Documenting the English Colossal Clean Crawled Corpus

Jesse Dodge[♣] **Maarten Sap**[♡] **Ana Marasović**[♣] **William Agnew**[♡]
Gabriel Ilharco[♡] **Dirk Groeneveld**[♣] **Matt Gardner**[♣]
[♡]Paul G. Allen School of Computer Science & Engineering, University of Washington
[♣]Allen Institute for Artificial Intelligence
jessed@allenai.org

Abstract

As language models are trained on ever more text, researchers are turning to some of the largest corpora available. Unlike most other types of datasets in NLP, large unlabeled text corpora are often presented with minimal documentation, and best practices for documenting them have not been established. In this work we provide the first documentation for the Colossal Clean Crawled Corpus (C4; Raffel et al., 2020), a dataset created by applying a set of filters to a single snapshot of Common Crawl. We begin with a high-level summary of the data, including distributions of where the text came from and when it was written. We then give more detailed analysis on salient parts of this data, including the most frequent sources of text (e.g. `patents.google.com`, which contains a significant percentage of machine translated and/or OCR'd text), the effect that the filters had on the data (they disproportionately remove text in AAE), and evidence that some other benchmark NLP dataset examples are contained in the text. We release a web interface to an interactive, indexed copy of this dataset, encouraging the community to continuously explore and report additional findings.

1 The English Colossal Clean Crawled Corpus (C4)

Unlike most other types of datasets in NLP, large unlabeled text corpora are often presented with minimal documentation. We provide some of the first documentation of the English Colossal Clean Crawled Corpus (C4; Raffel et al., 2020), one of the largest corpora of text available. C4 is created by taking a snapshot of Common Crawl[1] and applying a number of filters to remove text with the intention of retaining high-quality natural English. We host three different versions of the data: C4.EN.NOCLEAN (C4 with only a language ID filter applied), C4.EN.NOBLOCKLIST (result of all filters except one that discards documents from a list of banned words), and C4.EN (the result of all filters).[2] We also release some documentation throughout the rest of this document, motivated by recent calls for better documentation of datasets (Gebru et al., 2018; Bender et al., 2021; Hutchinson et al., 2021). However, we recognize that the reported items herein are only a fraction of the important and relevant items to report for a massive web-crawled dataset such as C4. To facilitate further discussion of the data we host an indexed version of C4 at `https://c4-search.apps.allenai.org/`, allowing anyone to search it. We also provide as a place to report and discuss examples of interest.[3]

Dataset	# documents	# tokens	size
C4.EN.NOCLEAN	1.1 billion	1.4 trillion	2.3 TB
C4.EN.NOBLOCKLIST	395 million	198 billion	380 GB
C4.EN	365 million	156 billion	305 GB

Table 1: Statistics for the three corpora we host. One "document" is the text scraped from a single URL. Tokens are counted using the SpaCy English tokenizer. Size is compressed JSON files.

1.1 What is C4.EN?

C4 (Raffel et al., 2020) is a massive corpus of unlabeled web-crawled text, created by taking the April 2019 snapshot of Common Crawl and applying a number of filters to remove text with the intention of retaining high-quality natural English. This includes filtering out lines which don't end in a terminal punctuation mark or have fewer than three words, discarding documents with less than five sentences or that contain "lorem ipsum" placeholder text, and removing documents with any word on the "List of Dirty, Naughty,

[1] `https://commoncrawl.org/`

[2] Instructions to download these datasets can be found at `https://github.com/allenai/allennlp/discussions/5056`.

