1st EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP 2018

Brussels, Belgium
1 November 2018

ISBN: 978-1-5108-7421-3

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.

Printed by Curran Associates, Inc. (2019)

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

Keynote Talks

Trying to Understand Recurrent Neural Networks for Language Processing.
Yoav Goldberg xvi

Learning with Latent Linguistic Structure.
Graham Neubig xvii

Language representations in human brains and artificial neural networks.
Leila Wehbe xviii

Archival Papers

When does deep multi-task learning work for loosely related document classification tasks?
Emma Kerinec, Chloé Braud and Anders Søgaard 1

Analyzing Learned Representations of a Deep ASR Performance Prediction Model
Zied Elloumi, Laurent Besacier, Olivier Galibert and Benjamin Lecouteux 9

Explaining non-linear Classifier Decisions within Kernel-based Deep Architectures
Danilo Croce, Daniele Rossini and Roberto Basili 16

Nightmare at test time: How punctuation prevents parsers from generalizing
Anders Søgaard, Miryam de Lhoneux and Isabelle Augenstein 25

Evaluating Textual Representations through Image Generation
Graham Spinks and Marie-Francine Moens 30

On the Role of Text Preprocessing in Neural Network Architectures: An Evaluation Study on Text Categorization and Sentiment Analysis
Jose Camacho-Collados and Mohammad Taher Pilehvar 40

Jump to better conclusions: SCAN both left and right
Joost Bastings, Marco Baroni, Jason Weston, Kyunghyun Cho and Douwe Kiela 47

Understanding Convolutional Neural Networks for Text Classification
Alon Jacovi, Oren Sar Shalom and Yoav Goldberg 56

Linguistic representations in multi-task neural networks for ellipsis resolution
Ola Rønning, Daniel Hardt and Anders Søgaard 66

Unsupervised Token-wise Alignment to Improve Interpretation of Encoder-Decoder Models
Shun Kiyono, Sho Takase, Jun Suzuki, Naoaki Okazaki, Kentaro Inui and Masaaki Nagata 74

Rule induction for global explanation of trained models
Madhumita Sushil, Simon Suster and Walter Daelemans 82

Can LSTM Learn to Capture Agreement? The Case of Basque
Shauli Ravfogel, Yoav Goldberg and Francis Tyers 98

Rearranging the Familiar: Testing Compositional Generalization in Recurrent Networks
Joao Loula, Marco Baroni and Brenden Lake 108

Evaluating the Ability of LSTMs to Learn Context-Free Grammars
Luzi Sennhauser and Robert Berwick 115

Interpretable Neural Architectures for Attributing an Ad's Performance to its Writing Style
Reid Pryzant, Sugato Basu and Kazoo Sone 125

Interpreting Neural Networks with Nearest Neighbors
Eric Wallace, Shi Feng and Jordan Boyd-Graber 136

'Indicatements' that character language models learn English morpho-syntactic units and regularities
Yova Kementchedjhieva and Adam Lopez 145

LISA: Explaining Recurrent Neural Network Judgments via Layer-wIse Semantic Accumulation and Example to Pattern Transformation
Pankaj Gupta and Hinrich Schütze 154

Analysing the potential of seq-to-seq models for incremental interpretation in task-oriented dialogue
Dieuwke Hupkes, Sanne Bouwmeester and Raquel Fernández 165

An Operation Sequence Model for Explainable Neural Machine Translation
Felix Stahlberg, Danielle Saunders and Bill Byrne 175

Introspection for convolutional automatic speech recognition
Andreas Krug and Sebastian Stober 187

Learning and Evaluating Sparse Interpretable Sentence Embeddings
Valentin Trifonov, Octavian-Eugen Ganea, Anna Potapenko and Thomas Hofmann 200

What do RNN Language Models Learn about Filler–Gap Dependencies?
Ethan Wilcox, Roger Levy, Takashi Morita and Richard Futrell 211

Do Language Models Understand Anything? On the Ability of LSTMs to Understand Negative Polarity Items
Jaap Jumelet and Dieuwke Hupkes 222

Closing Brackets with Recurrent Neural Networks
Natalia Skachkova, Thomas Trost and Dietrich Klakow 232

Under the Hood: Using Diagnostic Classifiers to Investigate and Improve how Language Models Track Agreement Information
Mario Giulianelli, Jack Harding, Florian Mohnert, Dieuwke Hupkes and Willem Zuidema 240

