# Tutorial Proposal: Interpretability and Analysis in Neural NLP

**Yonatan Belinkov**
Harvard University and MIT

**Sebastian Gehrmann**
Google AI

**Ellie Pavlick**
Brown University

## Abstract

While deep learning has transformed the natural language processing (NLP) field and impacted the larger computational linguistics community, the rise of neural networks is stained by their opaque nature: It is challenging to interpret the inner workings of neural network models, and explicate their behavior. Therefore, in the last few years, an increasingly large body of work has been devoted to the analysis and interpretation of neural network models in NLP.

This body of work is so far lacking a common framework and methodology. Moreover, approaching the analysis of modern neural networks can be difficult for newcomers to the field. This tutorial aims to fill this gap and introduce the nascent field of interpretability and analysis of neural networks in NLP.

The tutorial will cover the main lines of analysis work, such as structural analyses using probing classifiers, behavioral studies and test suites, and interactive visualizations. We will highlight not only the most commonly applied analysis methods, but also the specific limitations and shortcomings of current approaches, in order to inform participants where to focus future efforts.

## 1 Tutorial Description

Deep learning has transformed the NLP field and impacted the larger computational linguistics community. Neural networks have become the preferred modeling approach for various tasks, from language modeling, through morphological inflection and syntactic parsing, to machine translation, summarization, and reading comprehension.

The rise of neural networks is, however, stained by their opaque nature. In contrast to earlier approaches that made use of manually crafted features, it is more challenging to interpret the inner workings of neural network models, and explicate their behavior. Therefore, in the last few years, an increasingly large body of work has been devoted to the analysis and interpretation of neural network models in NLP.

The topic has so far been represented in two dedicated workshops (Blackbox 2018 and 2019) and was recently established as a track in the main *CL conferences. Due to these recent developments, methods for the analysis and interpretability of neural networks in NLP are so far lacking a common framework and methodology. Moreover, approaching the analysis of modern neural networks can be difficult for newcomers to the field, since it requires both a familiarity with recent work in neural NLP and with analysis methods which are not yet standardized. This tutorial aims to fill this gap and introduce the nascent field of interpretability and analysis of neural networks in NLP.

The tutorial will cover the main lines of analysis work, mostly drawing on the recent TACL survey by Belinkov and Glass (2019).[1] In particular, we will devote a large portion to work aiming to find linguistic information that is captured by neural networks, such as probing classifiers (Hupkes et al., 2018; Adi et al., 2017; Conneau et al., 2018a,b; Tenney et al., 2019b, inter alia), controlled behavior studies on language modelling (Gulordava et al., 2018; Linzen et al., 2016a; Marvin and Linzen, 2018) or inference tasks (Poliak et al., 2018a,b; White et al., 2017; Kim et al., 2019; McCoy et al., 2019; Ross and Pavlick, 2019), psycholinguistic methods (Ettinger et al., 2018; Chrupała and Alishahi, 2019), layerwise analyses (Peters et al., 2018; Tenney et al., 2019a), among other methods (Hewitt and Manning, 2019; Zhang

[1] A comprehensive bibliography is found in the accompanying website of the survey: `https://boknilev.github.io/nlp-analysis-methods/`.



*Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1–5

https://doi.org/10.18653/v1/P17

and Bowman, 2018; Shi et al., 2016). We will also present various interactive visualization methods such as neuron activations (Karpathy et al., 2015; Dalvi et al., 2019), attention mechanisms (Bahdanau et al., 2014; Strobelt et al., 2018), and saliency measures (Li et al., 2016; Murdoch et al., 2018; Arras et al., 2017), including a walkthrough on how to build a simple attention visualization. Next, we will discuss the construction and use of challenge sets for fine-grained evaluation in the context of different tasks (Conneau and Kiela, 2018; Wang et al., 2018; Isabelle and Kuhn, 2018; Sennrich, 2017, inter alia). Finally, we will review work on generating adversarial examples in NLP, focusing on the challenges brought upon by the discrete nature of textual input (Papernot et al., 2016b; Ebrahimi et al., 2018; Jia and Liang, 2017; Belinkov and Bisk, 2018, inter alia). A detailed outline is provided in Section 3.

Throughout the tutorial, we will highlight not only the most commonly applied analysis methods, but also the specific limitations and shortcomings of current approaches. By the end of the tutorial, participants will be better informed where to focus future research efforts.

## 2 Tutorial Type

This tutorial will cover cutting-edge research in interpretability and analysis of modern neural NLP models. The topic has not been previously covered in *CL tutorials.

