

Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP (BlackboxNLP 2021)

Punta Cana, Dominican Republic
11 November 2021

ISBN: 978-1-7138-9170-3

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© (2021) 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>To what extent do human explanations of model behavior align with actual model behavior?</i> Grusha Prasad, Yixin Nie, Mohit Bansal, Robin Jia, Douwe Kiela and Adina Williams	1
<i>Test Harder than You Train: Probing with Extrapolation Splits</i> Jenny Kunz and Marco Kuhlmann	15
<i>Does External Knowledge Help Explainable Natural Language Inference? Automatic Evaluation vs. Human Ratings</i> Hendrik Schuff, Hsiu-Yu Yang, Heike Adel and Ngoc Thang Vu	26
<i>The Language Model Understood the Prompt was Ambiguous: Probing Syntactic Uncertainty Through Generation</i> Laura Aina and Tal Linzen	42
<i>On the Limits of Minimal Pairs in Contrastive Evaluation</i> Jannis Vamvas and Rico Sennrich	58
<i>What Models Know About Their Attackers: Deriving Attacker Information From Latent Representations</i> Zhouhang Xie, Jonathan Brophy, Adam Noack, Wencong You, Kalyani Asthana, Carter Perkins, Sabrina Reis, Zayd Hammoudeh, Daniel Lowd and Sameer Singh	69
<i>ALL Dolphins Are Intelligent and SOME Are Friendly: Probing BERT for Nouns' Semantic Properties and their Prototypicality</i> Marianna Apidianaki and Aina Garí Soler	79
<i>ProSPer: Probing Human and Neural Network Language Model Understanding of Spatial Perspective</i> Tessa Masis and Carolyn Anderson	95
<i>Can Transformers Jump Around Right in Natural Language? Assessing Performance Transfer from SCAN</i> Rahma Chaabouni, Roberto Dessì and Eugene Kharitonov	136
<i>Transferring Knowledge from Vision to Language: How to Achieve it and how to Measure it?</i> Tobias Norlund, Lovisa Hagström and Richard Johansson	149
<i>Discrete representations in neural models of spoken language</i> Bertrand Higy, Lieke Gelderloos, Afra Alishahi and Grzegorz Chrupała	163
<i>Word Equations: Inherently Interpretable Sparse Word Embeddings through Sparse Coding</i> Adly Templeton	177
<i>A howling success or a working sea? Testing what BERT knows about metaphors</i> Paolo Pedinotti, Eliana Di Palma, Ludovica Cerini and Alessandro Lenci	192
<i>How Length Prediction Influence the Performance of Non-Autoregressive Translation?</i> Minghan Wang, GUO Jiaxin, Yuxia Wang, Yimeng Chen, Su Chang, Hengchao Shang, Min Zhang, Shimin Tao and Hao Yang	205
<i>On the Language-specificity of Multilingual BERT and the Impact of Fine-tuning</i> Marc Tanti, Lonneke van der Plas, Claudia Borg and Albert Gatt	214

<i>Relating Neural Text Degeneration to Exposure Bias</i> Ting-Rui Chiang and Yun-Nung Chen	228
<i>Efficient Explanations from Empirical Explainers</i> Robert Schwarzenberg, Nils Feldhus and Sebastian Möller	240
<i>Variation and generality in encoding of syntactic anomaly information in sentence embeddings</i> Qinxuan Wu and Allyson Ettinger	250
<i>Enhancing Interpretable Clauses Semantically using Pretrained Word Representation</i> Rohan Kumar Yadav, Lei Jiao, Ole-Christoffer Granmo and Morten Goodwin	265
<i>Analyzing BERT's Knowledge of Hypernymy via Prompting</i> Michael Hanna and David Mareček	275
<i>An in-depth look at Euclidean disk embeddings for structure preserving parsing</i> Federico Fancellu, Lan Xiao, Allan Jepsen and Afsaneh Fazly	283
<i>Training Dynamic based data filtering may not work for NLP datasets</i> Arka Talukdar, Monika Dagar, Prachi Gupta and Varun Menon	296
<i>Multi-Layer Random Perturbation Training for improving Model Generalization Efficiently</i> Lis Kanashiro Pereira, Yuki Taya and Ichiro Kobayashi	303
<i>Screening Gender Transfer in Neural Machine Translation</i> Guillaume Wisniewski, Lichao Zhu, Nicolas Bailler and François Yvon	311
<i>What BERT Based Language Model Learns in Spoken Transcripts: An Empirical Study</i> Ayush Kumar, Mukuntha Narayanan Sundararaman and Jithendra Vepa	322
<i>Assessing the Generalization Capacity of Pre-trained Language Models through Japanese Adversarial Natural Language Inference</i> Hitomi Yanaka and Koji Mineshima	337
<i>Investigating Negation in Pre-trained Vision-and-language Models</i> Radina Dobрева and Frank Keller	350
<i>Not all parameters are born equal: Attention is mostly what you need</i> Nikolay Bogoychev	363
<i>Not All Models Localize Linguistic Knowledge in the Same Place: A Layer-wise Probing on BERToids' Representations</i> Mohsen Fayyaz, Ehsan Aghazadeh, Ali Modarressi, Hosein Mohebbi and Mohammad Taher Pilehvar	375
<i>Learning Mathematical Properties of Integers</i> Maria Ryskina and Kevin Knight	389
<i>Probing Language Models for Understanding of Temporal Expressions</i> Shivin Thukral, Kunal Kukreja and Christian Kavouras	396
<i>How Familiar Does That Sound? Cross-Lingual Representational Similarity Analysis of Acoustic Word Embeddings</i> Badr Abdullah, Iuliia Zaitova, Tania Avgustinova, Bernd Möbius and Dietrich Klakow	407

<i>Perturbing Inputs for Fragile Interpretations in Deep Natural Language Processing</i> Sanchit Sinha, Hanjie Chen, Arshdeep Sekhon, Yangfeng Ji and Yanjun Qi	420
<i>An Investigation of Language Model Interpretability via Sentence Editing</i> Samuel Stevens and Yu Su	435
<i>Interacting Knowledge Sources, Inspection and Analysis: Case-studies on Biomedical text processing</i> Parsa Bagherzadeh and Sabine Bergler	447
<i>Attacks against Ranking Algorithms with Text Embeddings: A Case Study on Recruitment Algorithms</i> Anahita Samadi, debapriya banerjee and Shirin Nilizadeh	457
<i>Controlled tasks for model analysis: Retrieving discrete information from sequences</i> Ionut-Teodor Sorodoc, Gemma Boleda and Marco Baroni	468
<i>The Acceptability Delta Criterion: Testing Knowledge of Language using the Gradient of Sentence Acceptability</i> Héctor Vázquez Martínez	479
<i>How Does BERT Rerank Passages? An Attribution Analysis with Information Bottlenecks</i> Zhiying Jiang, Raphael Tang, Ji Xin and Jimmy Lin	496
<i>Do Language Models Know the Way to Rome?</i> Bastien Liétard, Mostafa Abdou and Anders Søgaard	510
<i>Exploratory Model Analysis Using Data-Driven Neuron Representations</i> Daisuke Oba, Naoki Yoshinaga and Masashi Toyoda	518
<i>Fine-Tuned Transformers Show Clusters of Similar Representations Across Layers</i> Jason Phang, Haokun Liu and Samuel R. Bowman	529
<i>BERT Has Uncommon Sense: Similarity Ranking for Word Sense BERTology</i> Luke Gessler and Nathan Schneider	539