

Second BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP at ACL 2019

Florence, Italy
1 August 2019

ISBN: 978-1-5108-9109-8

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© (2019) by the Association for Computational Linguistics
All rights reserved.

Printed by Curran Associates, Inc. (2019)

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>Transcoding Compositionally: Using Attention to Find More Generalizable Solutions</i> Kris Korrel, Dieuwke Hupkes, Verna Dankers and Elia Bruni	1
<i>Sentiment Analysis Is Not Solved! Assessing and Probing Sentiment Classification</i> Jeremy Barnes, Lilja Øvrelid and Erik Velldal	12
<i>Second-order Co-occurrence Sensitivity of Skip-Gram with Negative Sampling</i> Dominik Schlechtweg, Cennet Oguz and Sabine Schulte im Walde	24
<i>Can Neural Networks Understand Monotonicity Reasoning?</i> Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze and Johan Bos	31
<i>Multi-Granular Text Encoding for Self-Explaining Categorization</i> Zhiguo Wang, Yue Zhang, Mo Yu, Wei Zhang, Lin Pan, Linfeng Song, Kun Xu and Yousef El-Kurdi 41	
<i>The Meaning of "Most" for Visual Question Answering Models</i> Alexander Kuhnle and Ann Copestake	46
<i>Do Human Rationales Improve Machine Explanations?</i> Julia Strout, Ye Zhang and Raymond Mooney	56
<i>Analyzing the Structure of Attention in a Transformer Language Model</i> Jesse Vig and Yonatan Belinkov	63
<i>Detecting Political Bias in News Articles Using Headline Attention</i> Rama Rohit Reddy Gangula, Suma Reddy Duggenpudi and Radhika Mamidi	77
<i>Testing the Generalization Power of Neural Network Models across NLI Benchmarks</i> Aarne Talman and Stergios Chatzikyriakidis	85
<i>Character Eyes: Seeing Language through Character-Level Taggers</i> Yuval Pinter, Marc Marone and Jacob Eisenstein	95
<i>Faithful Multimodal Explanation for Visual Question Answering</i> Jialin Wu and Raymond Mooney	103
<i>Evaluating Recurrent Neural Network Explanations</i> Leila Arras, Ahmed Osman, Klaus-Robert Müller and Wojciech Samek	113
<i>On the Realization of Compositionality in Neural Networks</i> Joris Baan, Jana Leible, Mitja Nikolaus, David Rau, Dennis Ulmer, Tim Baumgärtner, Dieuwke Hupkes and Elia Bruni	127
<i>Learning the Dyck Language with Attention-based Seq2Seq Models</i> Xiang Yu, Ngoc Thang Vu and Jonas Kuhn	138
<i>Modeling Paths for Explainable Knowledge Base Completion</i> Josua Stadelmaier and Sebastian Padó	147
<i>Probing Word and Sentence Embeddings for Long-distance Dependencies Effects in French and English</i> Paola Merlo	158

<i>Derivational Morphological Relations in Word Embeddings</i> Tomáš Musil, Jonáš Vidra and David Mareček	173
<i>Hierarchical Representation in Neural Language Models: Suppression and Recovery of Expectations</i> Ethan Wilcox, Roger Levy and Richard Futrell	181
<i>Blackbox Meets Blackbox: Representational Similarity & Stability Analysis of Neural Language Models and Brains</i> Samira Abnar, Lisa Beinborn, Rochelle Choenni and Willem Zuidema	191
<i>An LSTM Adaptation Study of (Un)grammaticality</i> Shammur Absar Chowdhury and Roberto Zamparelli	204
<i>An Analysis of Source-Side Grammatical Errors in NMT</i> Antonios Anastasopoulos	213
<i>Finding Hierarchical Structure in Neural Stacks Using Unsupervised Parsing</i> William Merrill, Lenny Khazan, Noah Amsel, Yiding Hao, Simon Mendelsohn and Robert Frank 224	
<i>Adversarial Attack on Sentiment Classification</i> Yi-Ting Tsai, Min-Chu Yang and Han-Yu Chen	233
<i>Open Sesame: Getting inside BERT's Linguistic Knowledge</i> Yongjie Lin, Yi Chern Tan and Robert Frank	241
<i>GEval: Tool for Debugging NLP Datasets and Models</i> Filip Graliński, Anna Wróblewska, Tomasz Stanisławek, Kamil Grabowski and Tomasz Górecki 254	
<i>From Balustrades to Pierre Vinken: Looking for Syntax in Transformer Self-Attentions</i> David Mareček and Rudolf Rosa	263
<i>What Does BERT Look at? An Analysis of BERT's Attention</i> Kevin Clark, Urvashi Khandelwal, Omer Levy and Christopher D. Manning	276