Do Language Models Care About Text Quality? Evaluating Web-Crawled Corpora Across 11 Languages

Rik van Noord, Taja Kuzman, Peter Rupnik, Nikola Ljubešić,
Miquel Esplà-Gomis, Gema Ramírez-Sánchez and Antonio Toral

LREC-COLING 2024

Introduction

Introduction

- LLMs are revolutionizing the field of NLP
- Multilingual corpora play a pivotal role
- Some of the largest corpora are web crawled: e.g. CC100 or mC4
- Divergent levels and strategies to clean corpora
  - How does this affect models trained on them?

Related work

Corpora are crucial to train LLM, but only a handful of papers focus on their quality.

We extend upon the findings of two previous works:

- *Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets* (Kreutzer et al., 2022)
  - manual auditing of samples from several multilingual corpora
  - the work shows serious quality issues, especially for low-resource languages
- *Does Corpus Quality Really Matter for Low-Resource Languages?* (Artetxe et al., 2022)
  - both manual and automatic quality assessment for Basque on several corpora
  - no clear correlation found between either the size or the quality of corpora in the performance of derived LMs

Approach: corpora

- Two evaluation strategies: **manual** and **automatic**
- Four corpora evaluated:
  - **CC100** (Conneau et al., 2020): used to build the XLM-R model
  - **mC4** (Xue et al., 2021): used to build the mT5 model
  - **OSCAR** (Ortiz Suárez et al., 2019)
  - **MaCoCu** (Bañón et al., 2022)
- Corpora used as published, **without additional filtering**
  - **Exception** of OSCAR best practice: we only select paragraphs that are recognised as being in the correct language

Comparison of corpora

Corpus	Langs.	Source	Lang. ID	Filtering
CC100	100	Common Crawl	`fastText`	Paragraph deduplication
mC4	101	Common Crawl	`cld3`	Paragraph deduplication, too long/short paragraphs, paragraphs containing bad words
OSCAR	152	Common Crawl	`TLSH`	Near duplicate removal, perplexity from LM trained on harmful content
MaCoCu	11	Web crawling	`cld2`	Near duplicate removal, badly encoded or too short paragraphs,

Approach: languages

- **Human** annotation of 11 languages: *Albanian*, Bosnian, *Bulgarian*, Croatian, *Icelandic*, *Macedonian*, Maltese, Montenegrin, *Serbian*, *Slovenian* and *Turkish*.
  - Underlined languages covered by the four corpora
- **Automatic** evaluation on 5: Albanian, Croatian, Icelandic, Serbian, Slovenian

Manual evaluation

Human annotation

- 200 paragraphs per corpus/language: **from 200 to 800** per language
- Two annotators per language
- Samples are shuffled and authors do not know their origin
- Each paragraph ranked using the following scale:
  1. Wrong language (or not natural language) [WL]
  2. Not running text [NR]
  3. Partially running text [PR]
  4. Running text, but slightly not standard [RT]
  5. Publishable text [PT]

Summarised results

Aggregated results for 7 languages shared across all corpora:

	WL	NR	PR	RT	PT	RT+PT
MaCoCu	1.8	3.6	10.4	29.9	54.3	**84.2**
CC100	0.4	6.9	13.3	26.2	53.2	79.4
mC4	6.4	13.9	16.0	22.4	41.3	**63.7**
OSCAR	0.6	5.4	9.9	24.7	59.4	**84.1**

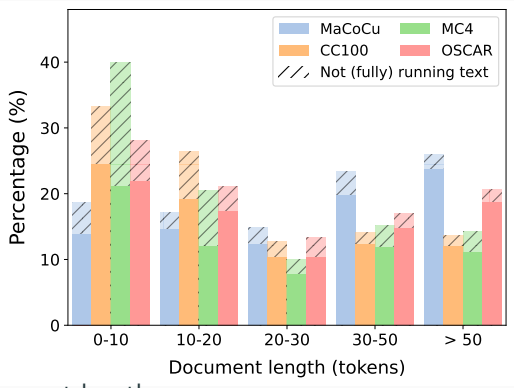
- Results per language in paper
- MaCoCu and OSCAR have the most RT+PT for most languages
- mC4 has the fewest RT+PT for almost all languages

Relation between length and quality

Are **longer text** fragments more prone to contain **not running text**?

Relation between length and quality

Are **longer text** fragments more prone to contain **not running text**?



- Results per text fragment length ranges
- For each corpus: percentage of running text (solid) vs. not fully running text (striped)
- Results confirm that longer texts contain mostly running text

Automatic evaluation

Automatic experiments

- Continue training XML-R base model with each corpus separately + all concatenated
- Models for Albanian, Croatian, Icelandic, Serbian and Slovenian
  - One single model trained for Croatian and Serbian, but evaluated separately
- Evaluation by fine-tuning on downstream tasks:
  - Part-of-Speech tagging (XPOS)
  - Named Entity Recognition (NER)
  - Choice of Plausible Alternatives (COPA)
  - Commitment Bank (CB)

Corpora sizes

Dataset sizes in GB of compressed text (in bold, largest corpus per language)

