NLP-OSS 2023

The 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS)

Proceedings of the Workshop

December 6, 2023

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 979-8-89176-045-5

Program Committee

Program Chairs

Geeticka Chauhan
Jeremy Gwinnup
Dmitrijs Milajevs
Elijah Rippeth
Liling Tan

Reviewers

Sina Ahmadi, Zaid Alyafeai, Abhinav Arora

Guillaume Becquin, Steven Bethard, Tenzin Singhay Bhotia, Francis Bond, Daniel Braun

Geeticka Chauhan, Won Ik Cho, Marco Cognetta

Steve DeNeefe, Gérard Dupont

Ignatius Ezeani

Michael Wayne Goodman, Jeremy Gwinnup, Jana Götze

David M Howcroft, Phu Mon Htut

Cassandra L Jacobs

Thomas H Kober, Philipp Koehn

Arun Balajiee Lekshmi Narayanan, Pasquale Lisena

Nitin Madnani, Shubhanshu Mishra, Wafaa Mohammed, Anish Mohan, John Xavier Morris

Aakanksha Naik

Odunayo Ogundepo, Atul Kr Ojha, Akintunde Oladipo

Aline Paes, Flammie A Pirinen, Matt Post

Elijah Rippeth, Alexander M Rush

Lane Schwartz, Micah Shlain, Mallika Singh, Sudhakar Singh, Aitor Soroa, Shilpa Suresh

Liling Tan, Raphael Tang, Christoph Teichmann, Tommaso Teofili, Jörg Tiedemann, Vijay Murari Tiyyala, Atnafu Lambebo Tonja

Taha Zerrouki

Table of Contents

Program

Thursday, May 26, 2022

09:15 - 10:15 *Invited Talk 1*

SEA-LION (Southeast Asian Languages In One Network): A Family of Southeast Asian Language Models
William Tjhi, David Ong and Peerat Limkonchotiwat

10:30 - 11:00 *Coffee Break*

11:00 - 11:30 *Lightning Session 1*

11:30 - 12:15 *Poster Session 1*

Jina Embeddings: A Novel Set of High-Performance Sentence Embedding Models
Michael Günther, Georgios Mastrapas, Bo Wang, Han Xiao and Jonathan Geuter

Deepparse : An Extendable, and Fine-Tunable State-Of-The-Art Library for Parsing Multinational Street Addresses
David Beauchemin

PyThaiNLP: Thai Natural Language Processing in Python
Wannaphong Phatthiyaphaibun, Korakot Chaovavanich, Charin Polpanumas, Arthit Suriyawongkul, Lalita Lowphansirikul, Pattarawat Chormai, Peerat Limkonchotiwat, Thanathip Suntorntip and Can Udomcharoenchaikit

Zelda Rose: a tool for hassle-free training of transformer models
Loïc Grobol

Kani: A Lightweight and Highly Hackable Framework for Building Language Model Applications
Andrew Zhu, Liam Dugan, Alyssa Hwang and Chris Callison-Burch

Beyond the Repo: A Case Study on Open Source Integration with GECToR
Sanjna Kashyap, Zhaoyang Xie, Kenneth Steimel and Nitin Madnani

Two Decades of the ACL Anthology: Development, Impact, and Open Challenges
Marcel Bollmann, Nathan Schneider, Arne Köhn and Matt Post

Thursday, May 26, 2022 (continued)

nanoT5: Fast & Simple Pre-training and Fine-tuning of T5 Models with Limited Resources
Piotr Nawrot

AWARE-TEXT: An Android Package for Mobile Phone Based Text Collection and On-Device Processing
Salvatore Giorgi, Garrick Sherman, Douglas Bellew, Sharath Chandra Guntuku, Lyle Ungar and Brenda Curtis

SOTASTREAM: A Streaming Approach to Machine Translation Training
Matt Post, Thamme Gowda, Roman Grundkiewicz, Huda Khayrallah, Rohit Jain and Marcin Junczys-Dowmunt

An Open-source Web-based Application for Development of Resources and Technologies in Underresourced Languages
Siddharth Singh, Shyam Ratan, Neerav Mathur and Ritesh Kumar

Rumour Detection in the Wild: A Browser Extension for Twitter
Andrej Jovanovic and Björn Ross

12:15 - 13:45 *Lunch Break*

13:45 - 14:45 *Invited Talk 2*

trlX: A Framework for Large Scale Open Source RLHF
Louis Castricato

14:45 - 15:15 *Lightning Session 2*

15:15 - 15:30 *Coffee Break*

15:30 - 16:15 *Poster Session 2*

GPT4All: An Ecosystem of Open Source Compressed Language Models
Yuvanesh Anand, Zach Nussbaum, Adam Treat, Aaron Miller, Richard Guo, Benjamin M Schmidt, Brandon Duderstadt and Andriy Mulyar

DeepZensols: A Deep Learning Natural Language Processing Framework for Experimentation and Reproducibility
Paul Landes, Barbara Di Eugenio and Cornelia Caragea

Improving NER Research Workflows with SeqScore
Constantine Lignos, Maya Kruse and Andrew Rueda

torchdistill Meets Hugging Face Libraries for Reproducible, Coding-Free Deep Learning Studies: A Case Study on NLP
Yoshitomo Matsubara

Using Captum to Explain Generative Language Models
Vivek Miglani, Aobo Yang, Aram H. Markosyan, Diego Garcia-Olano and Narine Kokhlikyan

nerblackbox: A High-level Library for Named Entity Recognition in Python
Felix Stollenwerk

News Signals: An NLP Library for Text and Time Series
Chris Hokamp, Demian Gholipour Ghalandari and Parsa Ghaffari

PyTAIL: An Open Source Tool for Interactive and Incremental Learning of NLP Models with Human in the Loop for Online Data
Shubhanshu Mishra and Jana Diesner

GPTCache: An Open-Source Semantic Cache for LLM Applications Enabling Faster Answers and Cost Savings
Fu Bang

The Vault: A Comprehensive Multilingual Dataset for Advancing Code Understanding and Generation
Dung Nguyen Manh, Nam Le Hai, Anh T. V. Dau, Anh Minh Nguyen, Khanh Nghiem, Jin Guo and Nghi D. Q. Bui

16:15 - 17:15 *Invited Talk 3*

Towards Explainable and Accessible AI
Brandon Duderstadt and Yuvanesh Anand

Wednesday, December 6, 2023

09:00 - 09:15 *Opening Remarks*

17:15 - 17:30 *Closing Remarks*