Explaining Neural Language Models from Internal Representations to Model Predictions

Lectures on Computational Linguistics 2023, May 31, Pisa

Alessio Miaschi
ItaliaNLP Lab, Istituto di Linguistica Computazionale (CNR-ILC), Pisa
alessio.miaschi@ilc.cnr.it
https://alemiaschi.github.io/

Gabriele Sarti
GroNLP, Center for Language and Cognition, University of Groningen
g.sarti@rug.nl
https://gsarti.com/

Explaining Neural Language Models from Internal Representations to Model Predictions

Part I:

- Assessing Transformer Models Syntactic Abilities
- Probing Transformer Internal Representations

Part II:

- Feature Attribution for NLP:
  - Attributing Classification Models with **ferret**
  - Attributing Generative Language Models with **Inseq**

Materials

Github Repository: https://github.com/gsarti/lcl23-xnlm-lab

Explaining Neural Language Models from Internal Representations to Model Predictions

Part I

About me and...

I am a PostDoc at the ItaliaNLP Lab, Institute for Computational Linguistics “A. Zampolli” (CNR-ILC, Pisa). In 2022, I received my PhD in Computer Science at the University of Pisa.

My research interests lie primarily in the context of Natural Language Processing. I am particularly interested in the analysis and the definition of methods for inferring and evaluating representations from data, as well as in the development of NLP tools for building educational applications.

About me and... the team!

I am a PostDoc at the ItaliaNLP Lab, Institute for Computational Linguistics "A. Zampolli" (CNR-ILC, Pisa). In 2022, I received my PhD in Computer Science at the University of Pisa.

My research interests lie primarily in the context of Natural Language Processing. I am particularly interested in the analysis and the definition of methods for inferring and evaluating representations from data, as well as in the development of NLP tools for building educational applications.

The **ItaliaNLP Lab** (**CNR-ILC**) gathers researchers, postdocs and students from computational linguistics, computer science and linguistics who work on developing resources and algorithms for processing and understanding human languages.

Permanent Researchers:

- Felice Dell'Orletta
- Simonetta Montemagni
- Dominique Brunato
- Franco Alberto Cardillo
- Giulia Venturi

Postdocs:

- Chiara Alzetta
- Alessio Miaschi
- Andrea Amelio Ravelli

PhD Students:

- Luca Dini
- Benedetta Iavarone
- Giovanni Puccetti

Research Fellows:

- Chiara Fazzone

Affiliated Researchers:

- Luca Bacco
- Mario Merone

Master/Undergraduate/Visiting Students

Link to website: http://www.italianlp.it/

Outline

- Introduction
- Lab 1.1: Assessing Transformer Models Syntactic Abilities
- Lab 1.2: Probing Transformer Internal Representations

Interpretability in NLP

- The analysis of the inner workings of NLMs has become one of the most addressed line of research in NLP

- Several methods have been implemented to obtain meaningful explanations and to understand how these models are able to capture syntax- and semantic- sensitive phenomena

Interpretability in NLP

- The analysis of the inner workings of NLMs has become one of the most addressed line of research in NLP

- Several methods have been implemented to obtain meaningful explanations and to understand how these models are able to capture syntax- and semantic- sensitive phenomena

- Several approaches:
    - Behavioural tests (e.g. Goldberg, 2019, Warstadt et al., 2020);
    - Probing tasks (e.g. Hewitt and Manning, 2019; Pimentel et al., 2020);
    - Analysis of attention mechanisms (e.g. Clark et al., 2019);
    - Feature Attribution methods (e.g. Ramnath, 2020);

Interpretability in NLP

- The analysis of the inner workings of NLMs has become one of the most addressed line of research in NLP