Language	CC100	MaCoCu	mC4	OSCAR	Comb
Albanian	2.1	1.4	**4.6**	0.9	9.0
Icelandic	1.3	1.6	**2.9**	0.6	6.3
Serbo-Croatian	10.8	**12.1**	4.9	1.4	29.2
Slovenian	4.2	4.7	**11.0**	0.4	20.1
Croatian	**8.6**	5.9	—	0.003	—
Serbian	2.2	**6.2**	4.9	1.4	—

Results (models fine-tuned for 50k steps)

- Table shows average ranking per task
- Position differences according to Mann-Whitney test (Mann and Whitney, 1947)

	hr	**sr**	**sq**	**is**	**sl**	**Avg.**
XLM-R	4.25	4.75	3.75	4.75	4.25	4.35
CC100	1.0	1.0	1.5	1.75	1.0	1.25
MaCoCu	2.0	3.0	2.75	1.0	3.25	2.40
mC4	3.5	3.5	2.5	2.75	1.75	2.80
OSCAR	4.0	3.5	2.5	2.75	2.5	3.05
Combined	1.0	1.5	3.25	1.75	2.25	1.95

- Surprisingly, CC100 performs best
- The other three models perform similarly, with OSCAR scoring the lowest; might be related to size?

Results with size controlled

- Same experiment with 10k instead of 50k steps to simulate size control (Muennighoff et al., 2023)

	hr	sr	sq	is	sl	Avg.
XLM-R	3.5	3.5	3.5	4.75	4.0	3.85
CC100	1.0	1.0	1.5	2.5	1.0	1.40
MaCoCu	1.5	2.5	2.0	2.75	4.25	2.60
mC4	2.25	2.5	2.25	3.0	2.5	2.50
OSCAR	3.75	2.5	2.0	3.0	2.25	2.70
Combined	1.0	1.0	2.25	1.5	2.25	1.60

- CC100 still performs best
- The other three models obtain closer results

Conclusions

Concluding remarks

- Our manual evaluation shows clear differences in quality between corpora
  - MaCoCu and OSCAR contain more high-quality texts
  - mC4 shows contains text with more issues (Maltese seems particularly problematic)
- These differences are not clearly reflected in the performance of the models trained
  - CC100 performs best
  - OSCAR performs worst, but with controlled size performance is similar to MaCoCu and mC4
- Impact of quality of corpora on derived models deserves further investigation
- Still open questions:
  - Unexpected performance of models trained on CC100
  - Low performance of mC4 taking into account it is much larger

Acknowledgement

Co-financed by the Connecting Europe Facility of the European Union

The MaCoCu project has received funding from the European Union's Connecting Europe Facility 2014-2020 - CEF Telecom, under Grant Agreement No. INEA/CEF/ICT/A2020/2278341. This communication reflects only the authors' views. The Agency is not responsible for any use that may be made of the information it contains.

References i

Artetxe, M., Aldabe, I., Agerri, R., Perez-de Viñaspre, O., and Soroa, A. (2022). Does corpus quality really matter for low-resource languages? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7383–7390, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Bañón, M., Esplà-Gomis, M., Forcada, M. L., García-Romero, C., Kuzman, T., Ljubešić, N., van Noord, R., Sempere, L. P., Ramírez-Sánchez, G., Rupnik, P., Suchomel, V., Toral, A., van der Werff, T., and Zaragoza, J. (2022). MaCoCu: Massive collection and curation of monolingual and bilingual data: focus on under-resourced languages. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 303–304, Ghent, Belgium. European Association for Machine Translation.

References ii

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Kreutzer, J., Caswell, I., Wang, L., Wahab, A., van Esch, D., Ulzii-Orshikh, N., Tapo, A., Subramani, N., Sokolov, A., Sikasote, C., Setyawan, M., Sarin, S., Samb, S., Sagot, B., Rivera, C., Rios, A., Papadimitriou, I., Osei, S., Suarez, P. O., Orife, I., Ogueji, K., Rubungo, A. N., Nguyen, T. Q., Müller, M., Müller, A., Muhammad, S. H., Muhammad, N., Mnyakeni, A., Mirzakhalov, J., Matangira, T., Leong, C., Lawson, N., Kudugunta, S., Jernite, Y., Jenny, M., Firat, O., Dossou, B. F. P., Dlamini, S., de Silva, N., Çabuk Ballı, S., Biderman, S., Battisti, A., Baruwa, A., Bapna, A., Baljekar, P., Azime, I. A., Awokoya, A., Ataman, D., Ahia, O., Ahia, O., Agrawal, S., and Adeyemi, M. (2022). Quality at a glance:

References iii

An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Mann, H. B. and Whitney, D. R. (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1):50 – 60.

Muennighoff, N., Rush, A., Barak, B., Le Scao, T., Tazi, N., Piktus, A., Pyysalo, S., Wolf, T., and Raffel, C. A. (2023). Scaling data-constrained language models. In Oh, A., Neumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, *Advances in Neural Information Processing Systems*, volume 36, pages 50358–50376. Curran Associates, Inc.

Ortiz Suárez, P. J., Sagot, B., and Romary, L. (2019). Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pages 9 – 16, Mannheim. Leibniz-Institut für Deutsche Sprache.

References iv

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2021). 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*, pages 483–498, Online. Association for Computational Linguistics.