[3] `https://github.com/allenai/c4-documentation`

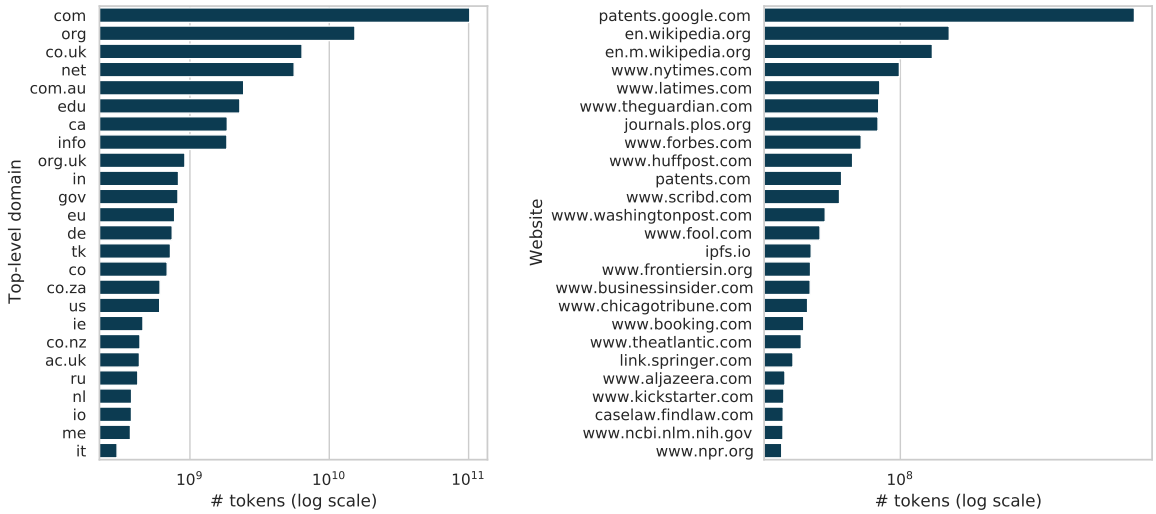

Figure 1: Number of tokens from the 25 most represented top-level domains (left) and websites (right) in C4.EN.

Obscene, or Otherwise Bad Words".[4] Additionally, `langdetect`[5] is used to remove documents which weren't classified as English with probability at least 0.99, so C4 is primarily comprised of English text. For brevity, we refer readers to Raffel et al. (2020) for a full list of the filters.

1.2 C4.EN.NOCLEAN and C4.EN.NOBLOCKLIST Variants

In addition to C4.EN, the "cleaned" version of C4 (created by applying all filters), we host the "uncleaned" version (C4.EN.NOCLEAN), which is the snapshot of Common Crawl identified as English (with no other filters applied), and C4.EN.NOBLOCKLIST, which is the same as C4.EN but without filtering out documents containing tokens from a blocklist of words (see §4 for more details). Table 1 contains some statistics for the three corpora. C4.EN and C4.EN.NOBLOCKLIST each have more than 350 million documents, and more than a 150 billion tokens. C4.EN.NOCLEAN is more than 2.3 TB of text.

2 Corpus-level statistics

We provide some high-level analyses of the provenance of documents in C4.EN, the cleaned corpus, and broadly characterize the prevalence of different internet sources.

[4] https://git.io/vSyEu
[5] https://pypi.org/project/langdetect/

2.1 Internet domains

We first analyze which internet domains and websites are most represented in C4.EN in terms of number of word tokens (measured using the SpaCy English tokenizer).[6] We use the TLDExtract[7] package to parse the URLs.

2.1.1 Internet top-level domains

Figure 1 (left) shows the 25 most represented top-level domains (TLD)[8] in C4.EN. Unsurprisingly, popular top-level domains such as `.com`, `.org`, and `.net` are well represented. We note that some top-level domains reserved for non-US, English-speaking countries are less represented, and even some domains for countries with a primary language other than English are represented in the top 25 (such as `ru`).

A significant portion of the text comes from `.gov` websites, reserved for the US government. Another potentially interesting top-level domain is `.mil`, reserved for the US government military. While not in the top 25 TLDs, C4.EN contains 33,874,654 tokens from `.mil` top-level domain sites, coming from 58,394 unique URLs. There are an additional 1,224,576 tokens (from 2,873 unique URLs) from `.mod.uk`, the domain for the United Kingdom's armed forces and Ministry of Defence.

2.1.2 Websites

In Figure 1 (right), we show the top 25 most represented websites in C4.EN, ranked by total number

[6] https://spacy.io/api/tokenizer
[7] https://pypi.org/project/tldextract/
[8] https://en.wikipedia.org/wiki/List_of_Internet_top-level_domains

of tokens. Surprisingly, the cleaned corpus contains substantial amounts of patent text documents, with the single-most represented website in the corpus is `patents.google.com` and `patents.com` being in the top 10. We discuss the implications of this in §3.1.

The second most frequent origin of tokens is English Wikipedia, which has been extensively used in the training of large language models (Devlin et al., 2019; Liu et al., 2019; Brown et al., 2020, e.g., BERT, RoBERTa, GPT-3).