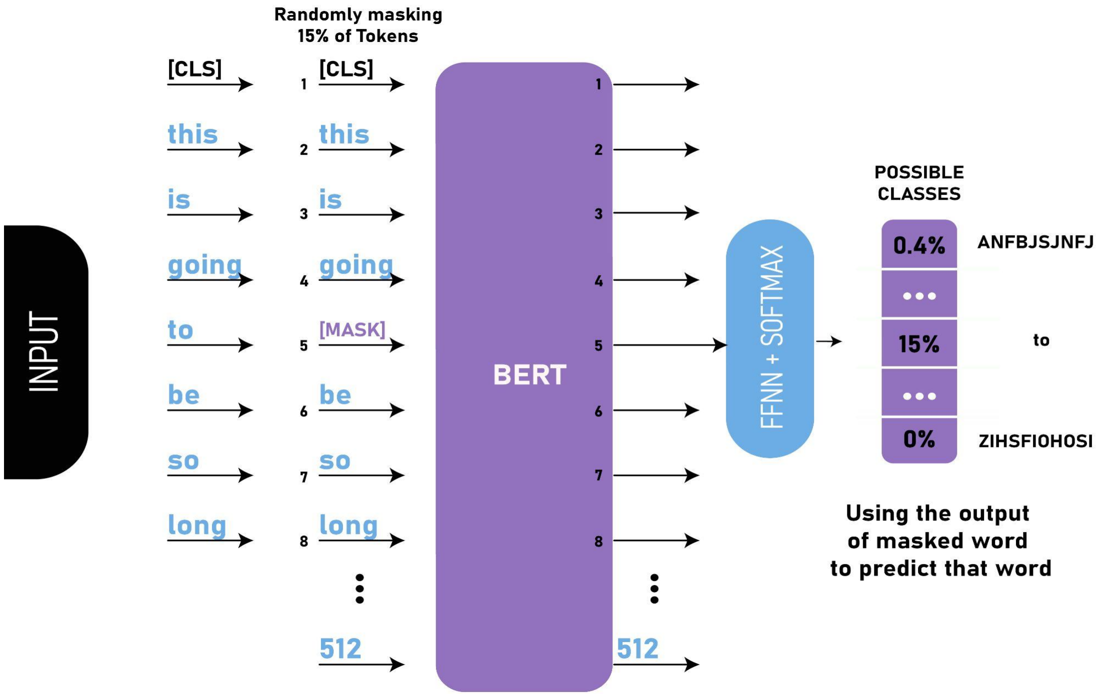
- Several methods have been implemented to obtain meaningful explanations and to understand how these models are able to capture syntax- and semantic- sensitive phenomena

- Several approaches:
    - **Behavioural tests (e.g. Goldberg, 2019, Warstadt et al., 2020);**
    - **Probing tasks (e.g. Hewitt and Manning, 2019; Pimentel et al., 2020);**
    - Analysis of attention mechanisms (e.g. Clark et al., 2019);
    - Feature Attribution methods (e.g. Ramnath, 2020);

Lab 1.1

Assessing Transformer Models Syntactic Abilities

Masked Language Modeling (MLM)



Source: https://www.geeksforgeeks.org/understanding-bert-nlp/

Assessing BERT’s Syntactic Abilities (Goldberg, 2019)

- Goldberg (2019) proposes a methodology for testing the implicit linguistic competence of BERT

- Specifically, two linguistic phenomena are considered:
    - Subject-Verb Agreement;
    - Reflexive Anaphora.

- **Approach**: masking target words and asking the model to “fill in the gap” with the words with high probability scores

Assessing BERT's Syntactic Abilities (Goldberg, 2019)

- Goldberg (2019) proposes a methodology for testing the implicit linguistic competence of BERT

- Specifically, two linguistic phenomena are considered:
  - Subject-Verb Agreement;
  - Reflexive Anaphora.
- **Approach**: masking target words and asking the model to "fill in the gap" with the words with high probability scores

Assessing BERT's Syntactic Abilities (Goldberg, 2019)

the game that the guard hates is bad

Assessing BERT's Syntactic Abilities (Goldberg, 2019)

the game that the guard hates **[MASK]** bad

Assessing BERT's Syntactic Abilities (Goldberg, 2019)

the game that the guard hates **[MASK]** bad

- *p(is)* = ?
- *p(are)* = ?

