First Workshop on Trustworthy Natural Language Processing (TrustNLP 2021)

Online
10 June 2021

ISBN: 978-1-7138-3032-0

Printed from e-media with permission by:

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

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

Interpretability Rules: Jointly Bootstrapping a Neural Relation Extractorwith an Explanation Decoder
Zheng Tang and Mihai Surdeanu 1

Measuring Biases of Word Embeddings: What Similarity Measures and Descriptive Statistics to Use?
Hossein Azarpanah and Mohsen Farhadloo 8

Private Release of Text Embedding Vectors
Oluwaseyi Feyisetan and Shiva Kasiviswanathan 15

Accountable Error Characterization
Amita Misra, Zhe Liu and Jalal Mahmud 28

xER: An Explainable Model for Entity Resolution using an Efficient Solution for the Clique Partitioning Problem
Samhita Vadrevu, Rakesh Nagi, JinJun Xiong and Wen-mei Hwu 34

Gender Bias in Natural Language Processing Across Human Languages
Abigail Matthews, Isabella Grasso, Christopher Mahoney, Yan Chen, Esma Wali, Thomas Middleton, Mariama Njie and Jeanna Matthews 45

Interpreting Text Classifiers by Learning Context-sensitive Influence of Words
Sawan Kumar, Kalpit Dixit and Kashif Shah 55

Towards Benchmarking the Utility of Explanations for Model Debugging
Maximilian Idahl, Lijun Lyu, Ujwal Gadiraju and Avishek Anand 68