EMNLP 2018

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

Proceedings of the First Workshop

November 1, 2018
Brussels, Belgium

Sponsored by:

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 978-1-948087-71-1

Introduction

BlackboxNLP is the first workshop on analyzing and interpreting neural networks for NLP, hosted by the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018) in Brussels, Belgium.

The goal of this workshop is to bring together people who are attempting to peek inside the neural network black box, taking inspiration from machine learning, psychology, linguistics and neuroscience. Neural networks have rapidly become a central component in language and speech understanding systems in the last few years. The improvements in accuracy and performance brought by the introduction of neural networks has typically come at the cost of our understanding of the system: what are the representations and computations that the network learns?

We received an impressive number of 76 submissions (including both archival papers and extended abstracts), suggesting that the issue of interpretability of neural networks is timely and important within the NLP community. The final program contains three keynote talks, eight oral presentations and 47 posters. We hope this workshop provides a venue for bringing together ideas and stimulate new ways of building methods and resources for facilitating better analysis and understanding of the inner-dynamics of neural networks for NLP.

BlackboxNLP would not have been possible without the dedication of its program committee. We would like to thank them for their invaluable effort in providing high-quality reviews in a very short period of time and for a higher number of submission originally expected. We are also grateful to our invited speakers, Leila Wehbe, Graham Neubig and Yoav Goldberg for contributing to our program. Finally, we are very thankful to our sponsors, Amazon and the Department of Cognitive Science, Johns Hopkins University for supporting the workshop.

Tal Linzen, Grzegorz Chrupała and Afra Alishahi

Organizers:

Tal Linzen, Johns Hopkins University
Grzegorz Chrupała, Tilburg University
Afra Alishahi, Tilburg University

Program Committee:

Željko Agić, IT University of Copenhagen
Niranjan Balasubramanian, Stony Brook University
Roberto Basili, University of Roma, Tor Vergata
Laurent Besacier, LIG
Yonatan Belinkov, MIT CSAIL
Or Biran, n-Join
Pravesh Biyani, IIIT Delhi
Arianna Bisazza, Leiden University
Samuel Bowman, New York University
Bill Byrne, University of Cambridge
Kyunghyun Cho, New York University
Ryan Cotterell, Johns Hopkins University
Barry Devereux, Queen's University, Belfast
Ewan Dunbar, Ecole Normale Supérieure et Ecole des Hautes Etudes en Sciences Sociales
Indranil Dutta, The English and Foreign Languages University
Allyson Ettinger, University of Maryland
Antske Fokkens, VU Amsterdam
Robert Frank, Yale University
Alona Fyshe, University of Alberta
Lieke Gelderloos, Tilburg University
Yoav Goldberg, Bar Ilan University
John Hale, Cornell University and Google DeepMind
David Harwath, Massachusetts Institute of Technology
Ákos Kádár, Tilburg University
Philipp Koehn, Johns Hopkins University
Adhiguna Kuncoro, University of Oxford and DeepMind
Ignacio Iacobacci, Sapienza University of Rome
Angeliki Lazaridou, DeepMind
Miryam de Lhoneux, Uppsala University
Nelson F. Liu, University of Washington
Adam Lopez, University of Edinburgh
David Mareček, Charles University in Prague
Rebecca Marvin, Johns Hopkins University
Paola Merlo, University of Geneva
Marie-Francine Moens, KU Leuven
Yves Peirsman, NLP Town
Mohammad Taher Pilehvar, University of Cambridge
Barbara Plank, IT University of Copenhagen
Delip Rao, Johns Hopkins University
Brian Roark, Google Inc.

Jan Šnajder, University of Zagreb
Whitney Tabor, University of Connecticut
Adina Williams, New York University
Fabio Massimo Zanzotto, University of Rome Tor Vergata
Willem Zuidema, University of Amsterdam

Invited Speakers:

Yoav Goldberg, Bar Ilan University
Graham Neubig, Carnegie Mellon University
Leila Wehbe, Carnegie Mellon University

Table of Contents

Keynote Talks

Archival Papers

Extended Abstracts

Conference Program

09:00-09:10 Opening Remarks

09:10-10:00 Invited Talk: Yoav Goldberg

10:00-11:00 Poster Session 1

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

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

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

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

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

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

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

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

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

Language Models Learn POS First
Naomi Saphra and Adam Lopez

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

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

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

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

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

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

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

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

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

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

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

Interpretable Word Embedding Contextualization
Kyoung-Rok Jang, Sung-Hyon Myaeng and Sang-Bum Kim

