

SustainLP 2020

**SustainLP: Workshop on Simple and Efficient  
Natural Language Processing**

**Proceedings of the Workshop**

November 20, 2020  
Online workshop

©2020 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](mailto:acl@aclweb.org)

ISBN 978-1-952148-77-4

## Introduction

It is our great pleasure to welcome you to the first edition of SustainNLP: Workshop on Simple and Efficient Natural Language Processing.

The Natural Language Processing community has, in recent years, demonstrated a notable focus on improving scores on standard benchmarks and taking the lead on community-wide leaderboards such as (Super)GLUE, SentEval or XTREME. While this led to improvements in benchmark performance of (predominantly neural) models, it also resulted in a worrisome increase in model complexity and the amount of computational resources required for training and using the current state-of-the-art models. Moreover, recent research efforts often fail to identify sources of performance gains in models and to justify model complexity beyond benchmark performance.

Because of these trends as well as the worrisome carbon footprint of (pre)training large neural models, we organized SustainNLP in order to promote simpler and more sustainable NLP research and practices, with two main objectives: (1) encouraging development of more resource-efficient NLP models; and (2) providing simpler architectures and empirical justification of model complexity. For both aspects, we encouraged submissions from all topical areas of NLP.

Besides the original research papers (short and long), we encouraged cross-submissions of work that has been published at other events as well as extended abstracts of work in progress that fit the scope and aims of the workshop (only the original research papers, however, are included in these workshop proceedings).

This first edition of the workshop also included a *shared task* encouraging an optimal trade-off between the model performance and efficiency during inference. The shared task focused on inference efficiency as it can be difficult to fairly evaluate training efficiency in the most general setting. Moreover, as large-scale pretrained models reach production, it is the computational cost of inference that will account for most of the cumulative lifetime environmental cost of these models.

We received overwhelming 48 submissions (and 1 shared task system description), proposing a multitude of viable resource-efficient NLP methods and spanning a wide range of NLP applications. We have selected 28 submissions for presentation at the workshop, yielding an acceptance rate of 58%). Additionally, the workshop will include presentations of 38 papers accepted for publication in EMNLP Findings, the content of which we judged to be in line with the scope and aims of the workshop.

Many thanks to our program committee for their thorough and thoughtful reviews. We would also like to thank to our panelists and invited speakers whose discussions and talks we strongly believe will make the workshop exciting and memorable.

We are looking forward to the first edition of the SustainNLP workshop!

SustainNLP Organizers

October 2020



**Organizers:**

Nafise Sadat Moosavi, TU Darmstadt  
Angela Fan, INRIA Nancy & Facebook AI Research Paris  
Goran Glavaš, University of Mannheim  
Vered Shwartz, Allen Institute for AI (AI2)  
Shafiq Joty, Nanyang Technological University  
Alex Wang, New York University  
Thomas Wolf, Huggingface Inc.

**Steering Committee:**

Sam Bowman, New York University  
Mona Diab, George Washington University  
Andrew McCallum, University of Massachusetts Amherst  
Alexander Rush, Cornell University  
Luke Zettlemoyer, University of Washington & Facebook

**Program Committee:**

Eneko Agirre, Mikel Artetxe, Dennis Aumiller, Guy Boudoukh, Samuel Cahyawijaya, Elizabeth Clark, Alexis Conneau, Rotem Dror, Gerard Dupont, Nouha Dziri, Kiril Gashteovski, Sebastian Gehrmann, Marjan Ghazvininejad, Samujjwal Ghosh, Andreas Hanselowski, Benjamin Heinzlerling, Ari Holtzman, Ozan Irsoy, Srinivasan Iyer, Peter Izsak, Mandar Joshi, Ehsan Kamaloo, Gyuwan Kim, Young Jin Kim, Guillaume Lample, Phong Le, Ji-Ung Lee, Louis Martin, Seyed Abolghasem Mirroshandel, Marius Mosbach, Myle Ott, Daraksha Parveen, Matthew Peters, Jonas Pfeiffer, Mohammad Taher Pilehvar, Barbara Plank, Simone Paolo Ponzetto, Ofir Press, Hannah Rashkin, Siva Reddy, Ines Rehbein, Andreas Rücklé, Victor Sanh, Roy Schwartz, Edwin Simpson, Gabriel Stanovsky, Anders Søgaard, Urmish Thakker, Prasetya Ajie Utama, Ivan Vulić, Moshe Wasserblat, Genta Indra Winata, Sam Wiseman, Caiming Xiong, Canwen Xu

