Six BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP (BlackboxNLP2023)

Singapore
7 December 2023

ISBN: 978-1-7138-8596-2

Printed from e-media with permission by:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571

Some format issues inherent in the e-media version may also appear in this print version.

Copyright© (2023) by the Association for Computational Linguistics
All rights reserved.

Printed with permission by Curran Associates, Inc. (2024)

For permission requests, please contact the Association for Computational Linguistics
at the address below.

Association for Computational Linguistics
209 N. Eighth Street
Stroudsburg, Pennsylvania 18360

Phone: 1-570-476-8006
Fax: 1-570-476-0860

acl@aclweb.org

Additional copies of this publication are available from:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571 USA
Phone: 845-758-0400
Fax: 845-758-2633
Email: curran@proceedings.com
Web: www.proceedings.com

Table of Contents

Knowledge-Grounded Natural Language Recommendation Explanation
Anthony Colas, Jun Araki, Zhengyu Zhou, Bingqing Wang and Zhe Feng 1

Emergent Linear Representations in World Models of Self-Supervised Sequence Models
Neel Nanda, Andrew Lee and Martin Wattenberg 16

Explaining Data Patterns in Natural Language with Language Models
Chandan Singh, John X. Morris, Jyoti Aneja, Alexander Rush and Jianfeng Gao 31

Probing Quantifier Comprehension in Large Language Models: Another Example of Inverse Scaling
Akshat Gupta 56

Disentangling the Linguistic Competence of Privacy-Preserving BERT
Stefan Arnold, Nils Kemmerzell and Annika Schreiner 65

"Honey, Tell Me What's Wrong", Global Explanation of Textual Discriminative Models through Cooperative Generation
Antoine Chaffin and Julien Delaunay 76

Self-Consistency of Large Language Models under Ambiguity
Henning Bartsch, Ole Jorgensen, Domenic Rosati, Jason Hoelscher-Obermaier and Jacob Pfau 89

Character-Level Chinese Backpack Language Models
Hao Sun and John Hewitt 106

Unveiling Multilinguality in Transformer Models: Exploring Language Specificity in Feed-Forward Networks
Sunit Bhattacharya and Ondřej Bojar 120

Why Bother with Geometry? On the Relevance of Linear Decompositions of Transformer Embeddings
Timothee Mickus and Raúl Vázquez 127

Investigating Semantic Subspaces of Transformer Sentence Embeddings through Linear Structural Probing
Dmitry Nikolaev and Sebastian Padó 142

Causal Abstraction for Chain-of-Thought Reasoning in Arithmetic Word Problems
Juanhe (TJ) Tan 155

Enhancing Interpretability Using Human Similarity Judgements to Prune Word Embeddings
Natalia Flechas Manrique, Wanqian Bao, Aurelie Herbelot and Uri Hasson 169

When Your Language Model Cannot Even Do Determiners Right: Probing for Anti-Presuppositions and the Maximize Presupposition! Principle
Judith Sieker and Sina Zarrieß 180

Introducing VULCAN: A Visualization Tool for Understanding Our Models and Data by Example
Jonas Groschwitz 199

The Self-Contained Negation Test Set
David Kletz, Pascal Amsili and Marie Candito 212

Investigating the Effect of Discourse Connectives on Transformer Surprisal: Language Models Understand Connectives, Even So They Are Surprised
Yan Cong, Emmanuele Chersoni, Yu-Yin Hsu and Philippe Blache 222

METAPROBE: A Representation- and Task-Agnostic Probe
Yichu Zhou and Vivek Srikumar 233

How Much Consistency Is Your Accuracy Worth?
Jacob K. Johnson and Ana Marasović 250

Investigating the Encoding of Words in BERT's Neurons Using Feature Textualization
Tanja Baeumel, Soniya Vijayakumar, Josef van Genabith, Guenter Neumann and Simon Ostermann 261

Evaluating Transformer's Ability to Learn Mildly Context-Sensitive Languages
Shunjie Wang and Shane Steinert-Threlkeld 271

Layered Bias: Interpreting Bias in Pretrained Large Language Models
Nirmalendu Prakash and Roy Ka-Wei Lee 284

Not Wacky vs. Definitely Wacky: A Study of Scalar Adverbs in Pretrained Language Models
Isabelle Lorge and Janet B. Pierrehumbert 296

Rigorously Assessing Natural Language Explanations of Neurons
Jing Huang, Atticus Geiger, Karel D'Oosterlinck, Zhengxuan Wu and Christopher Potts 317

NPIs Aren't Exactly Easy: Variation in Licensing across Large Language Models
Deanna DeCarlo, William Palmer, Michael Wilson and Bob Frank 332

Memory Injections: Correcting Multi-Hop Reasoning Failures During Inference in Transformer-Based Language Models
Mansi Sakarvadia, Aswathy Ajith, Arham Khan, Daniel Grzenda, Nathaniel Hudson, André Bauer, Kyle Chard and Ian Foster 342

Systematic Generalization by Finetuning? Analyzing Pretrained Language Models Using Constituency Tests
Aishik Chakraborty, Jackie CK Cheung and Timothy J. O'Donnell 357

On Quick Kisses and How to Make Them Count: A Study on Event Construal in Light Verb Constructions with BERT
Chenxin Liu and Emmanuele Chersoni 367

Identifying and Adapting Transformer-Components Responsible for Gender Bias in an English Language Model
Abhijith Chintam, Rahel Beloch, Willem Zuidema, Michael Hanna and Oskar van der Wal 379