BPPF 2021

The 1st Workshop on Benchmarking: Past, Present and Future

Proceedings of the Workshop

August 5–6, 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 978-1-954085-58-9

Message from the Program Chairs

Where have we been, and where are we going? It is easier to talk about the past than the future. These days, benchmarks evolve more bottom up (such as papers with code). There used to be more top-down leadership from government (and industry, in the case of systems, with benchmarks such as SPEC). Going forward, there may be more top-down leadership from organizations like MLPerf and/or influencers like David Ferrucci, who was responsible for IBM's success with Jeopardy, and has recently written a paper suggesting how the community should think about benchmarking for machine comprehension. Tasks such as reading comprehension become even more interesting as we move beyond English. Multilinguality introduces many challenges, and even more opportunities.

Organizing Committee

Workshop Organizers:

Kenneth Church (Baidu, USA)
Mark Liberman (Penn, USA)
Valia Kordoni (Humboldt, Germany)

Program Committee:

Eduardo Blanco (University of North Texas)
Nicoletta Calzolari (Italy)
Kenneth Church (Baidu, USA)
Christian Federmann (Microsoft Research, USA)
Valia Kordoni (Humboldt, Germany)
Julia Hirshberg (Columbia, USA)
Lori Lamel (LIMSI, France)
Mark Liberman (Penn, USA)
Phillip Koehn (JHU, USA)
Barbara Plank (IT University of Copenhagen, Denmark)
Preslav Nakov (Qatar Computing Research Institute (QCRI), HBKU)
Anette Frank (University of Heidelberg, Germany)
Roy Bar-Haim (IBM Research - Haifa, Israel)

Table of Contents

Benchmarking: Past, Present and Future
Kenneth Church, Mark Liberman and Valia Kordoni 1

Guideline Bias in Wizard-of-Oz Dialogues
Victor Petrén Bach Hansen and Anders Søgaard 8

We Need to Consider Disagreement in Evaluation
Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio and Alexandra Uma 15

How Might We Create Better Benchmarks for Speech Recognition?
Alëna Aksënova, Daan van Esch, James Flynn and Pavel Golik 22

Conference Program

Benchmarking: Past, Present and Future
Kenneth Church, Mark Liberman and Valia Kordoni

Guideline Bias in Wizard-of-Oz Dialogues
Victor Petrén Bach Hansen and Anders Søgaard

We Need to Consider Disagreement in Evaluation
Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio and Alexandra Uma

How Might We Create Better Benchmarks for Speech Recognition?
Alëna Aksënova, Daan van Esch, James Flynn and Pavel Golik