

BlackboxNLP Analyzing and Interpreting Neural Networks for NLP (BlackboxNLP 2022)

Abu Dhabi, United Arab Emirates
8 December 2022

ISBN: 978-1-7138-9104-8

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© (2022) 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

<i>A Minimal Model for Compositional Generalization on gSCAN</i> Alice Hein and Klaus Diepold	1
<i>Sparse Interventions in Language Models with Differentiable Masking</i> Nicola De Cao, Leon Schmid, Dieuwke Hupkes and Ivan Titov	16
<i>Where's the Learning in Representation Learning for Compositional Semantics and the Case of Thematic Fit</i> Mughilan Muthupari, Samrat Halder, Asad B. Sayeed and Yuval Marton	28
<i>Sentence Ambiguity, Grammaticality and Complexity Probes</i> Sunit Bhattacharya, Vilém Zouhar and Ondrej Bojar	40
<i>Post-Hoc Interpretation of Transformer Hyperparameters with Explainable Boosting Machines</i> Kiron Deb, Xuan Zhang and Kevin Duh	51
<i>Revisit Systematic Generalization via Meaningful Learning</i> Ning Shi, Boxin Wang, Wei Wang, Xiangyu Liu and Zhouhan Lin	62
<i>Is It Smaller Than a Tennis Ball? Language Models Play the Game of Twenty Questions</i> Maxime De Bruyn, Ehsan Lotfi, Jeska Buhmann and Walter Daelemans	80
<i>Post-hoc analysis of Arabic transformer models</i> Ahmed Abdelali, Nadir Durrani, Fahim Dalvi and Hassan Sajjad	91
<i>Universal Evasion Attacks on Summarization Scoring</i> Wenchuan Mu and Kwan Hui Lim	104
<i>How (Un)Faithful is Attention?</i> Hessam Amini and Leila Kosseim	119
<i>Are Multilingual Sentiment Models Equally Right for the Right Reasons?</i> Rasmus Kær Jørgensen, Fiammetta Caccavale, Christian Igel and Anders Søgaard	131
<i>Probing for Understanding of English Verb Classes and Alternations in Large Pre-trained Language Models</i> David K Yi, James V. Bruno, Jiayu Han, Peter Zukerman and Shane Steinert-Threlkeld	142
<i>Analyzing Gender Translation Errors to Identify Information Flows between the Encoder and Decoder of a NMT System</i> Guillaume Wisniewski, Lichao Zhu, Nicolas Ballier and François Yvon	153
<i>Human Ratings Do Not Reflect Downstream Utility: A Study of Free-Text Explanations for Model Predictions</i> Jenny Kunz, Martin Jirenius, Oskar Holmström and Marco Kuhlmann	164
<i>Analyzing the Representational Geometry of Acoustic Word Embeddings</i> Badr M. Abdullah and Dietrich Klakow	178
<i>Understanding Domain Learning in Language Models Through Subpopulation Analysis</i> Zheng Zhao, Yftah Ziser and Shay B Cohen	192
<i>Intermediate Entity-based Sparse Interpretable Representation Learning</i> Diego Garcia-Olano, Yasumasa Onoe, Joydeep Ghosh and Byron C Wallace	210

<i>Towards Procedural Fairness: Uncovering Biases in How a Toxic Language Classifier Uses Sentiment Information</i>	Isar Nejadgholi, Esmā Balkir, Kathleen C. Fraser and Svetlana Kiritchenko	225
<i>Investigating the Characteristics of a Transformer in a Few-Shot Setup: Does Freezing Layers in RoBERTa Help?</i>	Digvijay Anil Ingle, Rishabh Kumar Tripathi, Ayush Kumar, Kevin Patel and Jithendra Vepa	238
<i>It Is Not Easy To Detect Paraphrases: Analysing Semantic Similarity With Antonyms and Negation Using the New SemAntoNeg Benchmark</i>	Teemu Vahtola, Mathias Creutz and Jörg Tiedemann	249
<i>Controlling for Stereotypes in Multimodal Language Model Evaluation</i>	Manuj Malik and Richard Johansson	263
<i>On the Compositional Generalization Gap of In-Context Learning</i>	Arian Hosseini, Ankit Vani, Dzmitry Bahdanau, Alessandro Sordani and Aaron Courville . . .	272
<i>Explaining Translationese: why are Neural Classifiers Better and what do they Learn?</i>	Kwabena Amponsah-Kaakyire, Daria Pylypenko, Josef Van Genabith and Cristina España-Bonet	281
<i>Probing GPT-3’s Linguistic Knowledge on Semantic Tasks</i>	Lining Zhang, Mengchen Wang, Liben Chen and Wenxin Zhang	297
<i>Garden Path Traversal in GPT-2</i>	William Jurayj, William Rudman and Carsten Eickhof	305
<i>Testing Pre-trained Language Models’ Understanding of Distributivity via Causal Mediation Analysis</i>	Pangbo Ban, Yifan Jiang, Tianran Liu and Shane Steinert-Threlkeld	314
<i>Using Roark-Hollingshead Distance to Probe BERT’s Syntactic Competence</i>	Jingcheng Niu, Wenjie Lu, Eric Corlett and Gerald Penn	325
<i>DALLE-2 is Seeing Double: Flaws in Word-to-Concept Mapping in Text2Image Models</i>	Royi Rassin, Shauli Ravfogel and Yoav Goldberg	335
<i>Practical Benefits of Feature Feedback Under Distribution Shift</i>	Anurag Katakhar, Clay H. Yoo, Weiqin Wang, Zachary Chase Lipton and Divyansh Kaushik	346
<i>Identifying the Source of Vulnerability in Explanation Discrepancy: A Case Study in Neural Text Classification</i>	Ruixuan Tang, Hanjie Chen and Yangfeng Ji	356
<i>Probing Pretrained Models of Source Codes</i>	Sergey Troshin and Nadezhda Chirkova	371
<i>Probing the representations of named entities in Transformer-based Language Models</i>	Stefan Frederik Schouten, Peter Bloem and Piek Vossen	384
<i>Decomposing Natural Logic Inferences for Neural NLI</i>	Julia Rozanova, Deborah Ferreira, Mokbanarangan Thayaparan, Marco Valentino and Andre Freitas	394
<i>Probing with Noise: Unpicking the Warp and Weft of Embeddings</i>	Filip Klubicka and John D. Kelleher	404

<i>Look to the Right: Mitigating Relative Position Bias in Extractive Question Answering</i> Kazutoshi Shinoda, Saku Sugawara and Akiko Aizawa	418
<i>A Continuum of Generation Tasks for Investigating Length Bias and Degenerate Repetition</i> Darcey Riley and David Chiang	426
<i>Universal and Independent: Multilingual Probing Framework for Exhaustive Model Interpretation and Evaluation</i> Oleg Serikov, Vitaly Protasov, Ekaterina Voloshina, Viktoria Knyazkova and Tatiana Shavrina	441