MaCoCu: Massive collection and curation of monolingual and bilingual data: focus on under-resourced languages

Marta Bañón[†], Mălina Chichirău[♦], Miquel Esplà-Gomis[★], Mikel L. Forcada[★], Aarón Galiano-Jiménez[★], Taja Kuzman[‡], Nikola Ljubešić[‡], Rik van Noord[♦], Leopoldo Pla Sempere[★], Gema Ramírez-Sánchez[†], Peter Rupnik[‡], Vít Suchomel[‡], Antonio Toral[♦], Jaume Zaragoza[†]

[‡]Jožef Stefan Institute, [†]Prompsit, [♦]Rijksuniversiteit Groningen, [★]Universitat d'Alacant

[‡]{taja.kuzman,nikola.ljubesic,peter.rupnik}@ijs.si,
vit.suchomel@sketchengine.eu

[†]{mbanon,gramirez,jzaragoza}@prompsit.com

[♦]{r.i.k.van.noord,a.toral.ruiz,m.chichirau}@rug.nl

[★]{mespla,mlf,cgarcia,lpla}@dlsi.ua.es

Abstract

We present the most relevant results of the project *MaCoCu: Massive collection and curation of monolingual and bilingual data: focus on under-resourced languages* in its second year. Parallel and monolingual corpora have been produced for eleven low-resourced European languages by crawling large amounts of textual data from selected top-level domains of the Internet; both human and automatic evaluation show its usefulness. In addition, several large language models pretrained on MaCoCu data have been published, as well as the code used to collect and curate the data.

1 Introduction

This paper describes the main outcomes of the project *MaCoCu: Massive collection and curation of monolingual and bilingual data: focus on under-resourced languages* (Bañón et al., 2022), spanning from June 2021 to July 2023. MaCoCu is aimed at building large and high-quality monolingual and parallel (with English) corpora for ten low-resourced European languages (see Table 1). The international consortium behind this project consists of four partners: Jožef Stefan Institute (Slovenia), Rijksuniversiteit Groningen (Netherlands), Prompsit Language Engineering S.L. (Spain), and Universitat d'Alacant (Spain; coordinator).

Other existing initiatives, such as Paracrawl[1] or Oscar[2] exploit existing resources such as Common Crawl[3] or the Internet Archive.[4] Our strategy consists in automatically crawling top-level domains (TLD), potentially containing substantial amounts of text in the targeted languages,[5] and then applying a monolingual and a parallel curation pipelines. The evaluation of the first data release (van Noord et al., 2022a) confirms the usefulness of these data for different natural-language processing tasks.

2 Collected corpora

Monolingual and parallel corpora are built from crawled data by applying a thorough cleaning process, including noise fixing/filtering and removal of near-duplicate/boilerplate text. Corpora are then automatically annotated with: (a) document and paragraph IDs; (b) language variety (e.g. British/American English); (c) document-level affinity to DSIs identified through domain modelling (van Noord et al., 2022b); (d) personal information; and (e) identification of translated text: either human or machine translations (only for parallel corpora). Table 1 shows the size of the corpora for the second data release, published in April 2023.

2.1 Data evaluation

To the date, evaluation only covers the seven languages included in the first data release of the action, made public in April of 2022.

Mono-lingual A set of pre-trained language models (LMs)[6] has been built and released for Icelandic, Maltese and Bulgarian/Macedonian by continuing the training of multilingual XLM-RoBERTa-large (Conneau et al., 2020) using only MaCoCu data for all languages. These models outperform monolingual baselines, and XLM-R and large models on the POS, NER and COPA (Roemmele et al., 2011)

[1] https://paracrawl.eu/

[2] https://oscar-project.org/

[3] https://commoncrawl.org/

[4] https://archive.org/

[5] National TLDs such as .hr for Croatian, or .is for Icelandic, and also generic TLDs such as .com, .org, or .eu.

[6] https://huggingface.co/MaCoCu

Nurminen, Brenner, Koponen, Latomaa, Mikhailov, Schierl, Ranasinghe, Vanmassenhove, Vidal, Aranberri, Nunziatini, Escartín, Forcada, Popovic, Scarton, Moniz (eds.)
Proceedings of the 24th Annual Conference of the European Association for Machine Translation, p. 505–506
Tampere, Finland, June 2023.

