

Efficient and Equitable Natural Language Processing in the Age of Deep Learning

Jesse Dodge^{*1}, Iryna Gurevych^{*2}, Roy Schwartz^{*3}, Emma Strubell^{*4},
and Betty van Aken^{†5}

- 1 AI2 – Seattle, US. jessed@allenai.org
- 2 TU Darmstadt, DE. gurevych@cs.tu-darmstadt.de
- 3 The Hebrew University of Jerusalem, IL. roy.schwartz1@mail.huji.ac.il
- 4 Carnegie Mellon University – Pittsburgh, US. strubell@cmu.edu
- 5 (Berliner Hochschule für Technik, DE. Betty.vanAken@bht-berlin.de)

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 22232 “Efficient and Equitable Natural Language Processing in the Age of Deep Learning”. Since 2012, the field of artificial intelligence (AI) has reported remarkable progress on a broad range of capabilities including object recognition, game playing, speech recognition, and machine translation. Much of this progress has been achieved by increasingly large and computationally intensive deep learning models: training costs for state-of-the-art deep learning models have increased 300,000 times between 2012 and 2018 [1]. Perhaps the epitome of this trend is the subfield of natural language processing (NLP) that over the past three years has experienced even sharper growth in model size and corresponding computational requirements in the word embedding approaches (e.g. ELMo, BERT, openGPT-2, Megatron-LM, T5, and GPT-3, one of the largest models ever trained with 175B dense parameters) that are now the basic building blocks of nearly all NLP models. Recent studies indicate that this trend is both environmentally unfriendly and prohibitively expensive, raising barriers to participation in NLP research [2, 3]. The goal of this seminar was to mitigate these concerns and promote equity of access in NLP.

References

- 1 D. Amodei and D. Hernandez. 2018. AI and Compute. <https://openai.com/blog/ai-and-compute>
- 2 R. Schwartz, D. Dodge, N. A. Smith, and O. Etzioni. 2020. Green AI. Communications of the ACM (CACM)
- 3 E. Strubell, A. Ganesh, and A. McCallum. 2019. Energy and Policy Considerations for Deep Learning in NLP. In Proc. of ACL.

Seminar June 6–10, 2022 – <http://www.dagstuhl.de/22232>

2012 ACM Subject Classification Computing methodologies → Natural language processing; Computing methodologies → Neural networks; Social and professional topics → Sustainability

Keywords and phrases deep learning, efficiency, equity, natural language processing (nlp)

Digital Object Identifier 10.4230/DagRep.12.6.14

* Editor / Organizer

† Editorial Assistant / Collector



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 4.0 International license

Efficient and Equitable Natural Language Processing in the Age of Deep Learning, *Dagstuhl Reports*, Vol. 12, Issue 6, pp. 14–27

Editors: Jesse Dodge, Iryna Gurevych, Roy Schwartz, and Emma Strubell



Dagstuhl Reports

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany


1 Executive Summary

Roy Schwartz (The Hebrew University of Jerusalem, IL)

Jesse Dodge (AI2 – Seattle, US)

Iryna Gurevych (TU Darmstadt, DE)

Emma Strubell (Carnegie Mellon University – Pittsburgh, US)

License  Creative Commons BY 4.0 International license
© Roy Schwartz, Jesse Dodge, Iryna Gurevych, and Emma Strubell

For this seminar, we brought together a diverse group of researchers and practitioners in NLP and adjacent fields to develop actionable policies, incentives and a joint strategy towards more efficient and equitable NLP. This Dagstuhl Seminar covered a range of related topics, which we summarize as follows.

Efficient NLP models

A key method for mitigating the raised concerns is reducing costs by making models more efficient. We surveyed the different methods that exist for making NLP technology more efficient. We discussed their tradeoffs, prioritized them, and aimed to identify new opportunities to promote efficiency in NLP. During the seminar, we drafted a survey paper summarizing multiple methods for increasing the efficiency of NLP models. We aim to publish this work later this year.

Systemic issues

We also addressed systemic issues in the field relating to the reporting of computational budgets in NLP research, and how we can use incentive structures such as the NLP Reproducibility Checklist [1] to motivate researchers throughout the field to improve reporting. We discussed the survey responses for the reproducibility checklist used at four major NLP conferences, and we plan to release a report of this data.

