NLP4Prog 2021

The 1st Workshop on Natural Language Processing for Programming (NLP4Prog 2021)

Proceedings of the Workshop

August 6, 2021
Bangkok, Thailand (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 978-1-954085-64-0

Message from the Organizers

Welcome to NLP4Prog, the First Workshop on Natural Language Processing for Programming, co-located with ACL-IJCNLP 2021 online.

The proliferation of programming-related platforms such as GitHub and Stack Overflow has led to large amounts of rich, open-source data consisting of programs associated with natural language, such as natural language questions and answers with code snippets, open-source repositories with natural language comments, and communications between software developers. At the same time, deep learning based techniques have shown promising performance for modeling both natural language and computer programs. Driven by these revolutions on data and models, recent years have witnessed a major resurgence of using NLP techniques to assist programming (NLP4Prog).

As promising as the current developments are, there are still many challenges remaining. This workshop aims to bring related communities (e.g., NLP, Software Engineering, Programming Language, Human-Machine Interaction, Robotics) together to review the recent advances related to NLP4Prog and discuss the remaining challenges and what to expect in the short- and long-term future. While there are similar workshops such as NLP-SEA and NLPaSE held recently, most of them are in conjunction with software engineering venues; to the best of our knowledge, this is the first workshop focusing on NLP for programming and to be held in NLP venues.

A total of 31 papers were submitted and 25 were presented at the workshop. 10 of these papers appear in the proceedings, while the rest were submitted under a non-archival option. In addition, 4 papers from Findings of ACL were also offered presentation slots.

We are thankful to all reviewers for their help in the selection of the program, for their readiness in engaging in thoughtful discussions about individual papers, and for providing valuable feedback to the authors. We would also like to thank the ACL workshop organizers for all the valuable help and support with organizational aspects of the conference. Finally, we would like to thank all our authors and presenters for making this such an exciting event!

NLP4Prog Organizers: Royi Lachmy, Ziyu Yao, Greg Durrett, Milos Gligoric, Junyi Jessy Li, Ray Mooney, Graham Neubig, Yu Su, Huan Sun, Reut Tsarfaty

Organizing Committee

- Royi Lachmy (Bar-Ilan University)
- Ziyu Yao (The Ohio State University)
- Greg Durrett (UT Austin)
- Milos Gligoric (UT Austin)
- Junyi Jessy Li (UT Austin)
- Ray Mooney (UT Austin)
- Graham Neubig (Carnegie Mellon University)
- Yu Su (The Ohio State University/Microsoft Semantic Machines)
- Huan Sun (The Ohio State University)
- Reut Tsarfaty (Bar-Ilan University)

Program Committee

- Miltos Allamanis (MSR, Cambridge)
- Uri Alon (Technion, Israel)
- Jonathan Berant (Tel Aviv University)
- Ben Bogin (Tel Aviv University)
- Saikat Chakraborty (Columbia University)
- Xinyun Chen (UC Berkeley)
- Hanjun Dai (Google Brain)
- Xiang Deng (The Ohio State University)
- Elizabeth Dinella (Univ. of Pennsylvania)
- Li Dong (Microsoft Research Asia)
- Ahmed Elgohary (University of Maryland)
- Xiaodong Gu (HKUST)
- Vincent J. Hellendoorn (CMU)
- Julia Hockenmaier (UIUC)
- Toby Jia-Jun Li (CMU)
- Omer Levy (Tel Aviv University, Israel)
- Victoria Lin (Salesforce)
- Jian-Guang Lou (Microsoft Research Asia)

- Pengyu Nie (UT Austin)
- Sheena Panthaplackel (UT Austin)
- Ice Pasupat (Google AI; Stanford)
- Kyle Richardson (AI2)
- Richard Shin (UC Berkeley)
- Alane Suhr (Cornell)
- Ronen Tamari (Hebrew university)
- Fangzheng (Frank) Xu (CMU)
- Xi Ye (UT Austin)
- Pengcheng Yin (CMU)
- Tao Yu (Yale)
- Rui Zhang (Penn State)
- Ruiqi Zhong (UC Berkeley)

Table of Contents

Conference Program

Friday August 6, 2021

10:05am EDT Session 1

10:05am *Code to Comment Translation: A Comparative Study on Model Effectiveness & Errors*
Junayed Mahmud, Fahim Faisal, Raihan Islam Arnob, Antonios Anastasopoulos and Kevin Moran

10:17am *ConTest: A Unit Test Completion Benchmark featuring Context*
Johannes Villmow, Jonas Depoix and Adrian Ulges

10:30am *CommitBERT: Commit Message Generation Using Pre-Trained Programming Language Model*
Tae Hwan Jung

10:30am *Time-Efficient Code Completion Model for the R Programming Language*
Artem Popov, Dmitrii Orekhov, Denis Litvinov, Nikolay Korolev and Gleb Morgachev

10:30am *CoTexT: Multi-task Learning with Code-Text Transformer*
Long Phan, Hieu Tran, Daniel Le, Hieu Nguyen, James Annibal, Alec Peltekian and Yanfang Ye

10:30am *DIRECT : A Transformer-based Model for Decompiled Identifier Renaming*
Vikram Nitin, Anthony Saieva, Baishakhi Ray and Gail Kaiser

4:15pm EDT Session 2

4:40pm *Shellcode_IA32: A Dataset for Automatic Shellcode Generation*
Pietro Liguori, Erfan Al-Hossami, Domenico Cotroneo, Roberto Natella, Bojan Cukic and Samira Shaikh

4:40pm *Reading StackOverflow Encourages Cheating: Adding Question Text Improves Extractive Code Generation*
Gabriel Orlanski and Alex Gittens

4:40pm *Text-to-SQL in the Wild: A Naturally-Occurring Dataset Based on Stack Exchange Data*
Moshe Hazoom, Vibhor Malik and Ben Bogin

4:40pm *Bag-of-Words Baselines for Semantic Code Search*
Xinyu Zhang, Ji Xin, Andrew Yates and Jimmy Lin