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

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

10:30-11:00 Coffee Break

11:00-12:30 Oral Presentations

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

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

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

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

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

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

12:30-14:00 Lunch Break

14:00-14:50 Invited Talk: Graham Neubig

14:50-16:00 Poster Session 2

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

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

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

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

Introspection for convolutional automatic speech recognition
Andreas Krug and Sebastian Stober

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Diagnosing Failures in Question Answering Tasks with Attention
Aida Nematzadeh, Kaylee Burns, Erin Grant and Tom Griffiths

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

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

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

Limitations in learning an interpreted language with recurrent models
Denis Paperno

16:00-16:50 Invited Talk: Leila Wehbe

16:50-17:20 Oral Presentations Session 2

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

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

17:20-17:30 Best Paper Announcement and Closing Remarks

Keynote Talk

Trying to Understand Recurrent Neural Networks for Language Processing.

Yoav Goldberg

Bar Ilan University

Abstract

Recurrent neural networks (RNNs), and in particular LSTM networks, emerge as very capable learners for sequential data. Thus, my group started using them everywhere, achieving strong results on many language understanding and modeling tasks. However, little is known about how RNNs represent sequences, what they actually encode, and what they are capable representing. In this talk, I will describe some attempts at trying to shed light on the inner-working of RNNs. Particularly, I plan to describe at least two of the following: a method for comparing what is captured in vector representations of sentences based on different encoders (Adi et al, ICLR 2017, and more generally the notion of diagnostic classification), a framework for extracting a finite-state automata from trained RNNs (Weiss et al, ICML 2018), and a formal difference between the representation capacity of different RNN variants (Weiss et al, ACL 2018).

Biography of the Speaker

Yoav Goldberg is a Senior Lecturer at Bar Ilan University's Computer Science Department. Before that, he was a Research Scientist at Google Research New York. He works on problems related to Natural Language Processing and Machine Learning. In particular he is interested in syntactic parsing, structured-prediction models, learning for greedy decoding algorithms, multilingual language understanding, and cross domain learning. Lately, he is also interested in neural network based methods for NLP. He recently published a book on the subject.

Keynote Talk

Learning with Latent Linguistic Structure

Graham Neubig

Carnegie Mellon University

Abstract

Neural networks provide a powerful tool to model language, but also depart from standard methods of linguistic representation, which usually consist of discrete tag, tree, or graph structures. These structures are useful for a number of reasons: they are more interpretable, and also can be useful in downstream tasks. In this talk, I will discuss models that explicitly incorporate these structures as latent variables, allowing for unsupervised or semi-supervised discovery of interpretable linguistic structure, with applications to part-of-speech and morphological tagging, as well as syntactic and semantic parsing.

Biography of the Speaker

Graham Neubig is an assistant professor at the Language Technologies Intitute of Carnegie Mellon University. His work focuses on natural language processing, specifically multi-lingual models that work in many different languages, and natural language interfaces that allow humans to communicate with computers in their own language. Much of this work relies on machine learning to create these systems from data, and he is also active in developing methods and algorithms for machine learning over natural language data. He publishes regularly in the top venues in natural language processing, machine learning, and speech, and his work occasionally wins awards such as best papers at EMNLP, EACL, and WNMT. He is also active in developing open-source software, and is the main developer of the DyNet neural network toolkit.

Keynote Talk

Language representations in human brains and artificial neural networks

Leila Wehbe

Carnegie Mellon University

Abstract

When studying language in the brain, it has become more common to image the brain of humans while they process naturalistic language stimuli consisting of rich, natural text. To analyse the brain representation of such complex stimuli, vector representations derived from various NLP methods are extremely useful as a model of the information being processed in the brain. The recent deep learning revolution has ignited a lot of interest in using artificial neural networks as a source of high dimensional vector representation for modeling brain processes. However, these representations are hard to interpret and the problem becomes increasingly difficult: how do we study complex brain activity – a black box we want to understand – using hard-to-interpret artificial neural network representations – another black box we want to understand? In this talk, I will summarize the recent effort in modeling the brain processing of language, the use of artificial neural networks in this process, and how inferences about brain processes and about artificial neural network representations can still be made under this setup.

Biography of the Speaker

Leila Wehbe is an assistant professor of Machine Learning at Carnegie Mellon University. Previously, we was a postdoctoral researcher at the Gallant Lab in the Helen Wills Neuroscience Institute at UC Berkeley. She obtained her PhD from the Machine Learning Department and the Center for the Neural Basis of Cognition at Carnegie Mellon University, where she worked with Tom Mitchell. She works on studying language representations in the brain when subjects engage in naturalistic language tasks. Specifically, she combines functional neuroimaging with natural language processing and machine learning tools to build spatiotemporal maps of the information represented in the brain during language processing.