Equity of access

A third topic of discussion was the equity of access to computational resources and state-of-the-art NLP technologies. Prior to the seminar, we conducted a survey of different stakeholders across the NLP community. During the seminar, we analyzed and discussed the results of this survey to better understand who is most affected and how, and developed informed strategies and policies to mitigate this inequity moving forward. We are currently working on a paper summarizing the results of this survey, which we hope to publish later this year.

Measuring efficiency and equity

All of the above endeavors require establishing the right metrics and standards to measure our current status and progress towards efficiency and equity goals. We discussed multiple metrics and evaluation frameworks that capture the bigger picture of how different approaches compare in terms of energy efficiency not just in the research environment but in practice and over the entire ML model lifecycle (development, training and deployment), and that work under a wide range of computational budgets.

References

- 1 Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, Noah A. Smith: Show Your Work: Improved Reporting of Experimental Results. EMNLP/IJCNLP (1) 2019: 2185-2194

2 Table of Contents

Executive Summary

Roy Schwartz, Jesse Dodge, Iryna Gurevych, and Emma Strubell 15

Overview of Talks

Forays into Efficiency and Energy of NLP Models
Niranjan Balasubramanian 18

Faster Neural Network Training, Algorithmically
Jonathan Frankle 18

Evaluating Approximations is Hard; Efficient Machine Translation Shared Task
Kenneth Heafield 18

ML Efficiency: Open Challenges and Opportunities.
Sara Hooker 19

Neurosymbolic models in semantic parsing
Alexander Koller 19

Investigating Rational Activation Functions to Train Transformer Models
Ji-Ung Lee 20

Holistic model evaluation
Alexandra Sasha Luccioni 20

Deep Patient Representation
Alexander Löser 20

Is Sparsity a Path for Efficiency?
André F. T. Martins 21

The Sweet Lesson
Colin Raffel 21

On #Reviewer2 and paper-reviewer assignments
Anna Rogers 21

BigScience Large LMs and small labs
Thomas Wolf 22

Working groups

Efficiency benchmarking
Niranjan Balasubramanian, Leon Derczynski, Jesse Dodge, Jonathan Frankle, Iryna Gurevych, Kenneth Heafield, Sara Hooker, André F. T. Martins, Haritz Puerto, Colin Raffel, Roy Schwartz, Edwin Simpson, Noam Slonim, and Thomas Wolf . . . 22

Implementing changes in NLP research
Jesse Dodge, Niranjan Balasubramanian, Jessica Forde, Kenneth Heafield, Alexander Koller, Ji-Ung Lee, André F. T. Martins, Nils Reimers, Leonardo Ribeiro, Andreas Rücklé, and Betty van Aken 23

Breakout on “Making Change”
Roy Schwartz, Leon Derczynski, Jonathan Frankle, Iryna Gurevych, Alexander Koller, Alexandra Sasha Luccioni, André F. T. Martins, Colin Raffel, Anna Rogers, Noam Slonim, Emma Strubell, and Thomas Wolf 23

Panel discussions

Panel on Equity in NLP research


Colin Raffel, Iryna Gurevych, Alexandra Sasha Luccioni, Noah A. Smith, Emma Strubell, and Thomas Wolf 26

Participants 27

3 Overview of Talks

3.1 Forays into Efficiency and Energy of NLP Models


Niranjan Balasubramanian (Stony Brook University, US)

License  Creative Commons BY 4.0 International license
© Niranjan Balasubramanian

This talk presents forays into efficient QA models, modeling sparsity for hardware acceleration, and issues in measuring energy consumption.

3.2 Faster Neural Network Training, Algorithmically

Jonathan Frackle (Harvard University – Allston, US)

License  Creative Commons BY 4.0 International license
© Jonathan Frackle

Training modern neural networks is time-consuming, expensive, and energy-intensive. As neural network architectures double in size every few months, it is difficult for researchers and businesses without immense budgets to keep up. In this talk, I describe one approach for managing this challenge: changing the training algorithm itself. While many companies and researchers are focused on building hardware and systems to allow existing algorithms to run faster in a mathematically equivalent fashion, there is nothing sacred about this math. To the contrary, training neural networks is inherently approximate, relying on noisy data, convex optimizers in nonconvex regimes, and ad hoc tricks and hacks that seem to work well in practice for reasons that elude us.

