ACL 2018

NLP Open Source Software (NLP-OSS)

Proceedings of the Workshop

July 20, 2018
Melbourne, Australia

Gold Sponsors

NYU Center for Data Science

kaggle

Silver Sponsors

NAVER

Google

Bronze Sponsors

zendesk

©2018 The Association for Computational Linguistics

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 978-1-948087-38-4

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.

There are many workshops focusing open language resource/annotation creation and curation (e.g. BUCC, GWN, LAW, LOD, WAC). Moreover, we have the flagship LREC conference dedicated to linguistic resources. However, the engineering aspects of NLP OSS is overlooked and under-discussed within the community. There are open source conferences and venues (such as FOSDEM, OSCON, Open Source Summit) where discussions range from operating system kernels to air traffic control hardware but the representation of NLP related presentations is limited. In the Machine Learning (ML) field, the Journal of Machine Learning Research - Machine Learning Open Source Software (JMLR-MLOSS) is a forum for discussions and dissemination of ML OSS topics. We envision that the Workshop for NLP-OSS becomes a similar avenue for NLP OSS discussions.

To our best knowledge, this is the first workshop proposal in the recent years that focuses more on the building aspect of NLP and less on scientific novelty or state-of-art development. A decade ago, there was the SETQA-NLP (Software Engineering, Testing, and Quality Assurance for Natural Language Processing) workshop that raised awareness of the need for good software engineering practices in NLP. 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 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]

Řehůřek (2017) shared another example of insightful discussion on bridging the gap between the gap between academia and industry through creating open source and student incubation programs. Řehůřek discussed the need to look beyond the publish-or-perish culture to avoid the brittle "mummy effect" in SOTA research code/techniques.[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] https://rare-technologies.com/mummy-effect-bridging-gap-between-academia-industry/

Organizers:

Lucy Park, NAVER Corp.
Masato Hagiwara, Duolingo Inc.
Dmitrijs Milajevs, NIST and Queen Mary University of London
Liling Tan, Rakuten Institute of Technology

Program Committee:

Martin Andrews, Red Cat Labs
Steven Bird, Charles Darwin University
Francis Bond, Nanyang Technological University
Jason Baldridge, Google
Steven Bethard, University of Arizona
Fred Blain, University of Sheffield
James Bradbury, Salesforce Research
Denny Britz, Prediction Machines
Marine Carpuat, University of Maryland
Kyunghyun Cho, New York University
Grzegorz Chrupała, Tilburg University
Hal Daumé III, University of Maryland
Jon Dehdari, Think Big Analytics
Christian Federmann, Microsoft Research
Mary Ellen Foster, University of Glasgow
Michael Wayne Goodman, University of Washington
Arwen Twinkle Griffioen, Zendesk Inc.
Joel Grus, Allen Institute for Artificial Intelligence
Chris Hokamp, Aylien Inc.
Matthew Honnibal, Explosion AI
Sung Kim, Hong Kong University of Science and Technology
Philipp Koehn, Johns Hopkins University
Taku Kudo, Google
Christopher Manning, Stanford University
Diana Maynard, University of Sheffield
Tomas Mikolov, Facebook AI Research (FAIR)
Ines Montani, Explosion AI
Andreas Müller, Columbia University
Graham Neubig, Carnegie Mellon University
Vlad Niculae, Cornell CIS
Joel Nothman, University of Sydney
Matt Post, Johns Hopkins University
David Przybilla, Idio
Amandalynne Paullada, University of Washington

Delip Rao, Joostware AI Research Corp
Radim Řehůřek, RaRe Technologies
Elijah Rippeth, MITRE Corporation
Abigail See, Stanford University
Carolina Scarton, University of Sheffield
Rico Sennrich, University of Edinburgh
Dan Simonson, Georgetown University
Vered Shwartz, Bar-Ilan University
Ian Soboroff, NIST
Pontus Stenetorp, University College London
Rachael Tatman, Kaggle
Tommaso Teofili, Adobe
Emiel van Miltenburg, Vrije Universiteit Amsterdam
Maarten van Gompel, Radboud University
Gaël Varoquaux, INRIA
KhengHui Yeo, Institute for Infocomm Research
Marcos Zampieri, University of Wolverhampton