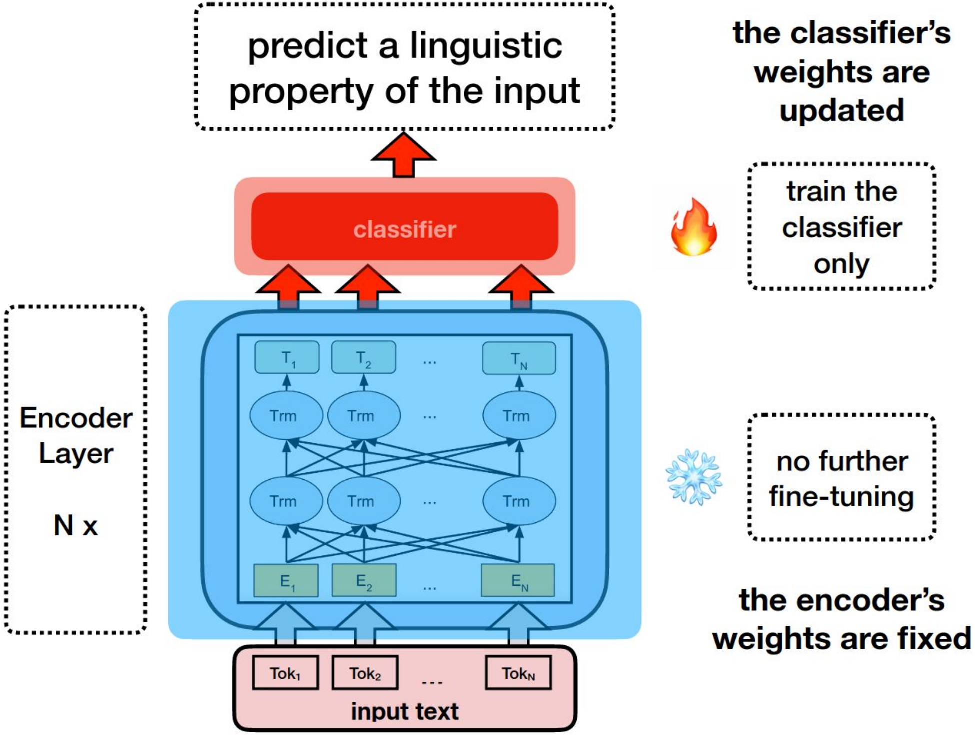
Assessing MLM Transformer Model Syntactic Abilities

https://colab.research.google.com/drive/1Z6G_DJNyYXAA8KMnOzle7OmC_nS-N_p5?usp=sharing