	Monolingual		Parallel	
Language	Docs.	Words	Segs.	Words
Turkish	16.0	4344.9	1.6	89.2
Bulgarian	10.5	3506.2	1.8	72.1
Croatian	8.1	2363.7	2.3	99.5
Slovenian	6.3	1920.1	1.9	85.0
Macedonian	2.0	524.1	0.4	18.3
Icelandic	1.7	644.5	0.3	10.6
Maltese	0.5	347.9	0.9	53.9
Albanian	1.7	625.7	0.5	24.3
Serbian	7.5	2491.0	2.1	95.9
Montenegrin	0.6	161.4	0.2	11.2
Bosnian	2.8	730.3	0.5	22.2

Table 1: Sizes for corpora in the 2nd data release. Monolingual corpora are measured in millions of documents (Docs.) and millions of words. Parallel corpora are measured in millions of parallel segments (Segs.) and millions of words. Bosnian is a bonus language as it was not initially covered in the action.

	bg	**is**	**mk**	**mt**	**tr**
XLM-R-base	56.9	55.2	55.3	52.2	53.2
XLM-R-large	53.1	54.3	52.5	54.0	50.5
Monolingual LM	—	54.6	—	55.6	56.4
XLM-R + MaCoCu	54.6	59.6	55.6	54.4	58.5

Table 2: Test set COPA scores for baseline LMs compared to continuing training XLM-R-large on MaCoCu data.

evaluation tasks. Table 2 shows the results for the COPA test set, the most challenging evaluation task. For Bulgarian/Macedonian we also train an LM from scratch using the RoBERTa (Liu et al., 2019) architecture, dubbed BERTovski, which reached competitive performance with XLM-R.

Parallel Parallel data were extrinsically evaluated first training neural machine translation systems on large data sets available on OPUS[7] (ParaCrawl, CommonCrawl, Tilde), and comparing the results obtained when adding the MaCoCu data to the training set. Results show improved performance for all languages across different evaluation sets and metrics. These results were confirmed by human evaluation (van Noord et al., 2022a).

3 Free/open-source pipeline

The curation pipelines used to produce MaCoCu corpora, Monotextor[8] and Bitextor,[9] have been released under free/open-source licences. Crawling and corpora-enrichment software have been also released under the MaCoCu[10] GitHub organisation.

4 Acknowledgment

This action has received funding from the European Union's Connecting Europe Facility 2014-2020 - CEF Telecom, under Grant Agreement No. INEA/CEF/ICT/A2020/2278341. The contents of this publication are the sole responsibility of its authors and do not necessarily reflect the opinion of the European Union.

References

Bañón, Marta, Miquel Esplà-Gomis, Mikel L. Forcada, Cristian García-Romero, Taja Kuzman, Nikola Ljubešić, Rik van Noord, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Peter Rupnik, Vít Suchomel, Antonio Toral, Tobias van der Werff, and Jaume Zaragoza. 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 EAMT*, pages 303–304, Ghent, Belgium, June.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the ACL*, pages 8440–8451, Online, July.

Liu, Yinhan, 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*.

Roemmele, Melissa, Cosmin Bejan, and Andrew Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI Spring Symposium - Technical Report*, pages 90–95.

van Noord, Rik, Miquel Esplà-Gomis, Nikola Ljubešić, Taja Kuzman, Gema Ramírez-Sánchez, Peter Rupnik, and Antonio Toral. 2022a. MaCoCu Evaluation Report.

van Noord, Rik, Cristian García-Romero, Miquel Esplà-Gomis, Leopoldo Pla Sempere, and Antonio Toral. 2022b. Building domain-specific corpora from the web: the case of European digital service infrastructures. In *Proceedings of the BUCC Workshop within LREC 2022*, pages 23–32, Marseille, France, June.

[7] https://opus.nlpl.eu/

[8] https://github.com/bitextor/monotextor

[9] https://github.com/bitextor/bitextor

[10] https://github.com/macocu