I discuss how we have put this approach into practice at MosaicML, including the dozens of algorithmic changes we have studied (which are freely available open source), the science behind how these changes interact with each other (the composition problem), and how we evaluate whether these changes have been effective. I will also detail several surprises we have encountered and lessons we have learned along the way. In the four months since we began this work in earnest, we have reduced the training times of standard computer vision models by 7x and standard language models by 2x on publicly available cloud instances, and we believe we are just scratching the surface.

3.3 Evaluating Approximations is Hard; Efficient Machine Translation Shared Task

Kenneth Heafield (University of Edinburgh, GB)

License  Creative Commons BY 4.0 International license
© Kenneth Heafield

Papers about a new approximation (i.e. faster for some loss in quality) often claim the quality loss is small, while better papers perform a Pareto comparison. Unfortunately, the baseline approximations used for the Pareto comparison are usually restricted to the same type of method, such as pruning. I argue the correct baseline is all approximations that already exist. Approximations are stackable, so the question is really whether the proposed

approximation belongs to a set of stacked approximations that advance the Pareto frontier. This is a high standard and difficult for the average paper to reach, so I present a partial solution. The efficient machine translation shared task establishes the state-of-the-art by soliciting competitive submissions and comparing them. Starting from a range of already efficient systems provides a much stronger baseline for evaluating a new approximation.

3.4 ML Efficiency: Open Challenges and Opportunities.

Sara Hooker (Google – Mountain View, US)

License  Creative Commons BY 4.0 International license
© Sara Hooker

Our field is currently characterized by a “bigger is better” trend in the size of deep neural networks. This talk posits that this is an unsustainable recipe – akin to building a ladder to the moon. We discuss some important directions for revisiting the efficiency of our representation learning approaches.

3.5 Neurosymbolic models in semantic parsing

Alexander Koller (Universität des Saarlandes, DE)


License  Creative Commons BY 4.0 International license
© Alexander Koller

There are many approaches to mapping natural-language sentences to symbolic meaning representations. The current dominant approach is with neural sequence-to-sequence models, which map the sentence to a string version of the meaning representation. Seq2seq models work well for many NLP tasks, including tagging and parsing, and deliver excellent accuracy on broad-coverage semantic parsing as well. However, it has recently been found that seq2seq models struggle with “compositional generalization”: They have a hard time generalizing from training examples to structurally similar unseen test sentences. I will show some new results that pinpoint this difficult more precisely, and discuss what this means for how to best evaluate semantic parsers.

I will then present our own research on compositional semantic parsing, which combines neural models with the Principle of Compositionality from theoretical semantics. Our semantic parser uses a neural supertagger to predict word meanings and a neural dependency parser to predict the compositional structure, and then evaluates this dependency structure in a graph algebra to obtain the meaning representation. We achieve state-of-the-art parsing accuracy across a number of graphbanks, at a speed of up to 10k tokens/second.

3.6 Investigating Rational Activation Functions to Train Transformer Models

Ji-Ung Lee (TU Darmstadt, DE)

License  Creative Commons BY 4.0 International license

© Ji-Ung Lee

Joint work of Haishuo Fang, Ji-Ung Lee, Nafise Sadat Moosavi, Iryna Gurevych

In this work, we explore rational activation functions for training transformer models. In contrast to activation functions such as GELU which remain fixed after initialization, rational activation functions are capable of approximating any arbitrary activation function during training. In preliminary experiments we find that using rational activation functions can lead to a faster convergence during pre-training as well as a higher performance on several downstream tasks.

3.7 Holistic model evaluation

Alexandra Sasha Luccioni (Hugging Face – Paris, FR)

License  Creative Commons BY 4.0 International license

© Alexandra Sasha Luccioni

In both research and industry, there are multiple factors to consider when comparing models. Our current ML benchmarks measure one aspect of this, e.g. NLI, NER, QA. How do we integrate different aspects of model performance when comparing models?

3.8 Deep Patient Representation

Alexander Löser (Berliner Hochschule für Technik, DE)

License  Creative Commons BY 4.0 International license

© Alexander Löser

Joint work of Betty van Aken, Jens-Michalis Papaioannou, Manuel Mayrdorfer, Klemens Budde, Felix A. Gers, Alexander Löser

Main reference Betty van Aken, Jens-Michalis Papaioannou, Manuel Mayrdorfer, Klemens Budde, Felix A. Gers, Alexander Löser: “Clinical Outcome Prediction from Admission Notes using Self-Supervised Knowledge Integration”, in Proc. of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 – 23, 2021, pp. 881–893, Association for Computational Linguistics, 2021.

