EMNLP 2020

The Second Workshop for NLP Open Source Software (NLP-OSS)

Proceedings of the Workshop

November 19, 2020
(Online)

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 978-1-952148-71-2

Introduction

With great scientific breakthrough comes solid engineering and open communities. The Natural Language Processing (NLP) community has benefited greatly from the open culture in sharing knowledge, data, and software. The primary objective of this workshop is to further the sharing of insights on the engineering and community aspects of creating, developing, and maintaining NLP open source software (OSS), which we seldom talk about in scientific publications. Our secondary goal is to promote synergies between different open source projects and encourage cross-software collaborations and comparisons.

We refer to Natural Language Processing OSS as an umbrella term that not only covers traditional syntactic, semantic, phonetic, and pragmatic applications; we extend the definition to include task-specific applications (e.g., machine translation, information retrieval, question-answering systems), low-level string processing that contains valid linguistic information (e.g. Unicode creation for new languages, language-based character set definitions) and machine learning/artificial intelligence frameworks with functionalities focusing on text applications.

In the earlier days of NLP, linguistic software was often monolithic and the learning curve to install, use, and extend the tools was steep and frustrating. More often than not, NLP OSS developers/users interact in siloed communities within the ecologies of their respective projects. In addition to the engineering aspects of NLP software, the open source movement has brought a community aspect that we often overlook in building impactful NLP technologies.

An example of precious OSS knowledge comes from SpaCy developer Montani (2017), who shared her thoughts and challenges of maintaining commercial NLP-OSS, such as handling open issues on the issue tracker, model release and packaging strategy and monetizing NLP OSS for sustainability.[1]

More recently, the Transformers library created by Hugging Face, has gathered much interest from the community by open sourcing implementations to use pretrained weights of BERT-like models, in a clean and well-organized structure. The interoperability of various pretrained models trained with different tools in one library enables quick benchmarking across the models, as well as developing best practices for reading/saving serialized interoperable.[2]

We hope that the NLP-OSS workshop becomes the intellectual forum to collate various open source knowledge beyond the scientific contribution, announce new software/features, promote the open source culture and best practices that go beyond the conferences.

[1] https://ines.io/blog/spacy-commercial-open-source-nlp

[2] models.https://github.com/huggingface/transformers

Organizers:

Lucy Park, NAVER Corp.
Masato Hagiwara, Octanove Labs LLC
Dmitrijs Milajevs, KPMG LLP
Nelson Liu, Stanford University
Geeticka Chauhan, Massachusetts Institute of Technology
Liling Tan, Rakuten Institute of Technology

Program Committee:

Aline Paes, Universidade Federal Fluminense
Amandalynne Paullada, University of Washington
Amittai Axelrod, DiDi Chuxing (Los Angeles)
Anca Dumitrache, FD Mediagroep
Arwen Twinkle Griffioen, Zendesk Inc.
Carolina Scarton, University of Sheffield
Chris Hokamp, AYLIEN
Christian Federmann, Microsoft Research
Dan Simonson, BlackBoiler, LLC
Daniel Braun, TU Muchen
Dave Howcroft, Heriot-Watt University
David Przybilla, Idio
Delip Rao, AI Foundation
Denny Britz, Prediction Machines
Ehsan Khoddammohammadi, Elsiever
Eleftherios Avramidis, German Research Center for Artificial Intelligence
Elijah Rippeth, MITRE Corporation
Emiel van Miltenburg, Vrije Universiteit Amsterdam
Emily Dinan, Facebook AI
Eric Schles, New York University & Sema4
Fabio Kepler, Unbabel
Francis Bond, Nanyang Technological University
Fred Blain, University of Sheffield
Gerard Dupont, Airbus
Ian Soboroff, NIST
Ignatius Ezeani, Lancaster Uni
Ines Montani, Explosion AI
James Bradbury, Google
Joel Nothman, University of Sydney
Karin Sim Smith, Lingo24
Kevin Cohen, University of Colorado Boulder
KhengHui Yeo, Institute for Infocomm Research
Laura Martinus, Explore AI

Madison May, Indico Data Solutions
Marcel Bollmann, University of Copenhagen
Marcos Zampieri, University of Wolverhampton
Mary Ellen Foster, University of Glasgow
Marzieh Fadaee, University of Amsterdam
Matthew Honnibal, Explosion AI
Micah Shlain, Allen Institute for Artificial Intelligence
Michael Wayne Goodman, Nanyang Technological University
Mohd Sanad Zaki Rizvi, Microsoft Research India
Moshe Wasserblat, Intel
Muthu Kumar Chandrasekaran, NUS, SG
Nahid Alam, Ople Inc
Paul P Liang, Carnegie Mellon University
Philipp Koehn, Johns Hopkins University
Sandya Mannarswamy , Independent Researcher
Shamil Chollampatt, Rakuten Institute of Technology
Sharat Chikkerur, Microsoft
Shilpa Suresh, Singapore Managment University
Shubhanshu Mishra, Twitter
Steve DeNeefe, SDL Research Labs
Steve Sloto, AWS AI
Steven Bethard, University of Arizona
Steven Bird, Charles Darwin University
Sung Kim, NAVER Corp.
Svitlana Vakulenko, University of Amsterdam
Tareq Al-Moslmi, University of Bergen
Thomas Kober, Rasa Technologies GmbH
Tilahun Abedissa, Addis Ababa University
Tommaso Teofili, Roma Tre University & Red Hat
Tommi A Pirinen, University of Hamburg
Varun Kumar, Amazon Alexa
Vlad Niculae, Instituto de Telecomunicações
Yves Peirsman, NLP Town

Invited Speaker:

Chip Huyen, Stanford & Snorkel AI
Spencer Kelly, Freelance Developer
Thomas Wolf, Huggingface

Invited Talks

Principles of Good Machine Learning Systems Design
Chip Huyen, Stanford & Snorkel AI

On Typing: Historical and Potential Interactions in Word-processing
Spencer Kelly, Freelance Developer

An Introduction to Transfer Learning in NLP and HuggingFace
Thomas Wolf, Huggingface

Principles of Good Machine Learning Systems Design

Chip Huyen
Stanford & Snorkel AI

Abstract

This talk covers what it means to operationalize Machine Learning (ML) models. It starts by analyzing the difference between ML in research vs. in production, ML systems vs. traditional software, as well as myths about ML production.