Lab 1.2

Probing Transformer Internal Representations

Probing Task Approach



predict a linguistic property of the input

the classifier's weights are updated

🔥 train the classifier only

Encoder Layer

N x

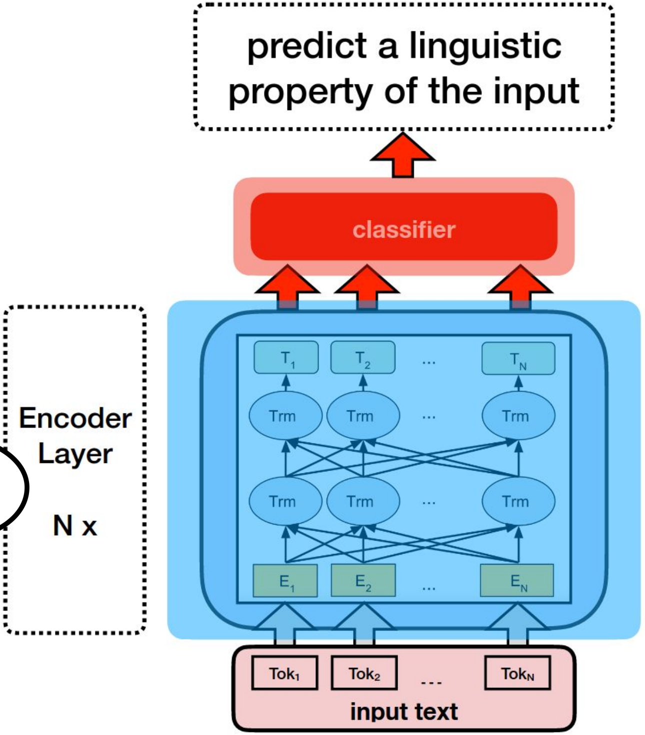
❄️ no further fine-tuning

the encoder's weights are fixed

input text

Source: https://people.cs.umass.edu/~miyyer/cs685_f20/slides/19-probes.pdf

Probing Task Approach



predict a linguistic property of the input

classifier

Encoder Layer

N x

input text

the classifier's weights are updated

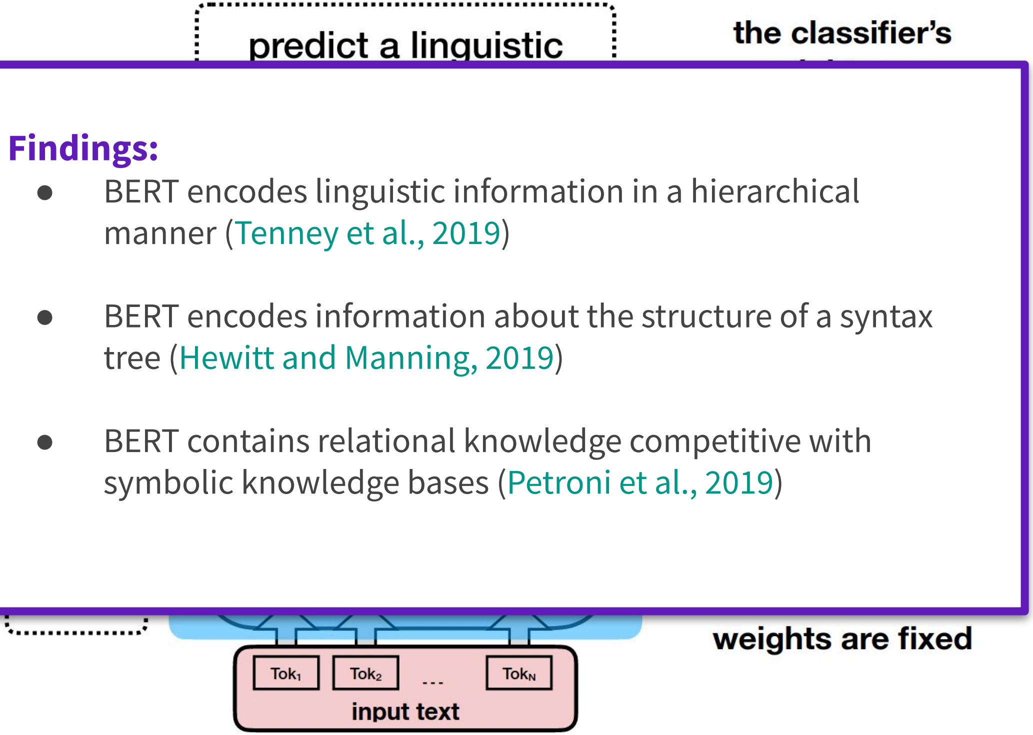
🔥 train the classifier only

❄️ no further fine-tuning

the encoder's weights are fixed

Source: https://people.cs.umass.edu/~miyyer/cs685_f20/slides/19-probes.pdf

Probing Task Approach



Findings:

- BERT encodes linguistic information in a hierarchical manner (Tenney et al., 2019)
- BERT encodes information about the structure of a syntax tree (Hewitt and Manning, 2019)
- BERT contains relational knowledge competitive with symbolic knowledge bases (Petroni et al., 2019)

Source: https://people.cs.umass.edu/~miyyer/cs685_f20/slides/19-probes.pdf

Profiling Neural Language Models

- The “*linguistic profiling*” methodology (van Halteren, 2004) assumes that wide counts of linguistic features are particularly helpful in the resolution of several NLP tasks, e.g.:
    - Text Profiling (e.g. text readability, textual genres)
    - Author Profiling (e.g. author’s age and native language)

Research Question:

Could the informative power of these features also be helpful to understand the behaviour of state-of-the-art NLMs?

Linguistic Profiling of a Neural Language Model (Miaschi et al., 2020)

- We investigated the linguistic knowledge implicitly encoded by BERT

Research questions:

1. What kind of linguistic properties are encoded in a pre-trained version of BERT?
2. How this knowledge is modified after a fine-tuning process
3. Whether this implicit knowledge affects the ability of the model to solve a specific downstream task

Linguistic Profiling of a Neural Language Model (Miaschi et al., 2020)

- We investigated the linguistic knowledge implicitly encoded by BERT

Research questions:

1. **What kind of linguistic properties are encoded in a pre-trained version of BERT?**

