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

Marta Bañón[†], 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[♦], 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,t.n.van.der.werff}@rug.nl

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

Abstract

We introduce the project *MaCoCu: Massive collection and curation of monolingual and bilingual data: focus on under-resourced languages*, funded by the Connecting Europe Facility, which is aimed at building monolingual and parallel corpora for under-resourced European languages. The approach followed consists of crawling large amounts of textual data from selected top-level domains of the Internet, and then applying a curation and enrichment pipeline. In addition to corpora, the project will release the free/open-source web crawling and curation software used.

1 Introduction

This paper describes the project *MaCoCu: Massive collection and curation of monolingual and bilingual data: focus on under-resourced languages*, funded by the Connecting Europe Facility in the 2020 CEF Telecom Call - Automated Translation (2020-EU-IA-0078).[1] This project started on June 1, 2021, and will last for two years. It is aimed at building large and high-quality monolingual and parallel (with English) corpora for five under-resourced official EU languages: Maltese, Bulgarian, Slovenian, Croatian, and Icelandic;[2] and for the languages of the five candidate states to become EU members: Turkish, Albanian, Macedonian, Montenegrin, and Serbian. Existing initiatives producing similar corpora, such as Paracrawl (Bañón et al., 2020) or Oscar (Abadji et al., 2022) exploit existing resources such as Common Crawl[3] or the Internet Archive.[4] In contrast, our strategy consists in automatically crawling top-level domains (TLD) with the potential to contain substantial amounts of textual data in the targeted languages,[5] and then applying a monolingual and a parallel curation pipelines on the downloaded data. This approach aims at obtaining more and higher-quality data than that available in existing compilations.[6]

One of the objectives of the project is to identify data relevant for Digital Service Infrastructures (DSIs). Our corpora will be enriched with information about the relevance of the data collected for ten DISs: e-Health, e-Justice, Online Dispute Resolution, Europeana, Open Data Portal, Business Registers Interconnection System, e-Procurement, Safer Internet, Cybersecurity, and Electronic Exchange of Social Security Information.

1.1 International consortium

Four partners are involved in this project: Institut Jožef Stefan (Slovenia), Rijksuniversiteit Groningen (Netherlands), Prompsit Language Engineering S.L. (Spain), and Universitat d'Alacant (Spain; coordinator). The consortium has a strong background in the task of building corpora, as several partners have been also part of the consortiums behind projects such as Paracrawl (Bañón et al., 2020), GoURMET (Birch et al., 2019), EuroPat[7] and Abu-MaTran (Toral et al., 2015).

[1] https://ec.europa.eu/inea/connecting-europe-facility/cef-telecom/2020-eu-ia-0078

[2] Maltese and Icelandic were chosen since they are especially under-resourced official EU languages; Bulgarian, Slovenian and Croatian were chosen due to the interest of the consortium on South-Slavic languages, a decision that extends previous efforts in the Abu-MaTran project (Toral et al., 2015).

[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] Preliminary automatic evaluation seem to confirm the quality of the data in the first data release (see Table 1).

[7] https://ec.europa.eu/inea/connecting-europe-facility/cef-telecom/2018-eu-ia-0061

2 Outcomes of the project

The main results of the project will be parallel and monolingual corpora, as well as the code used to build them. In this section, we briefly describe the most relevant features of these outcomes.

2.1 Corpora

The main goal of this project is to build monolingual and parallel corpora for the ten languages mentioned in Section 1. Since the project is aimed at producing high-quality corpora, a thorough cleaning process will be carried out, which will include automatic noise cleaning/fixing, removal of near-duplicates and irrelevant data, such as boilerplates, and automatic detection of machine translated content. The corpora produced will be enriched with:

- Identifiers that allow to re-construct the original paragraphs or documents from the segments in the corpora, enabling to leverage information beyond the sentence-level;
- Language variety (e.g. British/American English) for some covered languages;
- Document-level affinity to the DSIs covered, which will be automatically identified through domain modelling;
- Personal information identification, to allow final users to remove it for specific use cases;
- *Translationese*, or the identification of the translation direction (only for parallel data);
- Identification of machine translation (only for parallel data), so that such crawled documents can be filtered out by the user.

Currently, monolingual and parallel data have been released for seven out of the ten languages targeted. Table 1 provides information about the sizes of the current version of these corpora.

	Monolingual		Parallel	
Language	Docs.	Words	Segs.	Words
Turkish	16.0	4346.3	10.3	513.5
Bulgarian	10.5	3508.9	3.9	158.7
Croatian	7.3	2318.3	3.1	134.9
Slovene	5.8	1779.1	3.2	137.0
Macedonian	2.0	524.1	0.5	23.9
Icelandic	1.7	644.5	0.4	14.4
Maltese	0.5	347.9	1.2	69.6

Table 1: Sizes for the monolingual and parallel corpora for the first 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 in the language other than English.

2.2 Free/open-source pipeline

All the code developed within the project to crawl, curate and enrich the corpora built will be made available under free/open-source licences on MaCoCu[8] and Bitextor[9] GitHub organisations.[10]

3 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.

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

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

[10] Two code releases will be made, one at the end of the first year of the project, and the second one at the end of the project.

References

Abadji, Julien, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. Towards a Cleaner Document-Oriented Multilingual Crawled Corpus. *arXiv e-prints*, page arXiv:2201.06642, January.

Bañón, Marta, 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, July.

Birch, Alexandra, Barry Haddow, Ivan Tito, Antonio Valerio Miceli Barone, Rachel Bawden, Felipe Sánchez-Martínez, Mikel L. Forcada, Miquel Esplà-Gomis, Víctor Sánchez-Cartagena, Juan Antonio Pérez-Ortiz, Wilker Aziz, Andrew Secker, and Peggy van der Kreeft. 2019. Global under-resourced media translation (GoURMET). In *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*, pages 122–122, Dublin, Ireland, August.

Toral, Antonio, Tommi Pirinen, Andy Way, Raphaël Rubino, Gema Ramírez-Sánchez, Sergio Ortiz-Rojas, Víctor Sánchez-Cartagena, Jorge Ferrández-Tordera, Mikel Forcada, Miquel Espla-Gomis, Nikola Ljubešić, Filip Klubička, Prokopis Prokopidis, and Vassilis Papavassiliou. 2015. Automatic acquisition of machine translation resources in the AbuMaTran project. *Procesamiento del Lenguaje Natural*, (55):185–188.