URL <https://doi.org/10.18653/v1/2021.eacl-main.75>

Understanding clinical outcomes requires to integrate different modalities in a single latent representation. We present such operators that reuse clinical large language models in English language and integrate complementary medical latent representation from low resource languages, from ontologies, from time variant data and from set data. An example application is the differential diagnosis at <https://outcome-prediction.demo.dataxis.com>.

3.9 Is Sparsity a Path for Efficiency?

André F. T. Martins (IST – Lisbon, PT)

License © Creative Commons BY 4.0 International license
© André F. T. Martins

Current NLP models are increasingly larger and data-hungry, which poses important environmental challenges. In this talk, I discuss several ways in which sparsity might lead to more efficient NLP models. The current life cycle of NLP models offers several opportunities to improve memory and runtime efficiency at different stages: during pretraining, during finetuning, and during inference. I first distinguish between model sparsity and activation sparsity. Then, I focus on adaptive sparse attention approaches for the latter, where the softmax transformation is replaced by sparse transformations – entmax – which maintain end-to-end differentiability and have a learnable parameter which controls their sparsity. I finish by asking several open questions and inviting discussion.

3.10 The Sweet Lesson

Colin Raffel (University of North Carolina at Chapel Hill, US)

License © Creative Commons BY 4.0 International license
© Colin Raffel

Richard Sutton’s essay “The Bitter Lesson” argues that “general methods that leverage computation are ultimately the most effective”. In this talk, I will argue that the bitter lesson implies that, at a given point in time, it is often possible to outperform large-scale methods with methods that are more efficient and clever. Furthermore, actively working to develop more efficient methods has often uncovered new approaches that scale better. I call this perspective “the sweet lesson” and will present many examples of this principle. Finally, I will wrap up with some thoughts on how to internalize bitter and sweet lessons in NLP’s current era of scale.

3.11 On #Reviewer2 and paper-reviewer assignments

Anna Rogers (University of Copenhagen, DK)

License © Creative Commons BY 4.0 International license
© Anna Rogers

Joint work of Terne Thorn Jakobsen, Anna Rogers

Main reference Terne Thorn Jakobsen, Anna Rogers: “What Factors Should Paper-Reviewer Assignments Rely On? Community Perspectives on Issues and Ideals in Conference Peer-Review”, in Proc. of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4810–4823, Association for Computational Linguistics, 2022.

URL <https://doi.org/10.18653/v1/2022.naacl-main.354>

Some thoughts and new data on community preferences for how papers should be matched to reviewers.

3.12 BigScience Large LMs and small labs

Thomas Wolf (Hugging Face – Paris, FR)


License  Creative Commons BY 4.0 International license
© Thomas Wolf

In this talk, I'll be presenting the BigScience (<https://bigscience.huggingface.co>) project. A collaborative experiment in building a multilingual large scale dataset as well as a multilingual large language model, inspired by other fields of research like the Large Hadron Collider.

4 Working groups

4.1 Efficiency benchmarking

Niranjan Balasubramanian (Stony Brook University, US), Leon Derczynski (IT University of Copenhagen, DK), Jesse Dodge (AI2 – Seattle, US), Jonathan Frankle (Harvard University – Allston, US), Iryna Gurevych (TU Darmstadt, DE), Kenneth Heafield (University of Edinburgh, GB), Sara Hooker (Google – Mountain View, US), André F. T. Martins (IST – Lisbon, PT), Haritz Puerto (TU Darmstadt, DE), Colin Raffel (University of North Carolina at Chapel Hill, US), Roy Schwartz (The Hebrew University of Jerusalem, IL), Edwin Simpson (University of Bristol, GB), Noam Slonim (IBM – Haifa, IL), and Thomas Wolf (Hugging Face – Paris, FR)

License  Creative Commons BY 4.0 International license
© Niranjan Balasubramanian, Leon Derczynski, Jesse Dodge, Jonathan Frankle, Iryna Gurevych, Kenneth Heafield, Sara Hooker, André F. T. Martins, Haritz Puerto, Colin Raffel, Roy Schwartz, Edwin Simpson, Noam Slonim, and Thomas Wolf