It then goes over the principles of good ML systems design and introduces an iterative framework for ML systems design, from scoping the project, data management, model development, deployment, maintenance, to business analysis. It covers the differences between DataOps, ML Engineering, MLOps, and data science, and where each fits into the framework.

The talk ends with a survey of the ML production ecosystem, the economics of open source, and open-core businesses.

Biography

Chip Huyen is an engineer who develops tools and best practices for machine learning production. She's currently with Snorkel A and she'll be teaching Machine Learning Systems Design at Stanford. Previously, she was with Netflix, NVIDIA, Primer. She's also the author of four best-selling Vietnamese books.

On Typing: Historical and Potential Interactions in Word-processing

Spencer Kelly
Freelance Developer

Abstract

People love typing, in a surprising and universal way. In this talk we look at the development of word-processing, and the design-decisions in this historic interface. Can NLP contribute to word-processing, without making it worse? What would a text-centered computer really look like? We look at the history of punctuation, keyboards, and markup languages. We look at Wikipedia, text-editors, and data structures - with the goal of authoring usable data in text.

Biography

Spencer is the author of compromise - a small natural language processing library for the browser. He is a web developer, and maintainer of open-source libraries. His background is in the semantic web and Wikipedia. Today his work focuses on creating infographics. His open-source work is funded by freelance web development. He is from Toronto, Canada.

An Introduction to Transfer Learning in NLP and HuggingFace

Thomas Wolf
Huggingface

Abstract

In this talk I'll start by introducing the recent breakthroughs in NLP that resulted from the combination of Transfer Learning schemes and Transformer architectures. The second part of the talk will be dedicated to an introduction of the open-source tools released by HuggingFace, in particular our Transformers, Tokenizers and Datasets libraries and our models.

Biography

Thomas Wolf is co-founder and Chief Science Officer of HuggingFace. His team is on a mission to catalyze and democratize NLP research. Prior to HuggingFace, Thomas gained a Ph.D. in physics, and later a law degree. He worked as a physics researcher and a European Patent Attorney.

Table of Contents

Workshop Program

A Framework to Assist Chat Operators of Mental Healthcare Services
Thiago Madeira, Heder Bernardino, Jairo Francisco De Souza, Henrique Gomide, Nathália Munck Machado, Bruno Marcos Pinheiro da Silva and Alexandre Vieira Pereira Pacelli

ARBML: Democritizing Arabic Natural Language Processing Tools
zaid alyafeai and Maged Al-Shaibani

CLEVR Parser: A Graph Parser Library for Geometric Learning on Language Grounded Image Scenes
Raeid Saqur and Ameet Deshpande

End-to-end NLP Pipelines in Rust
Guillaume Becquin

Fair Embedding Engine: A Library for Analyzing and Mitigating Gender Bias in Word Embeddings
Vaibhav Kumar, Tenzin Bhotia and Vaibhav Kumar

Flexible retrieval with NMSLIB and FlexNeuART
Leonid Boytsov and Eric Nyberg

fugashi, a Tool for Tokenizing Japanese in Python
Paul McCann

Going Beyond T-SNE: Exposing whatlies in Text Embeddings
Vincent Warmerdam, Thomas Kober and Rachael Tatman

Howl: A Deployed, Open-Source Wake Word Detection System
Raphael Tang, Jaejun Lee, Afsaneh Razi, Julia Cambre, Ian Bicking, Jofish Kaye and Jimmy Lin

iNLTK: Natural Language Toolkit for Indic Languages
Gaurav Arora

jiant: A Software Toolkit for Research on General-Purpose Text Understanding Models [non-archival, published at ACL 2020]
Yada Pruksachatkun, Phil Yeres, Haokun Liu, Jason Phang, Phu Mon Htut, Alex Wang, Ian Tenney and Samuel R. Bowman

KLPT – Kurdish Language Processing Toolkit
Sina Ahmadi

No Day Set (continued)

Open Korean Corpora: A Practical Report
Won Ik Cho, Sangwhan Moon and Youngsook Song

Open-Source Morphology for Endangered Mordvinic Languages
Jack Rueter, Mika Hämäläinen and Niko Partanen

Pimlico: A toolkit for corpus-processing pipelines and reproducible experiments
Mark Granroth-Wilding

PySBD: Pragmatic Sentence Boundary Disambiguation
Nipun Sadvilkar and Mark Neumann

iobes: Library for Span Level Processing
Brian Lester

SacreROUGE: An Open-Source Library for Using and Developing Summarization Evaluation Metrics
Daniel Deutsch and Dan Roth

TextAttack: Lessons learned in designing Python frameworks for NLP
John Morris, Jin Yong Yoo and Yanjun Qi

TOMODAPI: A Topic Modeling API to Train, Use and Compare Topic Models
Pasquale Lisena, Ismail Harrando, Oussama Kandakji and Raphael Troncy

User-centered & Robust NLP OSS: Lessons Learned from Developing & Maintaining RSMTool
Nitin Madnani and Anastassia Loukina

WAFFLE: A Graph for WordNet Applied to FreeForm Linguistic Exploration
Berk Ekmekci and Blake Howald