CoCo4MT 2023

Proceedings of the Second Workshop on Corpus Generation and Corpus Augmentation for Machine Translation

September 5, 2023

Aim of the workshop

In this second version of the workshop on corpus generation and corpus augmentation for machine translation (CoCo4MT 2023), we attempt to further establish augmentation techniques that can be used for machine translation, especially in low-resource settings. Due to the overwhelming success with a variety of languages in CoCo4MT 2022[1], in this CoCo4MT workshop we further introduce unique low-resource languages like Urdi, Bengali, and Icelandic. Additionally, new machine learning techniques that based on segmentation, data mining, and deep learning are presented. As an extra addition, this year we introduce a shared task for the first time that focuses on the construction of corpora for machine translation.

The CoCo4MT 2023 submissions provide open source access to their code and corpus which is found directly in each submission. The CoCo4MT 2023 website[2] is available publicly. It contains all of the information for the previous year along with this year's workshop.

[1] Ortega, J. E., Carpuat, M., Chen, W., Kann, K., Lignos, C., Popovic, M., and Tafreshi, S., editors (2022). *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Workshop 2: Corpus Generation and Corpus Augmentation for Machine Translation).* Association for Machine Translation in the Americas.

[2] `https://sites.google.com/view/coco4mt`

Workshop scope and details

It is a well-known fact that machine translation systems, especially those that use deep learning, require massive amounts of data. Several resources for languages are not available in their human-created format. Some of the types of resources available are monolingual, multilingual, translation memories, and lexicons. Those types of resources are generally created for formal purposes such as parliamentary collections when parallel and more informal situations when monolingual. The quality and abundance of resources including corpora used for formal reasons is generally higher than those used for informal purposes. Additionally, corpora for low-resource languages, languages with less digital resources available, tends to be less abundant and of lower quality.

CoCo4MT is a workshop centered around research that focuses on manual and automatic corpus creation, cleansing, and augmentation techniques specifically for machine translation. We accept work that covers any language (including sign language) but we are specifically interested in those submissions that explici tly report on work with languages with limited existing resources (low-resource languages). Since techniques from high-resource languages are generally statistical in nature and could be used as generic solutions for any language, we welcome submissions on high-resource languages also.

CoCo4MT aims to encourage research on new and undiscovered techniques. We hope that the methods presented at this workshop will lead to the development of high-quality corpora that will in turn lead to high-performing MT systems and new dataset creation for multiple corpora. We hope that submissions will provide high-quality corpora that are available publicly for download and can be used to increase machine translation performance thus encouraging new dataset creation for multiple languages that will, in turn, provide a general workshop to consult for corpora needs in the future. The workshop's success will be measured by the following key performance indicators:

- Promotes the ongoing increase in quality of machine translation systems when measured by standard measurements,
- Provides a meeting place for collaboration from several research areas to increase the availability of commonly used corpora and new corpora,
- Drives innovation to address the need for higher quality and abundance of low-resource language data.

Topics of the workshop include but are not limited to:

- Difficulties with using existing corpora (e.g., political considerations or domain limitations) and their effects on final MT systems,
- Strategies for collecting new MT datasets (e.g., via crowdsourcing),
- Data augmentation techniques,
- Data cleansing and denoising techniques,
- Quality control strategies for MT data,
- Exploration of datasets for pretraining or auxiliary tasks for training MT systems.

This year, we also conducted the first CoCo4MT shared task, where we invited participants to develop and share their methods on identifying beneficial instances for machine translation without any existing

parallel data in a target language. The goal of the shared task was to encourage research on making the data curation process for machine translation more efficient, particularly for low-resource languages where collecting data to train high-performing MT systems is constrained by cost and scale. We used multi-way parallel data from the Bible to create training and evaluation data in nine languages, which are publicly available here: `https://github.com/ananyaganesh/coco4mt-shared-task`. We received two submissions to the shared task, and the details of both systems are published as part of the proceedings, along with the findings of the shared task.

Invited Speakers (listed alphabetically by first name)

We are happy our dear colleagues Jack Halpern, Manuel Mager, and Marta R. Costa-jussà have prepared talks on three important topics for CoCo4MT 2023.

Jack Halpern, The CJK Dictionary Institute

Jack Halpern, CEO of The CJK Dictionary Institute, is a lexicographer by profession. For sixteen years was engaged in the compilation of the New Japanese-English Character Dictionary, and as a research fellow at Showa Women's University (Tokyo), he was editor-in-chief of several kanji dictionaries for learners, which have become standard reference works. Jack Halpern, who has lived in Japan over 40 years, was born in Germany and has lived in six countries including France, Brazil, Japan, and the United States. An avid polyglot who specializes in Japanese and Chinese lexicography, he has studied 18 languages (speaks ten fluently) and has devoted several decades to the study of linguistics and lexicography. On a lighter note, Jack Halpern loves the sport of unicycling. Founder and long-time president of the International Unicycling Federation, he has promoted the sport worldwide and is a director of the Japan Unicycling Association. Currently, his passions are playing the quena and improving his Chinese, Esperanto, and Arabic.