This breakout session concerned questions about what we should measure and report for NLP experiments and how efficiency can be measured. A common problem with current practice in NLP is that efficiency is either not reported at all or that the used metrics are hard to compare. Different hardware environments often further require individual solutions. Shared tasks and benchmarks with fixed hardware were identified as one attempt to mitigate the problem of comparability. The MLPerf benchmarks [1] were mentioned as one positive example. However, participants raised the question whether such benchmarks lead to overfitting and distract from real life concerns. Also, different tasks require different constraints, e.g. the pre-training of a large language model entails different concerns than inferencing on this model. Use cases should therefore be viewed from different perspectives. The group agreed that pushing people to report and review efficiency measures can result in a culture shift and an acceleration of science in general.

References

- 1 Farrell, Steven, Murali Emani, Jacob Balma, Lukas Drescher, Aleksandr Drozd, Andreas Fink, Geoffrey Fox et al. “MLPerf™ HPC: A Holistic Benchmark Suite for Scientific Machine Learning on HPC Systems.” In 2021 IEEE/ACM Workshop on Machine Learning in High Performance Computing Environments (MLHPC), pp. 33-45. IEEE, 2021.

4.2 Implementing changes in NLP research

Jesse Dodge (AI2 – Seattle, US), Niranjan Balasubramanian (Stony Brook University, US), Jessica Forde (Brown University – Providence, US), Kenneth Heafield (University of Edinburgh, GB), Alexander Koller (Universität des Saarlandes, DE), Ji-Ung Lee (TU Darmstadt, DE), André F. T. Martins (IST – Lisbon, PT), Nils Reimers (Hugging Face – Paris), Leonardo Ribeiro (TU Darmstadt, DE), Andreas Rücklé (Amazon – Berlin, DE), and Betty van Aken (Berliner Hochschule für Technik, DE)

License © Creative Commons BY 4.0 International license

© Jesse Dodge, Niranjan Balasubramanian, Jessica Forde, Kenneth Heafield, Alexander Koller, Ji-Ung Lee, André F. T. Martins, Nils Reimers, Leonardo Ribeiro, Andreas Rücklé, and Betty van Aken

In this breakout session, the group discussed how to *implement* the ideas for more efficient and equitable NLP research within the community. A step toward that goal is the Checklist for Responsible NLP Research recently introduced in the ACL Rolling Review Submissions. The list contains questions, e.g. whether a submission discusses the risks of a work or mentions the computational budget of a solution. Jesse Dodge presented statistics from the first rounds of reviews using the checklist. The results showed that the acceptance rate was positively correlated with the number of items checked on the list, indicating that the initiative was successful in stimulating higher-quality research. Participants also agreed that templates for sections that mention limitations or reproducibility of experiments are a useful tool for early-stage researchers. The common view within the group was that current academic structures often encourage quantity over quality of publications, especially for junior researchers. Changing this culture and the incentives for doing research was identified as necessary for more carefully thought out and reproducible research.

4.3 Breakout on “Making Change”

Roy Schwartz (The Hebrew University of Jerusalem, IL), Leon Derczynski (IT University of Copenhagen, DK), Jonathan Frankle (Harvard University – Allston, US), Iryna Gurevych (TU Darmstadt, DE), Alexander Koller (Universität des Saarlandes, DE), Alexandra Sasha Luccioni (Hugging Face – Paris, FR), André F. T. Martins (IST – Lisbon, PT), Colin Raffel (University of North Carolina at Chapel Hill, US), Anna Rogers (University of Copenhagen, DK), Noam Slonim (IBM – Haifa, IL), Emma Strubell (Carnegie Mellon University – Pittsburgh, US), and Thomas Wolf (Hugging Face – Paris, FR)

License © Creative Commons BY 4.0 International license

© Roy Schwartz, Leon Derczynski, Jonathan Frankle, Iryna Gurevych, Alexander Koller, Alexandra Sasha Luccioni, André F. T. Martins, Colin Raffel, Anna Rogers, Noam Slonim, Emma Strubell, and Thomas Wolf

- Roy: Wrote a policy document to be adopted by the ACL exec
 - Iryna came to talk at Roy’s lab, mentioned she was part of ACL exec suggested going through ACL to try to influence things
 - Brought on Emma and Jesse and Andreas and had weekly meetings to think about what to do
 - Iryna knew next steps and how to promote – maybe not a general recipe since it relied on her expertise and position

