Opening Large Language Models

DCP23, 09/06/2023

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

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

PhD Students:

- Luca Dini
- Benedetta Iavarone
- Giovanni Puccetti

Research Fellows:

- Chiara Fazzone

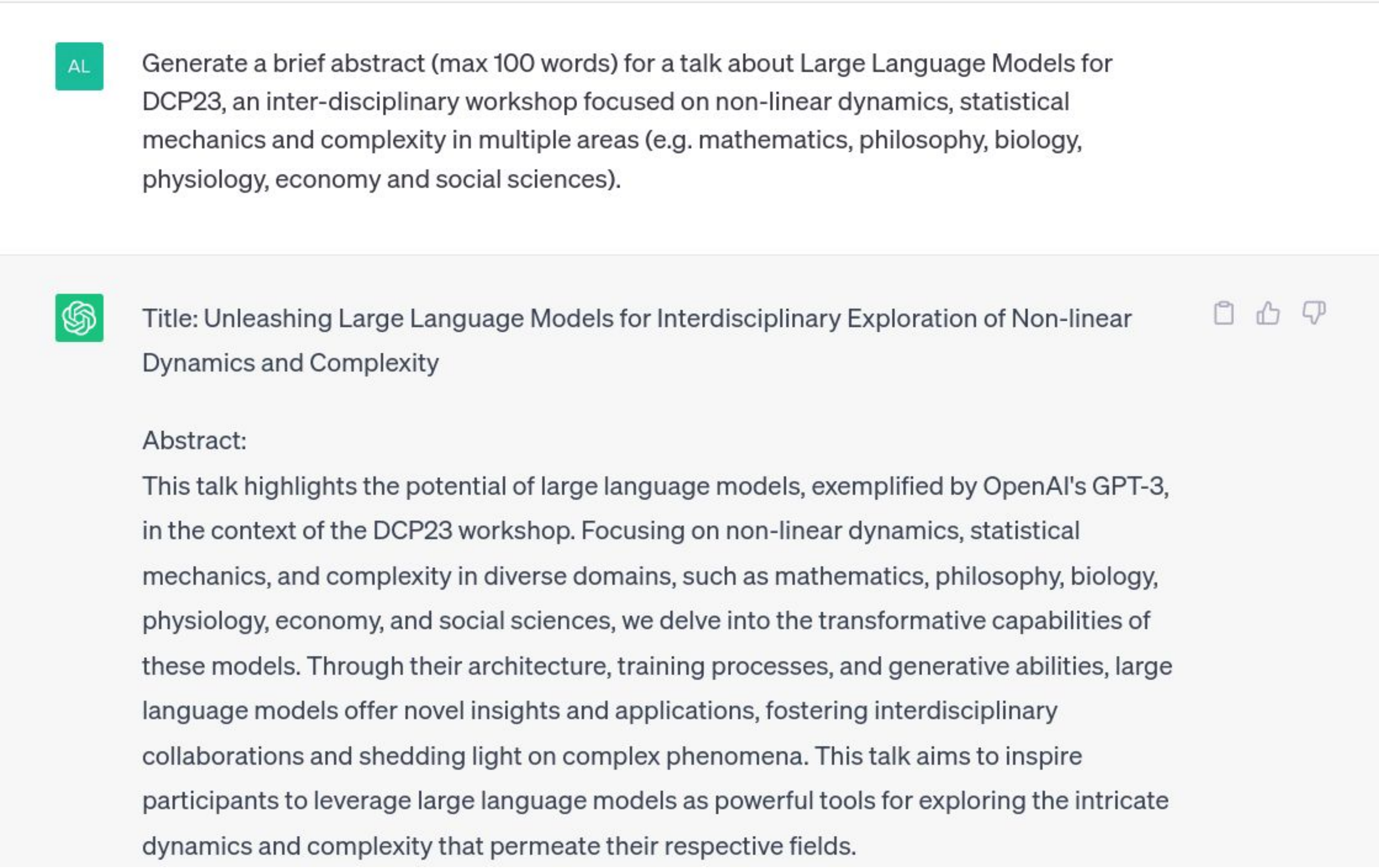
Affiliated Researchers:

- Luca Bacco
- Mario Merone

Master/Undergraduate/Visiting Students

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

Introduction



Generate a brief abstract (max 100 words) for a talk about Large Language Models for DCP23, an inter-disciplinary workshop focused on non-linear dynamics, statistical mechanics and complexity in multiple areas (e.g. mathematics, philosophy, biology, physiology, economy and social sciences).