Invited Speaker:

Joel Nothman, University of Sydney
Christopher Manning, Stanford University
Matthew Honnibal and Ines Montani, Explosion AI

Invited Talks

Open-Source Software's Responsibility to Science
Joel Nothman, University of Sydney

Stanford CoreNLP: 15 Years of Developing Academic Open Source Software
Christopher Manning, Stanford University

Reflections on Running spaCy: Commercial Open-source NLP
Matthew Honnibal and Ines Montani, Explosion AI

Open-Source Software's Responsibility to Science

Joel Nothman
University of Sydney

Abstract

Open-source software makes sophisticated technologies available to a wide audience. Arguably, most people applying language processing and machine learning techniques rely on popular open source tools targeted at these applications. Users may themselves be incapable of implementing the underlying algorithms. Users may or may not have extensive training to critically conduct experiments with these tools.

As maintainers of popular scientific software, we should be aware of our user base, and consider the ways in which our software design and documentation can lead or mislead users with respect to scientific best practices. In this talk, I will present some examples of these risks, primarily drawn from my experience developing Scikit-learn. For example: How can we help users avoid data leakage in cross-validation? How can we help users report precisely which algorithm or metric was used in an experiment?

Volunteer OSS maintainers have limited ability to see and manage these risks, and need the scientific community's assistance to get things right in design, implementation and documentation.

Biography

Joel Nothman began contributing to the Scientific Python ecosystem of open-source software as a research student at the University of Sydney in 2008. He has since made substantial contributions to the NLTK, Scipy, Pandas and IPython packages among others, but presently puts most of his open-source energies into maintaining Scikit-learn, a popular machine learning toolkit. Joel works as a data science research engineer at the University of Sydney, who fund some of his open-source development efforts. He completed his PhD on event reference there in 2014, and has been teaching their Natural Language Processing unit since 2016.

Stanford CoreNLP: 15 Years of Developing Academic Open Source Software

Christopher Manning
Stanford University

Abstract

My students and I at the Stanford NLP Group started releasing academic open source NLP software relatively early, in 2002. Over the years, the status and popularity of particular tools, and, since 2010, of the integrated Stanford CoreNLP offering has continually grown. It is not only used as a reliable tool — and easy mark to beat — in academic NLP, but it is widely used across government, non-profits, startups, and large companies. In this talk, I give my reflections on building academic open source software: what is required, what is important, and what is not so important; what we did right and what we did wrong; how a software project can be maintained long-term in such a context, how it adds to and detracts from doing academic research, narrowly defined; and how the world has changed and what the prospects are for the future.

Biography

Prof. Christopher Manning is the Thomas M. Siebel Professor in Machine Learning at Stanford University, in the Departments of Computer Science and Linguistics. He works on software that can intelligently process, understand, and generate human language material. He is a leader in applying Deep Learning to Natural Language Processing, with well-known research on the GloVe model of word vectors, Tree Recursive Neural Networks, sentiment analysis, neural network dependency parsing, neural machine translation, and deep language understanding. His computational linguistics work also covers probabilistic models of language, natural language inference and multilingual language processing, including being a principal developer of Stanford Dependencies and Universal Dependencies. Manning has coauthored leading textbooks on statistical approaches to Natural Language Processing (Manning and Schütze 1999) and information retrieval (Manning, Raghavan, and Schütze, 2008), as well as linguistic monographs on ergativity and complex predicates. Manning is an ACM Fellow, a AAAI Fellow, an ACL Fellow, and Past President of the ACL. Research of his has won ACL, Coling, EMNLP, and CHI Best Paper Awards. He has a B.A. (Hons) from The Australian National University, a Ph.D. from Stanford in 1994, and he held faculty positions at Carnegie Mellon University and the University of Sydney before returning to Stanford. He is a member of the Stanford NLP group and manages development of the Stanford CoreNLP software.