**Invited Speakers & Panelists:**

Mona Diab, George Washington University  
Heng Ji, University of Illinois at Urbana-Champaign  
Graham Neubig, Carnegie Mellon University  
Alexander Rush, Cornell University  
Emma Strubell, Carnegie Mellon University  
Armand Joulin, Facebook AI Research  
Kyunghyun Cho, New York University  
Yejin Choi, University of Washington & Allen Institute for Artificial Intelligence (AI2)  
Yoav Goldberg, Bar Ilan University  
Iryna Gurevych, TU Darmstadt



## Table of Contents

<i>Knowing Right from Wrong: Should We Use More Complex Models for Automatic Short-Answer Scoring in Bahasa Indonesia?</i>	
Ali Akbar Septiandri, Yosef Ardhito Winatmoko and Ilham Firdausi Putra . . . . .	1
<i>Rank and run-time aware compression of NLP Applications</i>	
Urmish Thakker, Jesse Beu, Dibakar Gope, Ganesh Dasika and Matthew Mattina . . . . .	8
<i>Learning Informative Representations of Biomedical Relations with Latent Variable Models</i>	
Harshil Shah and Julien Fauqueur . . . . .	19
<i>End to End Binarized Neural Networks for Text Classification</i>	
Kumar Shridhar, Harshil Jain, Akshat Agarwal and Denis Kleyko . . . . .	29
<i>Exploring the Boundaries of Low-Resource BERT Distillation</i>	
Moshe Wasserblat, Oren Pereg and Peter Izsak . . . . .	35
<i>Efficient Estimation of Influence of a Training Instance</i>	
Sosuke Kobayashi, Sho Yokoi, Jun Suzuki and Kentaro Inui . . . . .	41
<i>Efficient Inference For Neural Machine Translation</i>	
Yi-Te Hsu, Sarthak Garg, Yi-Hsiu Liao and Ilya Chatsviorkin . . . . .	48
<i>Sparse Optimization for Unsupervised Extractive Summarization of Long Documents with the Frank-Wolfe Algorithm</i>	
Alicia Tsai and Laurent El Ghaoui . . . . .	54
<i>Don't Read Too Much Into It: Adaptive Computation for Open-Domain Question Answering</i>	
Yuxiang Wu, Pasquale Minervini, Pontus Stenetorp and Sebastian Riedel . . . . .	63
<i>A Two-stage Model for Slot Filling in Low-resource Settings: Domain-agnostic Non-slot Reduction and Pretrained Contextual Embeddings</i>	
Cennet Oguz and Ngoc Thang Vu . . . . .	73
<i>Early Exiting BERT for Efficient Document Ranking</i>	
Ji Xin, Rodrigo Nogueira, Yaoliang Yu and Jimmy Lin . . . . .	83
<i>Keyphrase Generation with GANs in Low-Resources Scenarios</i>	
Giuseppe Lancioni, Saida S.Mohamed, Beatrice Portelli, Giuseppe Serra and Carlo Tasso . . . . .	89
<i>Quasi-Multitask Learning: an Efficient Surrogate for Obtaining Model Ensembles</i>	
Norbert Kis-Szabó and Gábor Berend . . . . .	97
<i>A Little Bit Is Worse Than None: Ranking with Limited Training Data</i>	
Xinyu Zhang, Andrew Yates and Jimmy Lin . . . . .	107
<i>Predictive Model Selection for Transfer Learning in Sequence Labeling Tasks</i>	
Parul Awasthy, Bishwaranjan Bhattacharjee, John Kender and Radu Florian . . . . .	113
<i>Load What You Need: Smaller Versions of Multilingual BERT</i>	
Amine Abdaoui, Camille Pradel and Grégoire Sigel . . . . .	119
<i>SqueezeBERT: What can computer vision teach NLP about efficient neural networks?</i>	
Forrest Iandola, Albert Shaw, Ravi Krishna and Kurt Keutzer . . . . .	124

<i>Analysis of Resource-efficient Predictive Models for Natural Language Processing</i> Raj Pranesh and Ambesh Shekhar .....	136
<i>Towards Accurate and Reliable Energy Measurement of NLP Models</i> Qingqing Cao, Aruna Balasubramanian and Niranjan Balasubramanian .....	141
<i>FastFormers: Highly Efficient Transformer Models for Natural Language Understanding</i> Young Jin Kim and Hany Hassan .....	149
<i>A comparison between CNNs and WFAs for Sequence Classification</i> Ariadna Quattoni and Xavier Carreras .....	159
<i>Counterfactual Augmentation for Training Next Response Selection</i> Seungtaek Choi, Myeongho Jeong, Jinyoung Yeo and Seung-won Hwang .....	164
<i>Do We Need to Create Big Datasets to Learn a Task?</i> Swaroop Mishra and Bhavdeep Singh Sachdeva .....	169
<i>Overview of the SustainLP 2020 Shared Task</i> Alex Wang and Thomas Wolf .....	174