Another category of documents that is substantially represented in news (NYTimes, LATimes, the Guardian, etc), which is also a domain that has been heavily used in pretraining of language models (e.g., GPT-2, Grover; Radford et al., 2019b; Zellers et al., 2019a).

Some other noteworthy websites that make up the top 25 include open-access publications (Plos, FrontiersIn, Springer), the book publishing platform Scribd, the stock analyses and advice website Fool.com, and the distributed file system ipsf.io.

Note that the distribution of websites in C4.EN is not necessarily representative of the most frequently used websites on the internet, as evidenced by the low overlap with the top 25 most visited websites as measured by Alexa.[9]

2.2 Utterance Date

Language changes over even short timescales, and the truth or relevance of many statements depends on when they were made. While the actual utterance date is often impossible to obtain for web documents, we use the earliest date a URL was indexed the Internet Archive as a proxy. We note that using the Internet Archive is not perfect, as it will sometimes index webpages many months after their creation, and only indexed approximately 60% of URLs in C4.EN. Additionally, due to rate-limiting on the Internet Archive's API, we estimate the date ranges using a random sample of 50,000 URLs from C4.EN.

Shown in Figure 2, we find that 92% of the documents in C4.EN are estimated to have been written in the last decade (2011-2019). However, there is a non-trivial amount of data that was written between 10-20 years before the collecting of the data.

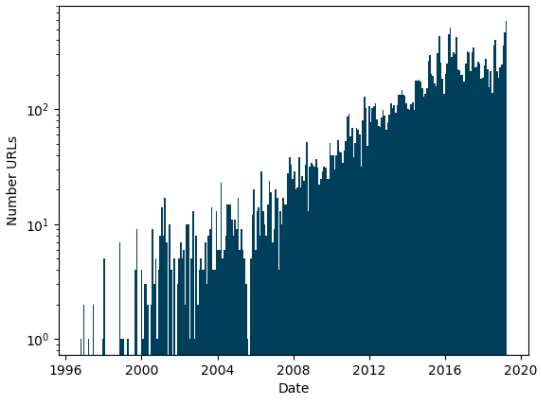

Figure 2: The earliest date a URL was indexed by either the Internet Archive or the Common Crawl snapshot, for only URLs with an index in the Internet Archive.

2.3 Geolocation

We also aim to assess which countries are represented in C4.EN, which we estimate using the location where a webpage is hosted as a proxy for the location of its creators. This information provides a view into who is represented in these datasets, as well as information about inclusion of regional dialects. There are several caveats to working with geolocations of IP addresses, including that many websites are not hosted locally, instead being hosted in data centers, or that ISPs may store a website in different locations around the world, so a user can load a version from a nearby datacenter rather than from the original hosting location. We use an IP-country database[10] and present country-level URL frequencies from 175,000 randomly sampled URLs from the cleaned common crawl dataset.

As shown in Figure 3, our findings show that 51.3% pages are hosted in the United States. The countries with the estimated 2nd, 3rd, 4th largest English speaking populations[11]–India, Pakistan, Nigeria, and The Philippines–have 3.4%, 0.06%, 0.03%, 0.1% the URLs of the United States, despite having many tens of millions of English speakers.

[9] https://www.alexa.com/topsites

[10] https://lite.ip2location.com/database/ip-country

[11] https://en.wikipedia.org/wiki/List_of_countries_by_English-speaking_population

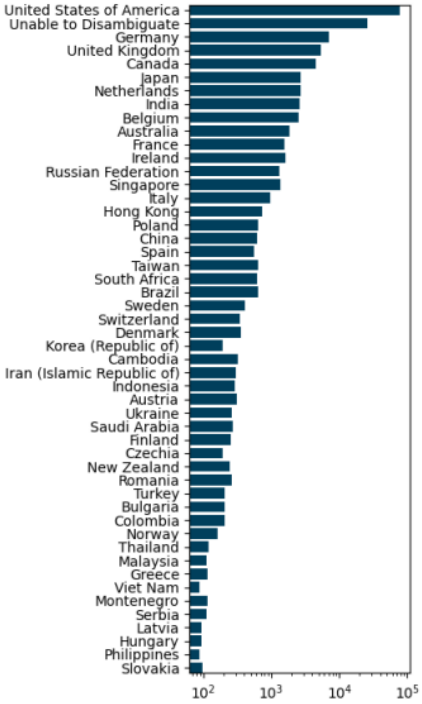

Figure 3: URL frequency by country for 175,000 randomly selected URLS from the cleaned common crawl dataset.