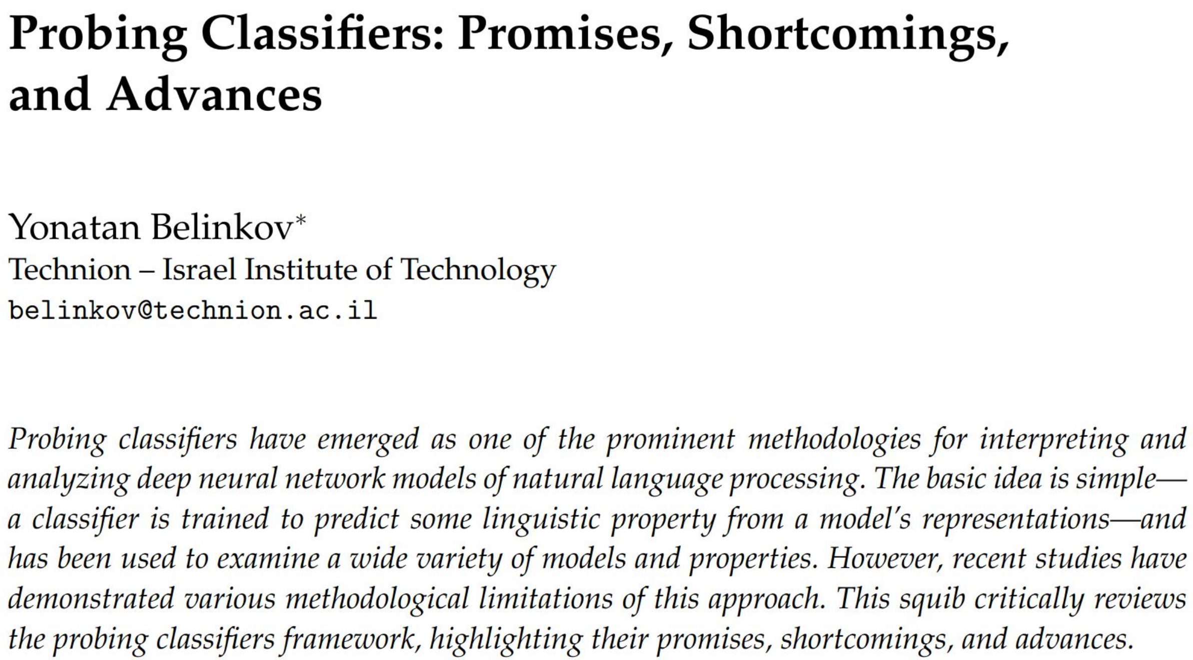
2. How this knowledge is modified after a fine-tuning process

3. Whether this implicit knowledge affects the ability of the model to solve a specific downstream task

Probing Transformer Internal Representations

https://colab.research.google.com/drive/1cYDaO9ymOZ58t2nhjYRaZqCBqHtp4dM7?usp=sharing

Probing Classifiers Framework



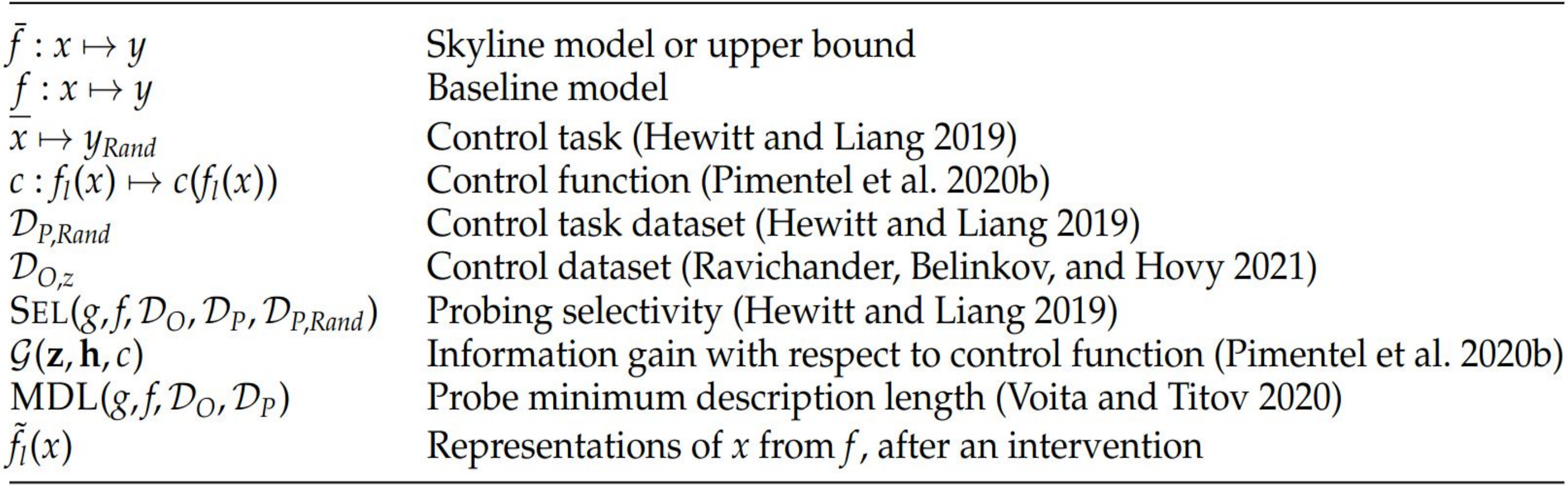
Probing Classifiers: Promises, Shortcomings, and Advances

