

Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP (BlackboxNLP2020)

Online
20 November 2020

ISBN: 978-1-7138-1988-2

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© (2020) by the Association for Computational Linguistics
All rights reserved.

Printed with permission by Curran Associates, Inc. (2021)

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>BERTering RAMS: What and How Much does BERT Already Know About Event Arguments? - A Study on the RAMS Dataset</i>	
Varun Gangal and Eduard Hovy	1
<i>Emergent Language Generalization and Acquisition Speed are not tied to Compositionality</i>	
Eugene Kharitonov and Marco Baroni	11
<i>Examining the rhetorical capacities of neural language models</i>	
Zining Zhu, Chuer Pan, Mohamed Abdalla and Frank Rudzicz	16
<i>What Happens To BERT Embeddings During Fine-tuning?</i>	
Amil Merchant, Elahe Rahimtoroghi, Ellie Pavlick and Ian Tenney	33
<i>It's not Greek to mBERT: Inducing Word-Level Translations from Multilingual BERT</i>	
Hila Gonen, Shauli Ravfogel, Yanai Elazar and Yoav Goldberg	45
<i>Leveraging Extracted Model Adversaries for Improved Black Box Attacks</i>	
Naveen Jafer Nizar and Ari Kobren	57
<i>On the Interplay Between Fine-tuning and Sentence-Level Probing for Linguistic Knowledge in Pre-Trained Transformers</i>	
Marius Mosbach, Anna Khokhlova, Michael A. Hedderich and Dietrich Klakow	68
<i>Unsupervised Evaluation for Question Answering with Transformers</i>	
Lukas Muttenthaler, Isabelle Augenstein and Johannes Bjerva	83
<i>Unsupervised Distillation of Syntactic Information from Contextualized Word Representations</i>	
Shauli Ravfogel, Yanai Elazar, Jacob Goldberger and Yoav Goldberg	91
<i>The Explanation Game: Towards Prediction Explainability through Sparse Communication</i>	
Marcos Treviso and André F. T. Martins	107
<i>Latent Tree Learning with Ordered Neurons: What Parses Does It Produce?</i>	
Yian Zhang	119
<i>Linguistically-Informed Transformations (LIT): A Method for Automatically Generating Contrast Sets</i>	
Chuanrong Li, Lin Shengshuo, Zeyu Liu, Xinyi Wu, Xuhui Zhou and Shane Steinert-Threlkeld	126
<i>Controlling the Imprint of Passivization and Negation in Contextualized Representations</i>	
Hande Celikkanat, Sami Virpioja, Jörg Tiedemann and Marianna Apidianaki	136
<i>The elephant in the interpretability room: Why use attention as explanation when we have saliency methods?</i>	
Jasmijn Bastings and Katja Filippova	149
<i>How does BERT capture semantics? A closer look at polysemous words</i>	
David Yenicelik, Florian Schmidt and Yannic Kilcher	156
<i>Neural Natural Language Inference Models Partially Embed Theories of Lexical Entailment and Negation</i>	
Atticus Geiger, Kyle Richardson and Christopher Potts	163

<i>BERTnesia: Investigating the capture and forgetting of knowledge in BERT</i>	
Jaspreet Singh, Jonas Wallat and Avishek Anand	N/A
<i>Probing for Multilingual Numerical Understanding in Transformer-Based Language Models</i>	
Devin Johnson, Denise Mak, Andrew Barker and Lexi Loessberg-Zahl	184
<i>Dissecting Lottery Ticket Transformers: Structural and Behavioral Study of Sparse Neural Machine Translation</i>	
Rajiv Movva and Jason Zhao	193
<i>Exploring Neural Entity Representations for Semantic Information</i>	
Andrew Runge and Eduard Hovy	204
<i>BERTs of a feather do not generalize together: Large variability in generalization across models with similar test set performance</i>	
R. Thomas McCoy, Junghyun Min and Tal Linzen	217
<i>Attacking Semantic Similarity: Generating Second-Order NLP Adversarial Examples</i>	
John Morris	228
<i>Discovering the Compositional Structure of Vector Representations with Role Learning Networks</i>	
Paul Soulos, R. Thomas McCoy, Tal Linzen and Paul Smolensky	238
<i>Structured Self-Attention Weights Encodes Semantics in Sentiment Analysis</i>	
Zhengxuan Wu, Thanh-Son Nguyen and Desmond Ong	255
<i>Investigating Novel Verb Learning in BERT: Selectional Preference Classes and Alternation-Based Syntactic Generalization</i>	
Tristan Thrush, Ethan Wilcox and Roger Levy	265
<i>The EOS Decision and Length Extrapolation</i>	
Benjamin Newman, John Hewitt, Percy Liang and Christopher D. Manning	276
<i>Do Language Embeddings capture Scales?</i>	
Xikun Zhang, Deepak Ramachandran, Ian Tenney, Yanai Elazar and Dan Roth	292
<i>Evaluating Attribution Methods using White-Box LSTMs</i>	
Yiding Hao	300
<i>Defining Explanation in an AI Context</i>	
Tejaswani Verma, Christoph Lingenfelder and Dietrich Klakow	314
<i>Searching for a Search Method: Benchmarking Search Algorithms for Generating NLP Adversarial Examples</i>	
Jin Yong Yoo, John Morris, Eli Lifland and Yanjun Qi	323
<i>This is a BERT. Now there are several of them. Can they generalize to novel words?</i>	
Coleman Haley	333
<i>diagNNose: A Library for Neural Activation Analysis</i>	
Jaap Jumelet	342