3 What is in the text?

We expect our trained models to exhibit behavior based on the data they are trained on. Most NLP datasets are released with some analysis of the content of the text within, but the scale of C4.EN makes most analyses challenging. Here we focus on machine-generated text and contamination.

3.1 Machine-generated text in `patents.google.com`

As the use of models which can generate natural language text proliferates, web-crawled data will increasingly contain data that was not written by humans. Using this data for language modeling can exacerbate biases, and fitting models on non-natural language can lead to issues in production. In general it can be difficult to distinguish between machine-generated language and natural language, but in some cases we can use meta-data from crawled websites to estimate the proportion of machine-generated text. Here we examine the Internet domain from which we get the most tokens: `patents.google.com`.

Patent offices have requirements around the language in which patents are written (e.g., the Japanese patent office requires patents be in Japanese). `patents.google.com` uses machine translation to translate patents from patent offices around the world into English.[12] Table 3 in Appendix A.2 includes the number of patents in C4.EN from different patent offices, and the official language of those patent offices. While the majority of the patents in this corpus are from the US patent office, more than ten percent are from patent offices which require patents be submitted in a language other than English[13].

While some patents in this corpus are native digital documents, many were physical documents scanned and run through Optical Character Recognition. Indeed, some older documents from other patent offices are first run through OCR then machine translation systems; some example documents are linked in Appendix A.2.

3.2 Benchmark data contamination

In this section, we study *benchmark data contamination* (Brown et al., 2020), i.e., to which extent downstream examples occur in the pretraining corpus. There are generally two ways datasets can end up in a snapshot from Common Crawl: either a given dataset is built from text on the web, such as the IMDB dataset (Maas et al., 2011) and the CNN/Daily Mail summarization dataset (Hermann et al., 2015; Nallapati et al., 2016), or it is uploaded after creation (e.g., to a github repository, for easy access). In this section, we explore both input and input-and-label contaminations of evaluation sets of popular datasets.

Unlike Brown et al. (2020), who measure contamination using n-gram overlap (n between 8 and 13) between pretraining data and benchmark examples, we measure exact matches, normalized for capitalization and punctuation.

Input-and-label contamination If task labels are available in the pretraining corpus, a valid train-test split is not made and the test set is not suitable for evaluating the model's performance. For tasks similar to language modeling (e.g., abstractive summarization) the task labels are target tokens. Thus, if target text occurs in the pretraining corpus, the model can learn to copy the text instead of actually solving the task (Meehan et al., 2020; Carlini et al.,

[12] "Patents with only non-English text have been machine-translated to English and indexed", from https://support.google.com/faqs/answer/7049585

[13] Many patent offices require a patent be filed in a particular language, but also allow translations into other languages be submitted, so this is an upper bound on the number of translated documents.

2020), and we measure the model's data-copying abilities rather than their abilities to solve the end-task.

In Table 2, we show that only about 1% of the target summaries ("highlights") in the CNN/Daily Mail corpus (Nallapati et al., 2016) appear in C4.EN. From the LAMA dataset (Petroni et al., 2019), where evaluation examples are comprised of template-generated sentences with a masked token, we find 4.6% and 3.3% of the examples in the T-REx and Google-RE sets, respectively, exist verbatim in C4.EN. While this is a tiny fraction of the C4.EN dataset, a language model pretrained on C4.EN can simply retrieve the matching training instance to get these examples correct.

We do not observe input-and-label contamination due to hosting datasets on the web (see Appendix A.3).

Input contamination Input contamination of development and test examples that does *not* include labels can also lead to downstream problems. For example, it can hinder testing how well pretrained models perform on new, previously unseen inputs such as new combinations of familiar syntactic structures (Kim and Linzen, 2020). In Table 2, we show that only around 2.4% of the test questions from BoolQ (Clark et al., 2019) have an exact match in C4.EN. We also examine CoLA (Warstadt et al., 2019), a dataset in which grammatical sentences have positive labels while ungrammatical examples have negative labels; we find exact matches in C4.EN for 14.4% of test instances. Test labels are not available for this dataset, but manual inspection reveals that the majority of these instances are grammatical, suggesting a (perhaps unsurprising) bias towards positive-labeled examples.