Iterative Recursive Attention Model for Interpretable Sequence Classification
Martin Tutek and Jan Šnajder 249

Interpreting Word-Level Hidden State Behaviour of Character-Level LSTM Language Models
Avery Hiebert, Cole Peterson, Alona Fyshe and Nishant Mehta 258

Importance of Self-Attention for Sentiment Analysis
Gaël Letarte, Frédérik Paradis, Philippe Giguère and François Laviolette 267

Firearms and Tigers are Dangerous, Kitchen Knives and Zebras are Not: Testing whether Word Embeddings Can Tell
Pia Sommerauer and Antske Fokkens 276

An Analysis of Encoder Representations in Transformer-Based Machine Translation
Alessandro Raganato and Jörg Tiedemann 287

Evaluating Grammaticality in Seq2seq Models with a Broad Coverage HPSG Grammar: A Case Study on Machine Translation
Johnny Wei, Khiem Pham, Brendan O'Connor and Brian Dillon 298

Context-Free Transductions with Neural Stacks
Yiding Hao, William Merrill, Dana Angluin, Robert Frank, Noah Amsel, Andrew Benz and Simon Mendelsohn 306

Extended Abstracts

Learning Explanations from Language Data
David Harbecke, Robert Schwarzenberg and Christoph Alt 316

How much should you ask? On the question structure in QA systems.
Barbara Rychalska, Dominika Basaj, Anna Wróblewska and Przemyslaw Biecek 319

Does it care what you asked? Understanding Importance of Verbs in Deep Learning QA System
Barbara Rychalska, Dominika Basaj, Anna Wróblewska and Przemyslaw Biecek 322

Interpretable Textual Neuron Representations for NLP
Nina Poerner, Benjamin Roth and Hinrich Schütze 325

Language Models Learn POS First
Naomi Saphra and Adam Lopez 328

Predicting and interpreting embeddings for out of vocabulary words in downstream tasks
Nicolas Garneau, Jean-Samuel Leboeuf and Luc Lamontagne 331

Probing sentence embeddings for structure-dependent tense
Geoff Bacon and Terry Regier 334

Collecting Diverse Natural Language Inference Problems for Sentence Representation Evaluation
Adam Poliak, Aparajita Haldar, Rachel Rudinger, J. Edward Hu, Ellie Pavlick, Aaron Steven White and Benjamin Van Durme 337

Interpretable Word Embedding Contextualization
Kyoung-Rok Jang and Sung-Hyon Myaeng 341

State Gradients for RNN Memory Analysis
Lyan Verwimp, Hugo Van hamme, Vincent Renkens and Patrick Wambacq 344

Extracting Syntactic Trees from Transformer Encoder Self-Attentions
David Mareček and Rudolf Rosa 347

Portable, layer-wise task performance monitoring for NLP models
Tom Lippincott 350

GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding
Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy and Samuel Bowman . . . 353

Explicitly modeling case improves neural dependency parsing
Clara Vania and Adam Lopez . . . 356

Language Modeling Teaches You More than Translation Does: Lessons Learned Through Auxiliary Syntactic Task Analysis
Kelly Zhang and Samuel Bowman . . . 359

Representation of Word Meaning in the Intermediate Projection Layer of a Neural Language Model
Steven Derby, Paul Miller, Brian Murphy and Barry Devereux . . . 362

Interpretable Structure Induction via Sparse Attention
Ben Peters, Vlad Niculae and André F. T. Martins . . . 365

Debugging Sequence-to-Sequence Models with Seq2Seq-Vis
Hendrik Strobelt, Sebastian Gehrmann, Michael Behrisch, Adam Perer, Hanspeter Pfister and Alexander Rush . . . 368

Grammar Induction with Neural Language Models: An Unusual Replication
Phu Mon Htut, Kyunghyun Cho and Samuel Bowman . . . 371

Does Syntactic Knowledge in Multilingual Language Models Transfer Across Languages?
Prajit Dhar and Arianna Bisazza . . . 374

Exploiting Attention to Reveal Shortcomings in Memory Models
Kaylee Burns, Aida Nematzadeh, Erin Grant, Alison Gopnik and Tom Griffiths . . . 378

End-to-end Image Captioning Exploits Distributional Similarity in Multimodal Space
Pranava Swaroop Madhyastha, Josiah Wang and Lucia Specia . . . 381

Limitations in learning an interpreted language with recurrent models
Denis Paperno . . . 384