Title: Unleashing Large Language Models for Interdisciplinary Exploration of Non-linear Dynamics and Complexity

Abstract:
This talk highlights the potential of large language models, exemplified by OpenAI's GPT-3, in the context of the DCP23 workshop. Focusing on non-linear dynamics, statistical mechanics, and complexity in diverse domains, such as mathematics, philosophy, biology, physiology, economy, and social sciences, we delve into the transformative capabilities of these models. Through their architecture, training processes, and generative abilities, large language models offer novel insights and applications, fostering interdisciplinary collaborations and shedding light on complex phenomena. This talk aims to inspire participants to leverage large language models as powerful tools for exploring the intricate dynamics and complexity that permeate their respective fields.

Neural Language Model (NLM)

- Neural Language Model → Neural Network trained to approximate the **language modeling** function

Neural Language Model (NLM)

- Neural Language Model → Neural Network trained to approximate the **language modeling** function

- **Language Modeling** → probability of a sentence s = *[w_1, w_2, …, w_n]* as:

$$P(s) = \prod_{i=1}^{N} P(w_i | w_1, w_2, ..., w_{i-1})$$

Neural Language Model (NLM)

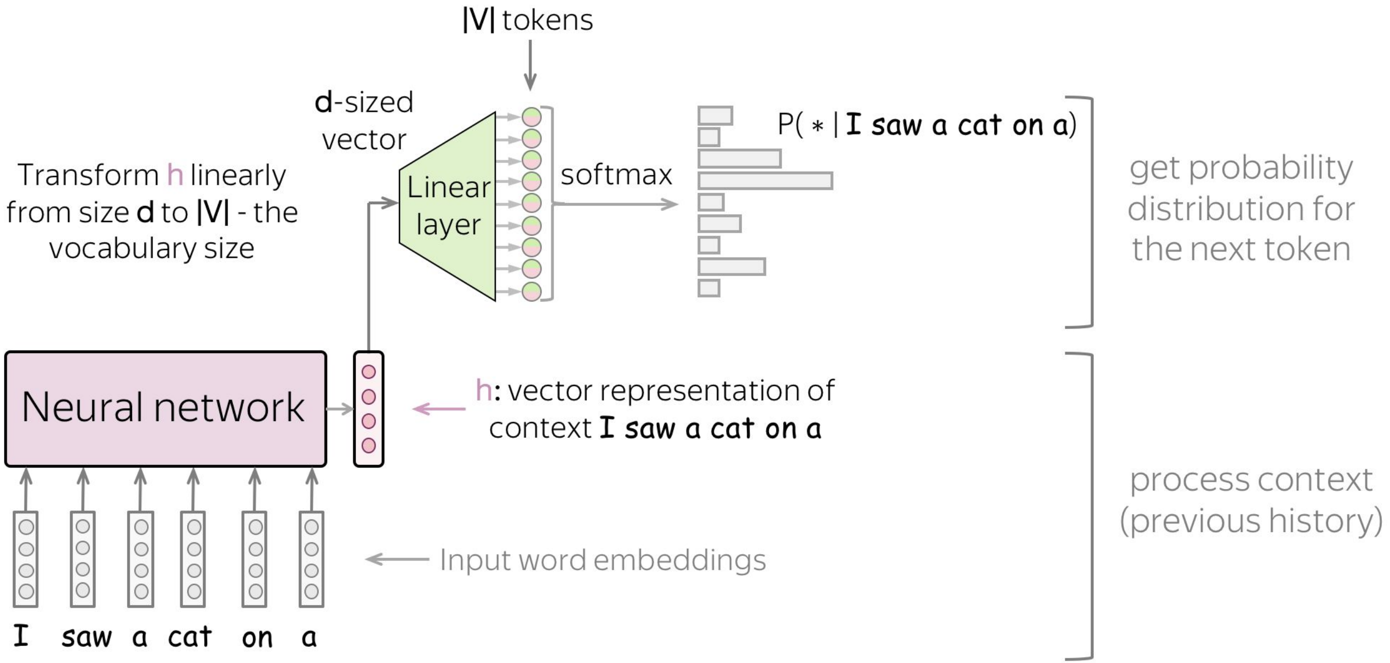
- Neural Language Model → Neural Network trained to approximate the **language modeling** function

- **Language Modeling** → probability of a sentence s = $[w_1, w_2, \ldots, w_n]$ as:

$$P(s) = \prod_{i=1}^{N} P(w_i | w_1, w_2, ..., w_{i-1})$$

- Bengio et al. (2003) proposed a model to learn this function relying on the architecture of a neural network → **Neural Probabilistic Language Model**

