3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)

Singapore
6 December 2023

ISBN: 978-1-7138-8605-1

Printed from e-media with permission by:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571

Some format issues inherent in the e-media version may also appear in this print version.

Association for Computational Linguistics
209 N. Eighth Street
Stroudsburg, Pennsylvania 18360

Phone: 1-570-476-8006
Fax: 1-570-476-0860

acl@aclweb.org

Additional copies of this publication are available from:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571 USA
Phone: 845-758-0400
Fax: 845-758-2633
Email: curran@proceedings.com
Web: www.proceedings.com

Table of Contents

calamanCy: A Tagalog Natural Language Processing Toolkit
Lester James Validad Miranda 1

Jina Embeddings: A Novel Set of High-Performance Sentence Embedding Models
Michael Günther, Georgios Mastrapas, Bo Wang, Han Xiao and Jonathan Geuter 8

Deepparse : An Extendable, and Fine-Tunable State-Of-The-Art Library for Parsing Multinational Street Addresses
David Beauchemin 19

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 25

Empowering Knowledge Discovery from Scientific Literature: A novel approach to Research Artifact Analysis
Petros Stavropoulos, Ioannis Lyris, Natalia Manola, Ioanna Grypari and Haris Papageorgiou . . 37

Zelda Rose: a tool for hassle-free training of transformer models
Loïc Grobol 54

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 59

Kani: A Lightweight and Highly Hackable Framework for Building Language Model Applications
Andrew Zhu, Liam Dugan, Alyssa Hwang and Chris Callison-Burch 65

Beyond the Repo: A Case Study on Open Source Integration with GECToR
Sanjna Kashyap, Zhaoyang Xie, Kenneth Steimel and Nitin Madnani 78

Two Decades of the ACL Anthology: Development, Impact, and Open Challenges
Marcel Bollmann, Nathan Schneider, Arne Köhn and Matt Post 83

nanoT5: Fast & Simple Pre-training and Fine-tuning of T5 Models with Limited Resources
Piotr Nawrot 95

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 102

SOTASTREAM: A Streaming Approach to Machine Translation Training
Matt Post, Thamme Gowda, Roman Grundkiewicz, Huda Khayrallah, Rohit Jain and Marcin Junczys-Dowmunt 110

An Open-source Web-based Application for Development of Resources and Technologies in Underresourced Languages
Siddharth Singh, Shyam Ratan, Neerav Mathur and Ritesh Kumar 120

Rumour Detection in the Wild: A Browser Extension for Twitter
Andrej Jovanovic and Björn Ross 130

DeepZensols: A Deep Learning Natural Language Processing Framework for Experimentation and Reproducibility
Paul Landes, Barbara Di Eugenio and Cornelia Caragea 141

Improving NER Research Workflows with SeqScore
Constantine Lignos, Maya Kruse and Andrew Rueda 147

torchdistill Meets Hugging Face Libraries for Reproducible, Coding-Free Deep Learning Studies: A Case Study on NLP
Yoshitomo Matsubara 153

Using Captum to Explain Generative Language Models
Vivek Miglani, Aobo Yang, Aram H. Markosyan, Diego Garcia-Olano and Narine Kokhlikyan 165

nerblackbox: A High-level Library for Named Entity Recognition in Python
Felix Stollenwerk 174

News Signals: An NLP Library for Text and Time Series
Chris Hokamp, Demian Gholipour Ghalandari and Parsa Ghaffari 179

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 190

Antarlekhaka: A Comprehensive Tool for Multi-task Natural Language Annotation
Hrishikesh Terdalkar and Arnab Bhattacharya 199

GPTCache: An Open-Source Semantic Cache for LLM Applications Enabling Faster Answers and Cost Savings
Fu Bang 212

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 219

SEA-LION (Southeast Asian Languages In One Network): A Family of Southeast Asian Language Models
William Tjhi, David Ong and Peerat Limkonchotiwat 245

trlX: A Framework for Large Scale Open Source RLHF
Louis Castricato 246

Towards Explainable and Accessible AI
Brandon Duderstadt and Yuvanesh Anand 247