## 3 Outline

1. Introduction
2. Structural Analyses
   (a) Methodology: Analysis by Probing Classifiers
   (b) Example Studies: Different Components and Linguistic Phenomena
   (c) Limitations
3. Behavioral Studies
   (a) Background on Test Suites and Challenge Sets
   (b) Types of Probing Tasks
   (c) Experimental Designs
   (d) Construction Methods
   (e) Languages
4. Interaction and Visualization
   (a) How Interaction can help and its limitations
   (b) Classification and Review of Related Efforts
   (c) Demo Walk-through: Simple Attention Visualization
   (d) Broader Perspectives and Opportunities
5. Other Methods
   (a) Generating Explanations
   (b) Psycholinguistic Methods
   (c) Testing on Formal Languages
6. Conclusion

## 4 Prerequisites

We would assume acquaintance with core linguistic concepts and basic knowledge of machine learning and neural networks, such as covered in most introductory NLP courses.

## 5 Reading List

In addition to the papers mentioned in this proposal, a comprehensive bibliography can be found in the following website: `https://boknilev.github.io/nlp-analysis-methods/`.

For trainees interested in reading important studies before the tutorial, we recommend the following: Belinkov and Glass (2019); Hupkes et al. (2018); Tenney et al. (2019b); Linzen et al. (2016b); Ettinger et al. (2018); Bahdanau et al. (2014); Li et al. (2016); Sennrich (2017); Papernot et al. (2016a); Ebrahimi et al. (2018).

## 6 Names and Affiliations

**Yonatan Belinkov**, Postdoctoral Fellow, Harvard University and MIT
email: `belinkov@seas.harvard.edu`
website: `http://people.csail.mit.edu/belinkov`

Yonatan Belinkov is a Postdoctoral Fellow at the Harvard School of Engineering and Applied Sciences (SEAS) and the MIT Computer Science and Artificial Intelligence Laboratory (CSAIL). His research interests are in interpretability and robustness of neural models of language. He has done previous work in machine translation, speech recognition, community question answering, and syntactic parsing. His research has been published at ACL, EMNLP, NAACL, CL, TACL, ICLR, and NeurIPS. His PhD dissertation at MIT analyzed internal language representations in deep learning models. He co-organized or co-organizes BlackboxNLP 2019, BlackboxNLP 2020, and the WMT 2019 machine translation robustness task, and serves as an area chair for the analysis and interpretability track at ACL and EMNLP 2020.

**Sebastian Gehrmann**, Research Scientist, Google AI
email: gehrmann@google.com
website: http://sebastiangehrmann.com

Sebastian is research scientist at Google AI. He received his PhD in 2020 from Harvard University. His research focuses on the development and evaluation of controllable and interpretable models for language generation. By applying methods from human-computer interaction and visualization to problems in NLP, he develops interactive interfaces that help with the interpretation and explanation of neural networks. His research has been published at ACL, NAACL, EMNLP, CHI, and IEEE VIS. He received an honorable mention at VAST 2018 and was nominated for ACL best demo 2019 for his work on interactive visualization tools. He co-organized INLG 2019 and served as an area chair in summarization for ACL 2020.

**Ellie Pavlick**, Assistant Professor of Computer Science, Brown University
email: ellie_pavlick@brown.edu
website: http://cs.brown.edu/people/epavlick

Ellie Pavlick is an Assistant Professor at Brown University and a Research Scientist at Google. She received her PhD in 2017 with her thesis on modeling compositional lexical semantics. Her current work focuses on computational models of semantics and pragmatics, with a focus on building cognitively-plausible representations. Her recent work has focused on "probing" distributional models in order to better understand the linguistic phenomena that are and are not encoded "for free" via language modelling. Her work has been published at ACL, NAACL, EMNLP, TACL, *SEM, and ICLR, including two best paper awards at *SEM 2016 and 2019. Ellie co-organized the 2018 JSALT summer workshop on building and evaluating general-purpose sentence representations. She also served as area chair for ACL's sentence-level semantics track.

## References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained Analysis of Sentence Embeddings Using Auxiliary Prediction Tasks. In *International Conference on Learning Representations (ICLR)*.

Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 2017. "What is relevant in a text document?": An interpretable machine learning approach. *PLOS ONE*, 12(8):1–23.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv preprint arXiv:1409.0473v7*.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and Natural Noise Both Break Neural Machine Translation. In *International Conference on Learning Representations (ICLR)*.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Grzegorz Chrupała and Afra Alishahi. 2019. Correlating neural and symbolic representations of language. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2952–2962, Florence, Italy. Association for Computational Linguistics.