Reflections on Running spaCy: Commercial Open-source NLP

Matthew Honnibal and Ines Montani
Explosion AI

Abstract

In this talk, I'll share some lessons we've learned from running spaCy, the fastest-growing library for Natural Language Processing in Python, and provide our perspective on how to make commercial open-source work for both users and developers. Every open-source project must strike a balance between the responsibilities and control of the maintainers, and the responsibilities and control of the users. Understanding and communicating the motivations for publishing software under an open-source license can put less pressure on maintainers, and help users select projects appropriate for their requirements.

Biography

Ines Montani is the lead developer of Prodigy, and a core contributor to spaCy. Although a full-stack developer, Ines has particular expertise in front-end development, having started building websites when she was 11. Before founding Explosion AI, she was a freelance developer and strategist, using her four years executive experience in ad sales and digital marketing.

Matthew Honnibal began his research career as a linguist, completing his PhD in 2009 on lexicalised parsing with Combinatory Categorial Grammar, before working on incremental speech parsing. These days he is best known as a software engineer, for his work on the spaCy NLP library. He grew up in Sydney, lives in Berlin, and still misses CCG.

Table of Contents

Conference Program

Friday, July 20, 2018

8:45–9:00 ***Loading Presentations to Computer***

9:00–9:05 ***Opening Remarks***

9:05–9:50 ***Invited Talk 1 (Joel Nothman)***

9:50–10:30 ***Lightning Presentation for Posters Session 1***

10:30–11:00 ***Coffee Break***

11:00–11:45 ***Poster Session 1***

AllenNLP: A Deep Semantic Natural Language Processing Platform
Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz and Luke Zettlemoyer

Stop Word Lists in Free Open-source Software Packages
Joel Nothman, Hanmin Qin and Roman Yurchak

Texar: A Modularized, Versatile, and Extensible Toolbox for Text Generation
Zhiting Hu, Zichao Yang, Tiancheng Zhao, Haoran Shi, Junxian He, Di Wang, Xuezhe Ma, Zhengzhong Liu, Xiaodan Liang, Lianhui Qin, Devendra Singh Chaplot, Bowen Tan, Xingjiang Yu and Eric Xing

The ACL Anthology: Current State and Future Directions
Daniel Gildea, Min-Yen Kan, Nitin Madnani, Christoph Teichmann and Martin Villalba

The risk of sub-optimal use of Open Source NLP Software: UKB is inadvertently state-of-the-art in knowledge-based WSD
Eneko Agirre, Oier Lopez de Lacalle and Aitor Soroa

12:00–14:00 ***Lunch***

14:00–14:45 ***Invited Talk 2 (Christopher Manning)***

14:45–15:30 ***Lightning Presentation for Posters 2***

15:30–16:00 ***Break***

16:00–16:45 ***Poster Session 2***

Baseline: A Library for Rapid Modeling, Experimentation and Development of Deep Learning Algorithms targeting NLP
Daniel Pressel, Sagnik Ray Choudhury, Brian Lester, Yanjie Zhao and Matt Barta

OpenSeq2Seq: Extensible Toolkit for Distributed and Mixed Precision Training of Sequence-to-Sequence Models
Oleksii Kuchaiev, Boris Ginsburg, Igor Gitman, Vitaly Lavrukhin, Carl Case and Paulius Micikevicius

Integrating Multiple NLP Technologies into an Open-source Platform for Multilingual Media Monitoring
Ulrich Germann, Renars Liepins, Didzis Gosko and Guntis Barzdins

The Annotated Transformer
Alexander Rush

16:45–17:30 ***Invited Talk 3 (Matthew Honnibal and Ines Montani)***

17:30–1735 ***Closing Remarks***