Marta R. Costa-jussà, Meta AI

Marta R. Costa-jussà is a research scientist at Meta AI since February 2022. She received her PhD from the UPC in 2008. Her research experience is mainly in Machine Translation. She has worked at LIMSI-CNRS (Paris), Barcelona Media Innovation Center, Universidade de São Paulo, Institute for Infocomm Research (Singapore), Instituto Politécnico Nacional (Mexico), the University of Edinburgh and at Universitat Politècnica de Catalunya (UPC, Barcelona), co-leading the MT-UPC Group. She has participated in 18 European/Spanish research projects; she has organised 12 workshops in top venues and she has published more than 100 papers. She has been part of the Editorial Board of the Computer Speech and Language journal. She has received an ERC Starting Grant and two Google Faculty Research Awards (2018 and 2019).

Manuel Mager, AWS AI Labs

Manuel Mager is an Applied Scientist at AWS AI Labs, and completing his Ph.D. candidate at the University of Stuttgart, Germany. He graduated in informatics from the National Autonomous University of Mexico (UNAM) and did a Master's in Computer Science at the Metropolitan Autonomous University, Mexico (UAM). His research is focused on Natural Language Processing for low resource languages, mainly indigenous languages of the American continent that are polysynthetic. He also worked on Graph-to-text generation and information extraction.

Other speakers and guests Due to its previous success, CoCo4MT will once again host a panel that includes several other researchers and notable speakers. The panel speakers will be announced in a future (post-edited) version of the proceedings.

Organizers

John E. Ortega, Northeastern University
Marine Carpuat, University of Maryland
William Chen, Carnegie Mellon University
Ananya Ganesh, University of Colorado Boulder
Katharina Kann, University of Colorado Boulder
Constantine Lignos, Brandeis University
Jonne Saleva, Brandeis University
Shabnam Tafreshi, University of Maryland
Rodolfo Zevallo, Universitat Pompeu Fabra

Program Committee (listed alphabetically by first name)

Abteen Ebrahimi, University of Colorado Boulder
Ananya Ganesh, University of Colorado Boulder
Bharathi Raja Chakravarthi, National University of Ireland Galway
Bonaventure F. P. Dossou, McGill University
Constantine Lignos, Brandeis University
Flammie Pirinen, UiT Norgga árktalaš universitehta
Jasper Kyle Catapang, University of Birmingham
John E. Ortega, Northeastern University
Jonne Sälevä, Brandeis University
Katharina Kann, University of Colorado Boulder
Kochiro Watanabe, The University of Tokyo
Koel Dutta Chowdhury, Saarland University
Majid Latifi, University of York
Maria Art Antonette Clariño, University of the Philippines Los Baños
Marine Carpuat, University of Maryland
Miquel Esplà-Gomis, Universitat d'Alacant
Pablo Gamallo, University of Santiago de Compostela - CITIUS
Patrick Simianer, Lilt
Rico Sennrich, University of Zurich
Rodolfo Zevallos, Universitat Pompeu Fabra
Sangjee Dondrub, Qinghai Normal University
Santanu Pal, Wipro
Shabnam Tafreshi, University of Maryland
Shiran Dudy, Northeastern University
Surafel Melaku Lakew, Amazon
Thepchai Supnithi, National Electronics and Computer Technology Center
William Chen, Carnegie Mellon University

Table of Contents

Do Not Discard – Extracting Useful Fragments from Low-Quality Parallel Data to Improve Machine Translation
Steinþór Steingrímsson, Pintu Lohar, Hrafn Loftsson and Andy Way 1

Development of Urdu-English Religious Domain Parallel Corpus
Sadaf Abdul Rauf and Noor e Hira 14

Findings of the CoCo4MT 2023 Shared Task on Corpus Construction for Machine Translation
Ananya Ganesh, Marine Carpuat, William Chen, Katharina Kann, Constantine Lignos, John E. Ortega, Jonne Saleva, Shabnam Tafreshi and Rodolfo Zevallos 22

Williams College's Submission for the Coco4MT 2023 Shared Task
Alex Root and Mark Hopkins 28

The AST Submission for the CoCo4MT 2023 Shared Task on Corpus Construction for Low-Resource Machine Translation
Steinþór Steingrímsson 33

Workshop Program

Do Not Discard – Extracting Useful Fragments from Low-Quality Parallel Data to Improve Machine Translation
Steinþór Steingrímsson, Pintu Lohar, Hrafn Loftsson and Andy Way

Development of Urdu-English Religious Domain Parallel Corpus
Sadaf Abdul Rauf and Noor e Hira

Findings of the CoCo4MT 2023 Shared Task on Corpus Construction for Machine Translation
Ananya Ganesh, Marine Carpuat, William Chen, Katharina Kann, Constantine Lignos, John E. Ortega, Jonne Saleva, Shabnam Tafreshi and Rodolfo Zevallos

Williams College's Submission for the Coco4MT 2023 Shared Task
Alex Root and Mark Hopkins

The AST Submission for the CoCo4MT 2023 Shared Task on Corpus Construction for Low-Resource Machine Translation
Steinþór Steingrímsson