Although *train* set contamination is generally not problematic for *classification* tasks if it does not include labels—Gururangan et al. (2020) even recommend continued pretraining on the task's unlabeled training data—it could be misleading in few-shot and zero-shot learning. These learning procedures have been highlighted as one of the most exciting and impactful NLP directions in the last year (Ruder, 2021; Brown et al., 2020), but how to properly evaluate models trained in these settings has not been established yet.[14] For instance, do we really evaluate few-shot or zero-shot learning on a given task if the model is adjusted on a notable number of end-task training and/or development inputs during pretraining? The LAMA dataset is one which is often used to evaluate zero-shot performance, and perhaps this practice should be considered carefully.

To recap, of the five datasets we examine, all have some level of contamination, though for most it's small. With open-sourced resources—a pretrained language model T5 (Raffel et al., 2020) and a pretraining corpus C4.EN (this work)—researchers are better positioned to construct more rigorous training and evaluation settings for T5-based models. Although we hope this motivates releasing other large language models and their pretraining corpora, the democritization of language models is undergoing discussion (Solaiman et al., 2019; Zellers et al., 2019b; Shevlane and Dafoe, 2020). Our observations support dynamically collecting data with the human-in-the-loop approach (Nie et al., 2020) that might reduce contamination of future benchmarks since (i) pretaining data is infrequently collected, and (ii) annotator-written examples for a given task are less likely to be (previously) crawled from the web.

4 The Impact of Blocklist Filtering

One of the main components of the C4 pipeline is the filtering out of documents that contain any word from a blocklist of "bad" words,[15] with the intent to remove hateful and toxic language as well as obscene, sexual, or lewd content. However, determining whether a document has toxic or lewd content is a more nuanced endeavor that goes beyond detecting "bad" words; hateful and lewd content can be expressed without negative keywords (e.g., microaggressions, innuendos; Breitfeller et al., 2019; Dinan et al., 2019). Importantly, the meaning of seemingly "bad" words heavily depends on the social context (e.g., impoliteness can serve prosocial functions; Wang et al., 2012), and *who* is saying certain words influences its offensiveness (e.g., the reclaimed slur "*n*gga*" is considered less offensive when uttered by a Black speaker than by a white speaker; Croom, 2013; Galinsky et al., 2013). As such, blocklist filtering risks removing harmless text by minority authors.

Whose English is included? We aim to investigate the extent to which minority voices are be-

[14] For example, Gao et al. (2020b) have recently argued that researchers should use only a few development examples instead of entire development sets for few-shot learning.

[15] https://git.io/vSyEu

Input or Label	Dataset	Percent Matched	Count Matched / Dataset Size
Label	LAMA T-REx	4.6%	1,585 / 34,014
Label	LAMA Google-RE	3.3%	184 / 5,527
Label	CNN/Daily Mail Highlights	1.1%	470 / 44618
Input	BoolQ	2.4%	79 / 3,245
Input	CoLA	14.4%	153 / 1,063

Table 2: The number of exact matches from evaluation sets of various benchmarks in C4.EN. 4.6% of the filled-in LAMA T-REx templates (e.g., "Dante was born in Florence") exist in C4.EN, as do about 3.3% of LAMA Google-RE templates. Only about 1% of CNN/Daily Mail highlights (the summary models are trained to generate) match text in C4.EN. 2.4% of the BoolQ test set questions appear verbatim, and 14.4% of CoLA test sentences appear.

ing removed due to blocklist filtering. Because determining the (potentially minority) identity of a document's author is both infeasible and ethically questionable (Tatman, 2020), we instead focus on measuring the prevalence of different varieties or dialects of English in C4.EN and C4.EN.NOBLOCKLIST. We use a dialect-aware topic model from Blodgett et al. (2016), which was trained on 60M geolocated tweets and relies on US census race/ethnicity data as topics. The model yields posterior probabilities of a given document being in African American English (AAE), Hispanic-aligned English (Hisp), White-aligned English (WAE),[16] and an "other" dialect category (initially intended by the model creators to capture Asian-aligned English). We extract the posterior probabilities of the four dialects for all documents in the C4.EN and C4.EN.NOBLOCKLIST corpora, and assign a dialect to a document based on which has the highest probability.

