource language (approximately 2GB). We observe a term occurrences evolution that reflects the linked events’ timing and importance.

3.4 License

This new corpus will be released under a research-only license that is compliant with the EU’s exceptions for research in text and data mining. Contrarily to the original OSCAR, no shuffled version of the corpus will be distributed, instead we will put in place an authentication system that will allow us to verify that requests for the corpus come from research institutions. A contact form will be also provided for independent researchers so that we can study their particular cases and determine if the utilization of the corpus corresponds to a legitimate research use.

Moreover, the introduction of metadata makes our corpus far more queryable, thus simplifying and speeding up the handling of take-down GDPR requests. For this reason, we will be releasing the complete set of metadata under a CC0 public domain license, so that any individual can check if their personal or even copyrighted data is in our new corpus and make a request accordingly.

4 Conclusion

We show that our solution is able to generate an OSCAR-like corpus that is augmented with metadata without breaking compatibility, while being faster, better tested and thoroughly documented. We believe our new pipeline and corpus will be useful for applications in computational linguistics as well as in corpus linguistics in general.

The generated corpus is of a larger size when including metadata and without deduplication. However, deduplicated textual content is of the same magnitude between OSCAR 2018 and OSCAR 2021, while reflecting topic changes from all over the world. This fact suggests that old data may be lost with the time passing, and could be resolved by using CommonCrawl releases to build an incremental corpus, with every version augmenting the corpus size.

Metadata enables queries and statistics on the generated data, and we believe that it can be used to filter OSCAR to generate corpora that respond to certain criteria.

We plan to make this new version of OSCAR available under research constraints, with split versions of both textual content and metadata along with tools to operate on the corpus, enabling fast and easy operation on the corpus for researchers.

References


Spain (Online). International Committee on Computational Linguistics.