Neural Language Model (NLM)



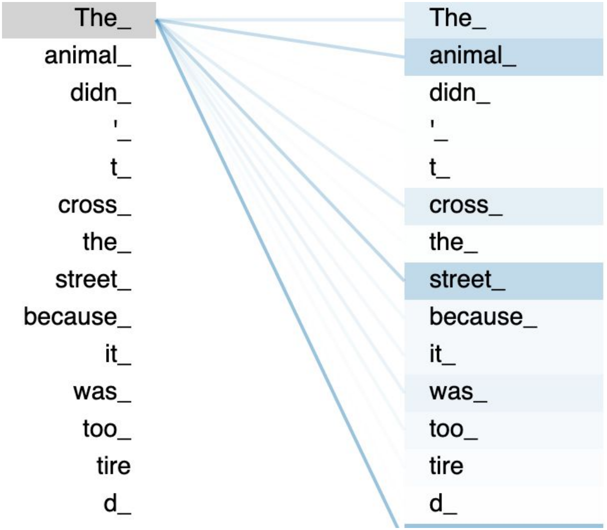
Source: https://lena-voita.github.io/nlp_course/language_modeling.html

Transformer Models

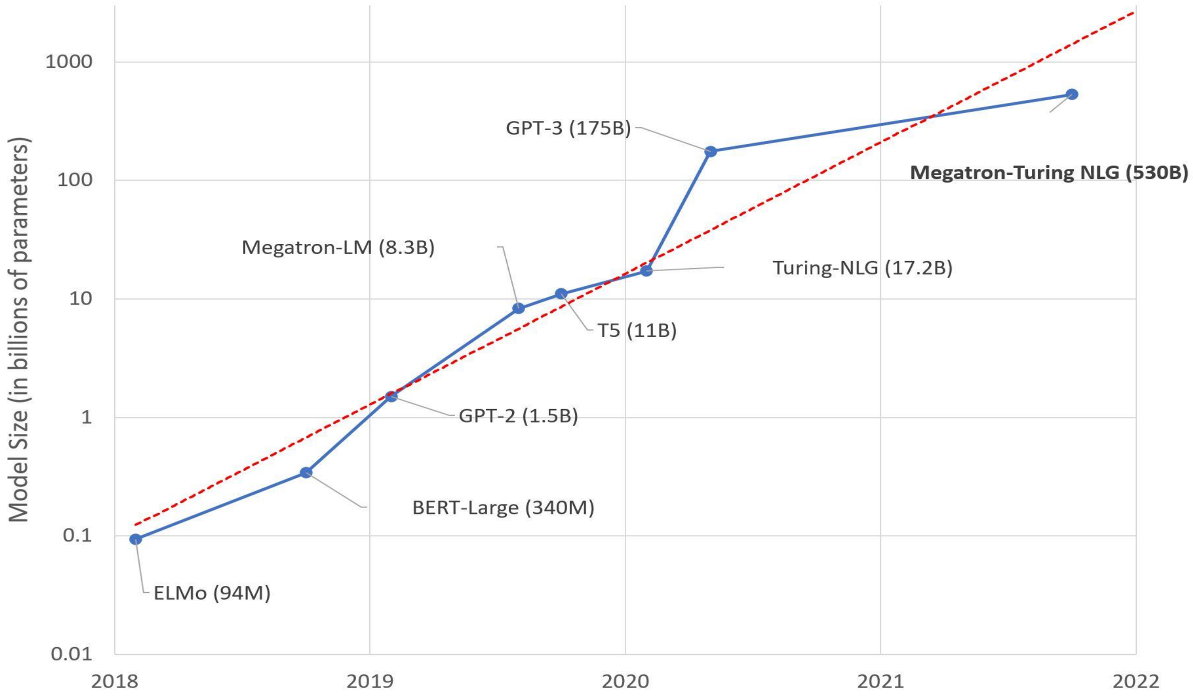
- Nowadays, the Transformer is the most commonly used architecture for the development of NLMs

- The Transformer (Vaswani et al., 2017) exploits the **attention mechanism** to create contextual representations of words and learn the relations among them