Yonatan Belinkov*
Technion – Israel Institute of Technology
belinkov@technion.ac.il

Probing classifiers have emerged as one of the prominent methodologies for interpreting and analyzing deep neural network models of natural language processing. The basic idea is simple—a classifier is trained to predict some linguistic property from a model's representations—and has been used to examine a wide variety of models and properties. However, recent studies have demonstrated various methodological limitations of this approach. This squib critically reviews the probing classifiers framework, highlighting their promises, shortcomings, and advances.

Yonatan Belinkov. 2022. Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics*, 48(1):207–219.

Probing Classifiers Framework



$\bar{f} : x \mapsto y$	Skyline model or upper bound
$\underline{f} : x \mapsto y$	Baseline model
$x \mapsto y_{Rand}$	Control task (Hewitt and Liang 2019)
$c : f_l(x) \mapsto c(f_l(x))$	Control function (Pimentel et al. 2020b)
$\mathcal{D}_{P,Rand}$	Control task dataset (Hewitt and Liang 2019)
$\mathcal{D}_{O,z}$	Control dataset (Ravichander, Belinkov, and Hovy 2021)
$\textsc{Sel}(g, f, \mathcal{D}_O, \mathcal{D}_P, \mathcal{D}_{P,Rand})$	Probing selectivity (Hewitt and Liang 2019)
$\mathcal{G}(\mathbf{z}, \mathbf{h}, c)$	Information gain with respect to control function (Pimentel et al. 2020b)
$\mathrm{MDL}(g, f, \mathcal{D}_O, \mathcal{D}_P)$	Probe minimum description length (Voita and Titov 2020)
$\tilde{f}_l(x)$	Representations of x from f, after an intervention

Yonatan Belinkov. 2022. Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics*, 48(1):207–219.

university of
groningen

Thanks for the attention!

🌐 https://alemiaschi.github.io/

🐦 @AlessioMiaschi

🌐 http://www.italianlp.it/

🐦 @ItaliaNLP_Lab

References

- Devlin, Jacob, et al. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *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)*.
- Goldberg, Yoav (2019). "Assessing BERT's syntactic abilities." *arXiv preprint arXiv:1901.05287*.
- Warstadt, Alex, et al. (2020). “BLiMP: The Benchmark of Linguistic Minimal Pairs for English”. *Transactions of the Association for Computational Linguistics* 2020; 8 377–392.
- Hewitt, John, and Christopher D. Manning (2019). "A structural probe for finding syntax in word representations." *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)*.
- Clark, Kevin, et al. (2019) "What Does BERT Look at? An Analysis of BERT’s Attention." *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*.
- Ramnath, Sahana, et al. (2020). Towards Interpreting BERT for Reading Comprehension Based QA. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (EMNLP), pages 3236–3242, Online. Association for Computational Linguistics.
- Pimentel, Tiago et al. (2020). “Information-Theoretic Probing for Linguistic Structure”. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4609–4622, Online. Association for Computational Linguistics.
- Tenney, Ian et al. (2019). “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.

References

- Petroni, Fabio et al. (2019). “Language Models as Knowledge Bases?”. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.
- van Halteren, Hans (2004). “Linguistic Profiling for Authorship Recognition and Verification”. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 199–206, Barcelona, Spain.
- Miaschi, Alessio, et al. (2020) "Linguistic Profiling of a Neural Language Model." *Proceedings of the 28th International Conference on Computational Linguistics*.
- Belinkov, Y. (2022). “Probing Classifiers: Promises, Shortcomings, and Advances”. *Computational Linguistics*, 48(1):207–219.