Alexis Conneau and Douwe Kiela. 2018. SentEval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan. European Languages Resources Association (ELRA).

Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018a. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136. Association for Computational Linguistics.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018b. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of* *the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Fahim Dalvi, Avery Nortonsmith, D. Anthony Bau, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, and James Glass. 2019. NeuroX: A Toolkit for Analyzing Individual Neurons in Neural Networks. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI): Demonstrations Track*.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. HotFlip: White-Box Adversarial Examples for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 31–36. Association for Computational Linguistics.

Allyson Ettinger, Ahmed Elgohary, Colin Phillips, and Philip Resnik. 2018. Assessing composition in sentence vector representations. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1790–1801, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. 2018. Visualisation and 'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *Journal of Artificial Intelligence Research*, 61:907–926.

Pierre Isabelle and Roland Kuhn. 2018. A Challenge Set for French–¿ English Machine Translation. *arXiv preprint arXiv:1806.02725v2*.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031. Association for Computational Linguistics.

Andrej Karpathy, Justin Johnson, and Fei-Fei Li. 2015. Visualizing and Understanding Recurrent Networks. *arXiv preprint arXiv:1506.02078v2*.

Najoung Kim, Roma Patel, Adam Poliak, Patrick Xia, Alex Wang, Tom McCoy, Ian Tenney, Alexis Ross, Tal Linzen, Benjamin Van Durme, Samuel R. Bowman, and Ellie Pavlick. 2019. Probing what different NLP tasks teach machines about function word comprehension. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 235–249, Minneapolis, Minnesota. Association for Computational Linguistics.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. Understanding Neural Networks through Representation Erasure. *arXiv preprint arXiv:1612.08220v3*.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016a. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016b. Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

W. James Murdoch, Peter J. Liu, and Bin Yu. 2018. Beyond Word Importance: Contextual Decomposition to Extract Interactions from LSTMs. In *International Conference on Learning Representations*.

Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. 2016a. Transferability in Machine Learning: From Phenomena to Black-Box Attacks using Adversarial Samples. *arXiv preprint arXiv:1605.07277v1*.

Nicolas Papernot, Patrick McDaniel, Ananthram Swami, and Richard Harang. 2016b. Crafting Adversarial Input Sequences for Recurrent Neural Networks. In *Military Communications Conference (MILCOM)*, pages 49–54. IEEE.

Matthew Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018. Dissecting contextual word embeddings: Architecture and representation.

In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509, Brussels, Belgium. Association for Computational Linguistics.

Adam Poliak, Yonatan Belinkov, James Glass, and Benjamin Van Durme. 2018a. On the evaluation of semantic phenomena in neural machine translation using natural language inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 513–523, New Orleans, Louisiana. Association for Computational Linguistics.

Adam Poliak, Aparajita Haldar, Rachel Rudinger, J. Edward Hu, Ellie Pavlick, Aaron Steven White, and Benjamin Van Durme. 2018b. Collecting diverse natural language inference problems for sentence representation evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 67–81, Brussels, Belgium. Association for Computational Linguistics.

Alexis Ross and Ellie Pavlick. 2019. How well do nli models capture verb veridicality? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Hong Kong, China. Association for Computational Linguistics.

Rico Sennrich. 2017. How Grammatical is Character-level Neural Machine Translation? Assessing MT Quality with Contrastive Translation Pairs. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 376–382. Association for Computational Linguistics.

Xing Shi, Inkit Padhi, and Kevin Knight. 2016. Does string-based neural MT learn source syntax? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1526–1534, Austin, Texas. Association for Computational Linguistics.

Hendrik Strobelt, Sebastian Gehrmann, Michael Behrisch, Adam Perer, Hanspeter Pfister, and Alexander M Rush. 2018. Seq2seq-vis: A visual debugging tool for sequence-to-sequence models. *IEEE transactions on visualization and computer graphics*, 25(1):353–363.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019a. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. 2019b. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Alex Wang, Amapreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. *arXiv preprint arXiv:1804.07461v1*.

Aaron Steven White, Pushpendre Rastogi, Kevin Duh, and Benjamin Van Durme. 2017. Inference is everything: Recasting semantic resources into a unified evaluation framework. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 996–1005, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Kelly Zhang and Samuel Bowman. 2018. Language modeling teaches you more than translation does: Lessons learned through auxiliary syntactic task analysis. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 359–361, Brussels, Belgium. Association for Computational Linguistics.