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

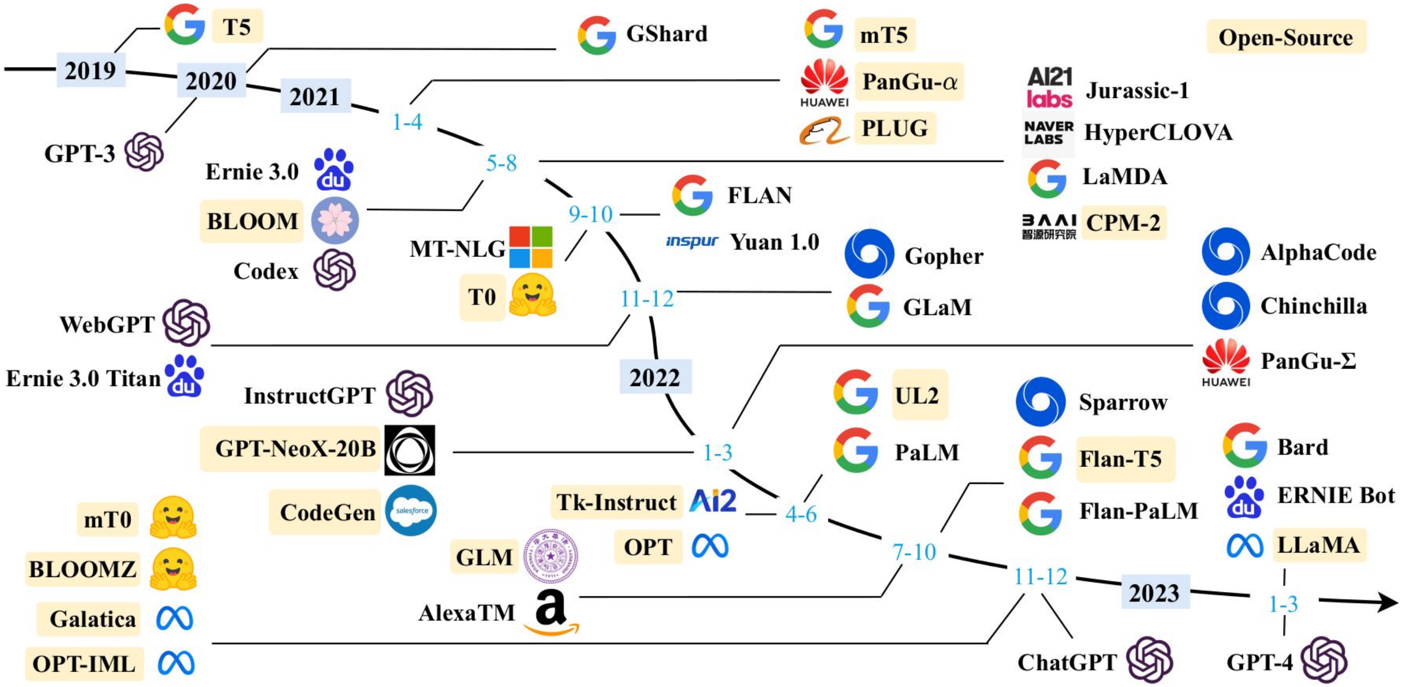


Parameters are all you need (?)



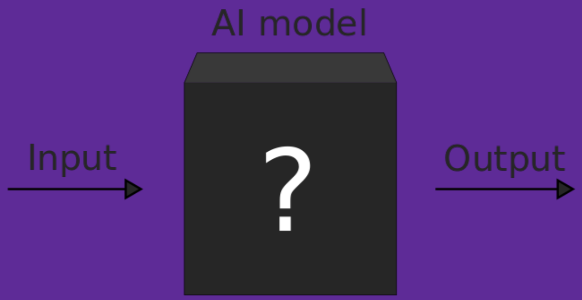
Source: https://huggingface.co/blog/large-language-models

Large Language Models (LLMs)



Source: https://huggingface.co/blog/large-language-models

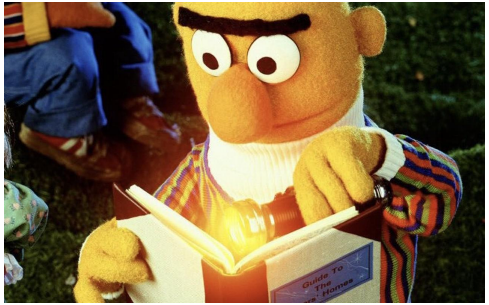
Interpreting Neural Language Models



Interpretability in NLP

> *“In the context of NLP, this question needs to be understood in light of earlier NLP work. [...] In some of these systems, features are more easily understood by humans. [...] In contrast, it is more difficult to understand what happens in an end-to-end neural network model that takes input (say, word embeddings) and generates an output.”*

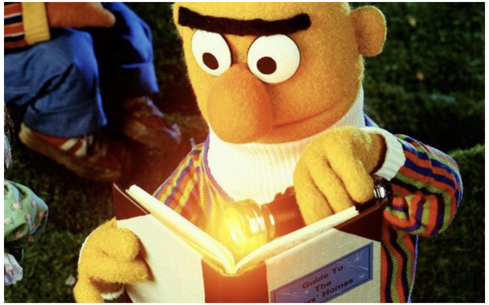
Belinkov and Glass, Analysis Methods in Neural Language Processing: A Survey (2019). In Transactions of ACL, Volume 7, pages 49-72.