## Workshop Program

*Knowing Right from Wrong: Should We Use More Complex Models for Automatic Short-Answer Scoring in Bahasa Indonesia?*

Ali Akbar Septiandri, Yosef Ardhito Winatmoko and Ilham Firdausi Putra

*Rank and run-time aware compression of NLP Applications*

Urmish Thakker, Jesse Beu, Dibakar Gope, Ganesh Dasika and Matthew Mattina

*Incremental Neural Coreference Resolution in Constant Memory*

Patrick Xia, João Sedoc and Benjamin Van Durme

*Learning Informative Representations of Biomedical Relations with Latent Variable Models*

Harshil Shah and Julien Fauqueur

*End to End Binarized Neural Networks for Text Classification*

Kumar Shridhar, Harshil Jain, Akshat Agarwal and Denis Kleyko

*Large Product Key Memory for Pre-trained Language Models*

Gyuwan Kim and Tae Hwan Jung

*P-SIF: Document Embeddings using Partition Averaging*

Vivek Gupta, Ankit Saw, Pegah Nokhiz, Praneeth Netrapalli, Piyush Rai and Partha Talukdar

*Exploring the Boundaries of Low-Resource BERT Distillation*

Moshe Wasserblat, Oren Pereg and Peter Izsak

*Efficient Estimation of Influence of a Training Instance*

Sosuke Kobayashi, Sho Yokoi, Jun Suzuki and Kentaro Inui

*Efficient Inference For Neural Machine Translation*

Yi-Te Hsu, Sarthak Garg, Yi-Hsiu Liao and Ilya Chatsviorkin

*Sparse Optimization for Unsupervised Extractive Summarization of Long Documents with the Frank-Wolfe Algorithm*

Alicia Tsai and Laurent El Ghaoui

*Don't Read Too Much Into It: Adaptive Computation for Open-Domain Question Answering*

Yuxiang Wu, Pasquale Minervini, Pontus Stenetorp and Sebastian Riedel

November 20, 2020 (continued)

*Doped Structured Matrices for Extreme Compression of LSTM Models*  
Urmish Thakker, Paul Whatmough, Zhi-Gang Liu, Matthew Mattina and Jesse Beu

*A Two-stage Model for Slot Filling in Low-resource Settings: Domain-agnostic Non-slot Reduction and Pretrained Contextual Embeddings*  
Cennet Oguz and Ngoc Thang Vu

*Early Exiting BERT for Efficient Document Ranking*  
Ji Xin, Rodrigo Nogueira, Yaoliang Yu and Jimmy Lin

*Keyphrase Generation with GANs in Low-Resources Scenarios*  
Giuseppe Lancioni, Saida S.Mohamed, Beatrice Portelli, Giuseppe Serra and Carlo Tasso

*Quasi-Multitask Learning: an Efficient Surrogate for Obtaining Model Ensembles*  
Norbert Kis-Szabó and Gábor Berend

*A Little Bit Is Worse Than None: Ranking with Limited Training Data*  
Xinyu Zhang, Andrew Yates and Jimmy Lin

*Predictive Model Selection for Transfer Learning in Sequence Labeling Tasks*  
Parul Awasthy, Bishwaranjan Bhattacharjee, John Kender and Radu Florian

*Load What You Need: Smaller Versions of Multilingual BERT*  
Amine Abdaoui, Camille Pradel and Grégoire Sigel

*SqueezeBERT: What can computer vision teach NLP about efficient neural networks?*  
Forrest Iandola, Albert Shaw, Ravi Krishna and Kurt Keutzer

*Analysis of Resource-efficient Predictive Models for Natural Language Processing*  
Raj Pranesh and Ambesh Shekhar

*Towards Accurate and Reliable Energy Measurement of NLP Models*  
Qingqing Cao, Aruna Balasubramanian and Niranjana Balasubramanian

*FastFormers: Highly Efficient Transformer Models for Natural Language Understanding*  
Young Jin Kim and Hany Hassan

November 20, 2020 (continued)

*A comparison between CNNs and WFAs for Sequence Classification*

Ariadna Quattoni and Xavier Carreras

*Counterfactual Augmentation for Training Next Response Selection*

Seungtaek Choi, Myeongho Jeong, Jinyoung Yeo and Seung-won Hwang

*Do We Need to Create Big Datasets to Learn a Task?*

Swaroop Mishra and Bhavdeep Singh Sachdeva

*Guiding Attention for Self-Supervised Learning with Transformers*

Ameet Deshpande and Karthik Narasimhan

Besides the papers listed above, 38 **EMNLP Findings** papers will be presented at SustainNLP as well.