- Wrote a document and took time to get it right, including feedback at the ACL business meetings, assembled an advisory panel for feedback too (including people you disagreed with)
- What were the three recommendations?
 - * Add instructions to reviewers and authors and question for review form
 - * Add efficiency track permanently to all future conferences
 - * Encourage submission of code and data using the badge system
- ACL exec ultimately decides whether they will adopt it, and then it is technically a new set of recommendations for all of the rest of the conferences
- There is no set of centralized rules – PCs of individual conferences can ultimately choose to ignore the policy documents. May therefore need to also talk to individual PCs to get the changes implemented.
- There is also a conference handbook – also hoping to have the recommendations put in the conference handbook. The person responsible for it is part of the ACL exec – need to work with the person to say “this is the paragraph that needs to be included here, this is the paragraph that needs to be included there”.
- How do PCs get elected? They are invited by the exec.
- How does the ACL exec get elected? Candidates are nominated/selected and then the community votes.
- Anna: Since PCs rotate, there may be no continuity. Therefore things that stick are ones that are perceived as being a positive change.
- Jonathan: Given this, what are the most “durable” changes? The efficiency track?
- Alexander: Things are more durable thanks to ARR, which is controlled by the exec rather than the individual PCs
- Noam: Meta-point – make OKRs?
 - What are the objectives and key results required for that?
 - Need to figure out how we are going to measure whether we were successful.
 - André: Agree, seems necessary to separate the what from the how.
 - Leon: Super hard to truly do in a super exact way – “does the change in reviewing have an impact” – ultimately you have subjective judgements, you can’t really measure a lot of these changes.
 - Alexander: Certain things are easier to measure – e.g. are people releasing more code?
 - Emma: Survey every year to ask things like “are things getting better?”
 - Are other communities dealing with this? E.g. quantum computing.
- Jonathan: Split off?
 - Why are you fighting to change the communities you have rather than split off and create a new subcommunity?
 - Good example: FACCT.
 - Emma: Strongly disagree – sort of like checking out, would rather change the community rather than make a new community that cares about different issues. Worry that FACCT makes those issues not first-priority issues in the ML community.
 - Jonathan: FACCT is changing the machine learning community, through influence. For example, with mlsys. Arguably it’s even more impactful than having them be lost in the shuffle at NeurIPS.
 - Jonathan: Not creating a new community – the community exists, really about how to get the message out there.
 - Thom: Keep them connected so that there’s cultural exchange.

- Alexander: Meta-point – levels of change
 - There are multiple levels at which change can be caused – e.g. individual (blog posts, tweets, whatever) and top-down (e.g. policy doc) and community building (there is a community that agrees with us)
 - Thom: Build tools – easy to use, best thing to use – that causes change.
 - Anna: Importance of the human interface – what about putting “efficiency” badges on the HF hub?
- Leon: How much of the community are you reaching/representing?
- Thom: Peer review?
 - Less worried about efficiency, seems like a lot of people are interested in it.
 - More worried about carbon emissions especially in contrast to the desire for more GPUs.
 - Worried about the reviewing process because it can mean big groups leave the reviewing process.
 - Alexander: Super worrying development that people are circumventing the review process through arxiv and press releases.
 - What about connecting other communities? E.g. EleutherAI, can we do a way to connect the communities.
 - Anna: I want anonymous pre-prints.
 - Jonathan: Can we publicly peer review the non-peer reviewed papers?
 - André: Shouldn't assume peer review is the correct thing.
 - André: Don't make peer review too complicated – e.g. don't use super-structured review forms.
 - Emma: Structured review forms are a super effective tool for effecting change. TACL “Single box and you write what you want” vs. ARR prompting about the science specifically. Prompting is important, not complexity.
 - Colin: Any examples of successful public critiques?
 - Leon: Yes, about a stock market paper.
 - Roy: Yes, Yoav about a language generation paper.
 - André: Blog posts don't solve the problem because in the end it still depends on influential people.
 - Anna: Sort of like fake news and fact checking. E.g. post a negative results/can't reproduce paper.
 - Sasha: Use openreview.
- Roy: How do we create a community towards these initiatives/keep the community alive?
 - Jonathan: Start a non-archival workshop that accepts minimum standards to have a big tent.
 - Jonathan: Trying to find ways to sustain conversations can be hard, e.g. dead Slacks – needs to be an event to have people meet regularly.
 - Iryna: Reflecting on argument mining – also a Dagstuhl Seminar done several times. Longest-lasting effect were the people – students in the seminar and the students of the PIs in the seminar.
 - Iryna: Tutorials, summer schools.
 - Anna: Does this work for norms vs. research areas?
 - Iryna: For norms you also have the institutional factor and scientific debate.
 - André: Have there been any workshops or tutorials?
 - Roy: SustaiNLP.
 - André: Make it about research and not about policies.