Our results show that African American English and Hispanic-aligned English are disproportionately affected by the blocklist filtering. Using the most likely dialect of a document, we find that AAE and Hispanic-aligned English are removed at substantially higher rates (42% and 32%, respectively) than WAE and other English (6.2% and 7.2%, respectively). Additionally, we find that 97.8% documents in C4.EN are assigned the WAE dialect category, with only 0.07% AAE and 0.09% Hispanic-aligned English documents. These findings suggest that the blocklist disproportionately removes documents that are detected to be in dialects associated with minority identities (specifically, Black and Hispanic).

[16]We acknowledge that there is disagreement on the choice of terminology to refer to different varieties of English. Here, we use the terms from Blodgett et al. (2016).

5 Related Work

Pretraining corpora BERT (Devlin et al., 2019) was trained on BOOKSCORPUS (Zhu et al., 2015) and English-language WIKIPEDIA. It was soon improved with additional data (ROBERTA; Liu et al., 2019): a portion of CC-NEWS (Nagel, 2016), OPENWEBTEXT (Gokaslan and Cohen, 2019; Radford et al., 2019a), and STORIES (Trinh and Le, 2018). Raffel et al. (2020) filtered Common Crawl (henceforth referred to as CC) to construct C4 and trained T5 with it. CC and C4 are used as a case study in this paper. Since then, other corpora have been (partially) constructed from CC, e.g., PILE (Gao et al., 2020a), CCNET (Wenzek et al., 2020), and MC4 (Xue et al., 2020). One of the largest language models, GPT-3 (Brown et al., 2020), was trained on a mixture of filtered CC (60% of GPT-3's data), WEBTEXT2 (22%; Kaplan et al., 2020), BOOKS1 and BOOKS2 (8% each; Brown et al., 2020), and English-language WIKIPEDIA (3%). Documents from CC and BOOKS2 are sampled less frequently for training GPT-3 because these two corpora are deemed to have lower-quality. GPT-3's CC data was downloaded from 41 shards of monthly CC from 2016–2019, and it constitutes 45TB of compressed text before filtering and 570GB after (∼400 billion byte-pair-encoded tokens). The filtered version is produced by discarding: (i) the CC documents with insufficient similarity to documents from WEBTEXT, WIKIEDIA, and BOOKS corpora according to a logistic regression, and (ii) documents with high overlap with other documents (deduplication).

Audits Since analyzing pretraining corpora is challenging due to their size, their documentation is often missing (Bender et al., 2021; Paullada et al., 2020). To bridge this gap, researchers started to publish systematic posy-hoc studies of these cor-

pora. Gehman et al. (2020) provide an in-depth analysis with respect to toxicity and fake news of OPENWEBTEXT. Caswell et al. (2021) recruited 51 volunteers speaking 70 languages to judge whether five publicly available multilingual web-crawled corpora (El-Kishky et al., 2020; Xue et al., 2020; Ortiz Suárez et al., 2020; Bañón et al., 2020; Schwenk et al., 2019) contain text in languages they report, as well as their quality. We focus on C4 and present a variety of previously unreported observations.

Brown et al. (2020) raise the issue of "benchmark data contamination" (§3.2)—they aimed to reduce contamination before training GPT-3, but due to a bug in their filtering, they estimated that a quarter of 20+ benchmarks have over 50% examples contaminated.[17] Due to the cost, they did not repeat the training, but they provide a post-hoc analysis. Namely, they report that GPT-3's performance on the "clean" benchmarks (contaminated examples discarded) is mostly comparable to the performance on the contaminated benchmarks. While their analysis is useful, we believe our exact match evaluation is more appropriate.

While mentioned studies give post-hoc audits, other works propose frameworks that could be adopted for better collection of pretraining data in the future. Jo and Gebru (2020) discuss parallels between creating historical archives and the curation of machine learning datasets including pretraining corpora. Hutchinson et al. (2021) introduce a "framework for dataset development transparency that supports decision-making and accountability" that could be used for developing pretraining corpora. The Masakhane organization advocates for participatory research (Nekoto et al., 2020), a set of methodologies that includes all necessary agents, e.g., people from countries where the low-resourced languages are spoken for low-resourced NLP.