- Sasha: These are longer-term things – what about short-term things? E.g. the fact that large LM papers keep winning best paper awards.
- Alexander: Are people in other communities interested in the same things?
- Alexander: Really like workshops and tutorials.
- Alexander: Transformative papers should be recognized as transformative.
- Alexander: Octopus paper was a nice example where a best paper award brought a lot of attention to the paper, beyond a large Twitter following. But it didn't change people's minds – just connected people who already thought the same thing.
- Iryna: Money!
 - Set up large-scale funding programs to support the work. In Germany, funding scheme where you can propose a special topic and they fund the faculty.

5 Panel discussions

5.1 Panel on Equity in NLP research

Colin Raffel (University of North Carolina at Chapel Hill, US), Iryna Gurevych (TU Darmstadt, DE), Alexandra Sasha Luccioni (Hugging Face – Paris, FR), Noah A. Smith (University of Washington – Seattle, US), Emma Strubell (Carnegie Mellon University – Pittsburgh, US), and Thomas Wolf (Hugging Face – Paris, FR)

License © Creative Commons BY 4.0 International license
 © Colin Raffel, Iryna Gurevych, Alexandra Sasha Luccioni, Noah A. Smith, Emma Strubell, and Thomas Wolf

The panel, together with contributions from the audience, discussed a number of aspects relating to equity in NLP research. Two of the most prominent discussion items were as follows. (1) An unequal allocation of resources can lead to a misalignment between real-world problems and research work. Sharing of resources (HPC cluster usage, collaborations), and hiring of researchers with experience and passion for real-world problems may provide some mitigation. (2) The lack of diversity in the research community – e.g., geographically and institutionally – can lead to an over-exposure and hype for certain types of work and research agendas. This leaves little attention for progress being made in domains that deviate from the mainstream. Mitigation strategies can include changing the incentive structures for publication, actively endorsing research work on a personal level, promoting the inclusion of researchers in discussions with different backgrounds, and simplifying communication with people affected by real-world problems relating to NLP.

Participants

- Yuki Arase
Osaka University, JP
- Niranjan Balasubramanian
Stony Brook University, US
- Leon Derczynski
IT University of
Copenhagen, DK
- Jesse Dodge
AI2 – Seattle, US
- Jessica Forde
Brown University –
Providence, US
- Jonathan Frankle
Harvard University –
Allston, US
- Iryna Gurevych
TU Darmstadt, DE
- Michael Hassid
The Hebrew University of
Jerusalem, IL
- Kenneth Heafield
University of Edinburgh, GB
- Sara Hooker
Google – Mountain View, US
- Alexander Koller
Universität des Saarlandes, DE
- Ji-Ung Lee
TU Darmstadt, DE
- Alexander Löser
Berliner Hochschule für
Technik, DE
- Alexandra Sasha Luccioni
Hugging Face – Paris, FR
- André F. T. Martins
IST – Lisbon, PT
- Haritz Puerto
TU Darmstadt, DE
- Colin Raffel
University of North Carolina
at Chapel Hill, US
- Nils Reimers
Hugging Face – Paris
- Leonardo Ribeiro
TU Darmstadt, DE
- Anna Rogers
University of Copenhagen, DK
- Andreas Rücklé
Amazon – Berlin, DE
- Roy Schwartz
The Hebrew University of
Jerusalem, IL
- Edwin Simpson
University of Bristol, GB
- Noam Slonim
IBM – Haifa, IL
- Noah A. Smith
University of Washington –
Seattle, US
- Emma Strubell
Carnegie Mellon University –
Pittsburgh, US
- Betty van Aken
Berliner Hochschule für
Technik, DE
- Thomas Wolf
Hugging Face – Paris, FR