6 Conclusion

We present some of the first documentation for the provenience of C4.EN, originally introduced by Raffel et al. (2020), and some of the first analysis of its contents. We host three versions of the data for download (https://github.com/allenai/allennlp/discussions/5056), in addition to an indexed version for easy searching (https://c4-search.apps.allenai.org/), and a repository for public discussion of findings (https://github.com/allenai/c4-documentation).

[17] They judged an example as contaminated if it has a 13-gram overlap with anything in the pretraining corpus, or if the entire example overlaps if it is shorter than 13 tokens.

References

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic Dialectal Variation in Social Media: A Case Study of African-American English. In *Proceedings of EMNLP*.

Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. 2019. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. 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 1664–1674.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Nicholas Carlini, Florian Tramèr, Eric Wallace, M. Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ú. Erlingsson, Alina Oprea, and Colin Raffel. 2020. Extracting training data from large language models. arXiv:2012.07805.

Isaac Caswell, Julia Kreutzer, Lisa Wang, Ahsan Wahab, D. V. Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, S. Sarin, Sokhar Samb, B. Sagot, C. Rivera, Annette Rios Gonzales, Isabel Papadimitriou, S. Osei, Pedro Javier Ortiz Suárez, Iroro Orife, Kelechi Ogueji, Rubungo Andre Niyongabo, Toan Q. Nguyen, Mathias Muller, A. Muller, S. Muhammad, N. Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, M. Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, N. D. Silva, Sakine cCabuk Balli, Stella Rose Biderman, Alessia Battisti, A. Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2021. Quality at a glance: An audit of web-crawled multilingual datasets. In *Proceedings of the AfricanNLP Workshop*.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

Adam M Croom. 2013. How to do things with slurs: Studies in the way of derogatory words. *Language & Communication*, 33(3):177–204.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Emily Dinan, Samuel Humeau, Bharath Chintagunta, and Jason Weston. 2019. Build it break it fix it for dialogue safety: Robustness from adversarial human attack. *arXiv preprint arXiv:1908.06083*.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5960–5969, Online. Association for Computational Linguistics.

Adam D Galinsky, Cynthia S Wang, Jennifer A Whitson, Eric M Anicich, Kurt Hugenberg, and Galen V Bodenhausen. 2013. The reappropriation of stigmatizing labels: the reciprocal relationship between power and self-labeling. *Psychol. Sci.*, 24(10):2020–2029.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020a. The pile: An 800gb dataset of diverse text for language modeling.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2020b. Making pre-trained language models better few-shot learners. arXiv:2012.1572.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé, and Kate Crawford. 2018. Datasheets for datasets. In *Proceedings of the 5th Workshop on Fairness, Accountability, and Transparency in Machine Learning*.

Sam Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Findings of EMNLP*.

Aaron Gokaslan and Vanya Cohen. 2019. OpenWebText Corpus.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2021. Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 560–575.

Eun Seo Jo and Timnit Gebru. 2020. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 306–316.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom Brown, Benjamin Chess, Reqon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. arXiv:2001.0836.

Najoung Kim and Tal Linzen. 2020. COGS: A compositional generalization challenge based on semantic interpretation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9087–9105, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Casey Meehan, Kamalika Chaudhuri, and Sanjoy Dasgupta. 2020. A non-parametric test to detect data-copying in generative models. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Sebastian Nagel. 2016. CC-NEWS.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online. Association for Computational Linguistics.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online. Association for Computational Linguistics.

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily L. Denton, and A. Hanna. 2020. Data and its (dis)contents: A survey of dataset development and use in machine learning research. In *The ML-Retrospectives, Surveys & Meta-Analyses NeurIPS 2020 Workshop*.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2019. Alanguage models as knowledge bases? In *EMNLP*.

Alec Radford, Jeffrey Wu, R. Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019a. Language models are unsupervised multitask learners.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019b. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*.

Sebastian Ruder. 2021. ML and NLP Research Highlights of 2020. `http://ruder.io/research-highlights-2020`.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. arXiv:1907.05791.

Toby Shevlane and Allan Dafoe. 2020. The offense-defense balance of scientific knowledge: Does publishing AI research reduce misuse? *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*.

Irene Solaiman, M. Brundage, J. Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, A. Radford, and J. Wang. 2019. Release strategies and the social impacts of language models. arXiv:1908.09203.

Rachael Tatman. 2020. What i won't build. WiNLP WOrkshop @ ACL.

Trieu H. Trinh and Quoc V. Le. 2018. A simple method for commonsense reasoning. arXiv:1806.02847.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019a. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019b. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *the International Conference on Learning Representations*.

William Yang Wang, Samantha Finkelstein, Amy Ogan, Alan W Black, and Justine Cassell. 2012. "love ya, jerkface": Using sparse log-linear models to build positive and impolite relationships with teens. In *Proceedings of the 13th annual meeting of the special interest group on discourse and dialogue*, pages 20–29.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mT5: A massively multilingual pre-trained text-to-text transformer.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019a. Defending against neural fake news.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019b. Defending against neural fake news. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Yukun Zhu, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *ICCV*.

A Appendix

A.1 Tokenization

The SentencePiece tokenizer for T5 is described in Section 3.3.1 of Raffel et al. (2020). They train this tokenizer and generate their WordPieces and vocabulary from a 10:1:1:1 ratio of English:French:German:Romainian, for a total of 32,000 word pieces. This English vocabulary is generated from the cleaned English C4, and thus does not contain the tokens in the banned word list; this can lead to some unexpected tokenizations, such as "sex" being tokenized as "s" + "ex".

A.2 Patents from different patent offices

An example patent originally in Chinese: https://patents.google.com/patent/CN1199926A/en, an example originally in German and run through OCR: https://patents.google.com/patent/WO1998039809A1/en.

A.3 Classification label contamination

We observe that a large portion of GLUE (Wang et al., 2019b) and SuperGLUE (Wang et al., 2019a) datasets can be easily found on Github (see a list below). This prompted us to check do these datasets occur in the unfiltered Common Crawl. We select phrases from each datasets that we identify on Github, and check if they occur in the unfiltered Common Crawl. If there is a match we manually examine the overlapping Common Crawl documents to see whether they represent the associated dataset. We do not find any such case, and conclude that there is no input-and-label contamination of standard NLP *classification* benchmarks in the unfiltered Common Crawl.

- https://github.com/nyu-mll/CoLA-baselines/blob/master/acceptability_corpus/
- https://github.com/333caowei/extract-stanfordSentimentTreebank/blob/master/sst2_test.csv
- https://github.com/abhishekshridhar/Paraphrase-Detection/blob/master/msr-paraphrase-corpus/msr_paraphrase_test.txt
- https://github.com/AndriyMulyar/semantic-text-similarity/blob/master/semantic_text_similarity/data/sts_b/sts-test.csv
- https://raw.githubusercontent.com/qinxinlei/QNLI/master/glue_data/QNLI/dev.tsv
- https://github.com/himanshushivhare/RTE/blob/master/RTE3-TEST/RTE3-TEST.xml

Count	Country or WIPO Code	Country or Office Name	Language
70489	US	USA	English
4583	EP	European Patent Office	English, French, or German
4554	JP	Japan	Japanese
2283	CN	China	Chinese (Simplified)
2154	WO	World Intellectual Property Organization	Various
1554	KR	Republic of Korea	Korean
1417	CA	Canada	English
982	AU	Australia	English
747	GB	United Kingdom	English
338	DE	Germany	German
332	TW	Taiwan	Traditional Chinese
271	FR	France	French
138	MX	Mexico	Spanish
118	SE	Sweden	Swedish
711	Other	Various	Various

Table 3: The number of patents from different different patent offices from `patents.google.com`, the largest single Internet domain (in terms of tokens) for C4. Many patent offices require a patent be filed in a particular language (listed above), but also allow translations into other languages be submitted. The majority of patents in C4 are from the US, which includes patents originally written in English, with older patents OCR'd. "Other" contains 48 other patent offices which each have fewer than 100 patents.

- https://github.com/zdwls/boolqQA/blob/main/datafile/test.jsonl
- https://github.com/mcdm/CommitmentBank/blob/master/CommitmentBank-items.csv
- https://github.com/drwiner/COPA/blob/master/datasets/copa-test.xml
- https://raw.githubusercontent.com/eitanhaimashiah/multibidaf/master/data/multirc_dev.json
- https://github.com/aEE25/Testing-WiC-with-ERNIE/blob/main/WiC_dataset/test/test.data.txt
- https://github.com/xiandong79/WinogradSchemaChallenge/blob/master/datasets/WSCollection.xml