Interpretability in NLP

> *"In the context of NLP, this question needs to be understood in light of earlier NLP work. [...] In some of these systems, features are more easily understood by humans. [...] In contrast, it is more difficult to understand what happens in an end-to-end neural network model that takes input (say, word embeddings) and generates an output."*

Belinkov and Glass, Analysis Methods in Neural Language Processing: A Survey (2019). In Transactions of ACL, Volume 7, pages 49-72.



Research questions:

- What happens in an end-to-end neural network model when trained on a language modeling task?
- What kind of linguistic knowledge (i.e. features) is encoded within their representations?
- Is there a relationship between the linguistic knowledge implicitly encoded and the ability to solve a specific task?

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)
    - Probing tasks (e.g. Hewitt and Manning, 2019; Pimentel et al., 2020);
    - Analysis of attention mechanisms (e.g. Clark et al., 2019);
    - Explainability via Integrated Gradients (e.g. Ramnath, 2020);

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.

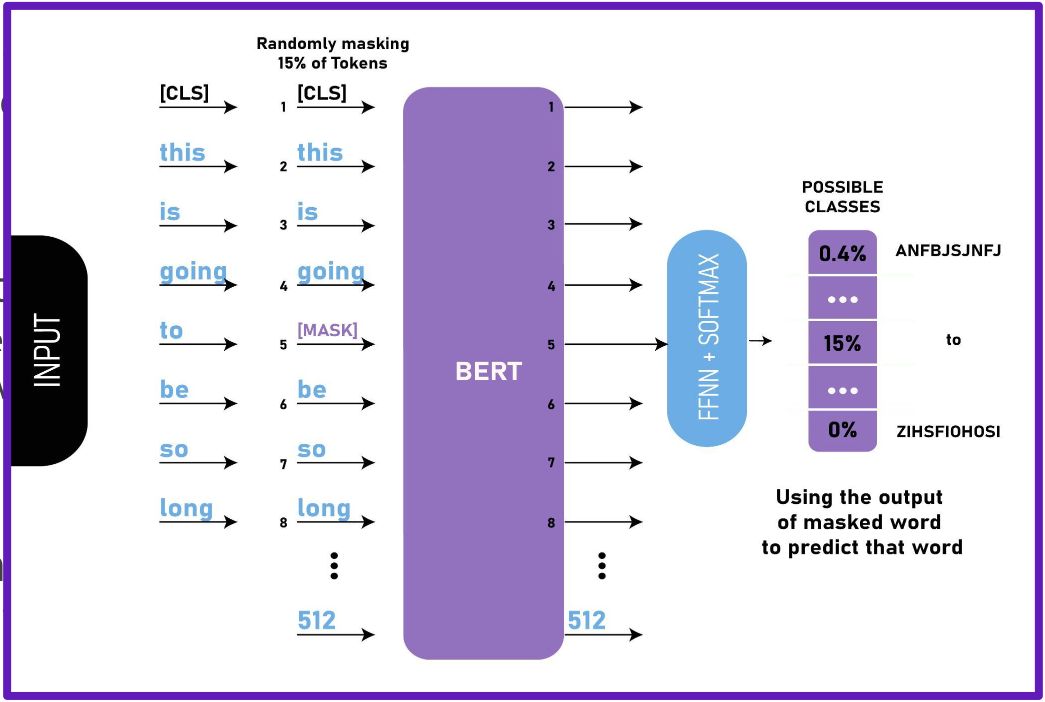
- **Approach**: masking target words and asking the model to "fill in the gap" with the words with high probability scores

Assessing BERT’s Syntactic Abilities (Goldberg, 2019)

- Goldberg (20 inguistic competence

- Specifically, t
  - Subject-Ve
  - Reflexive A

- **Approach**: m he gap” with the words with h



Assessing BERT's Syntactic Abilities (Goldberg, 2019)

the game that the guard hates is bad

Assessing BERT’s Syntactic Abilities (Goldberg, 2019)

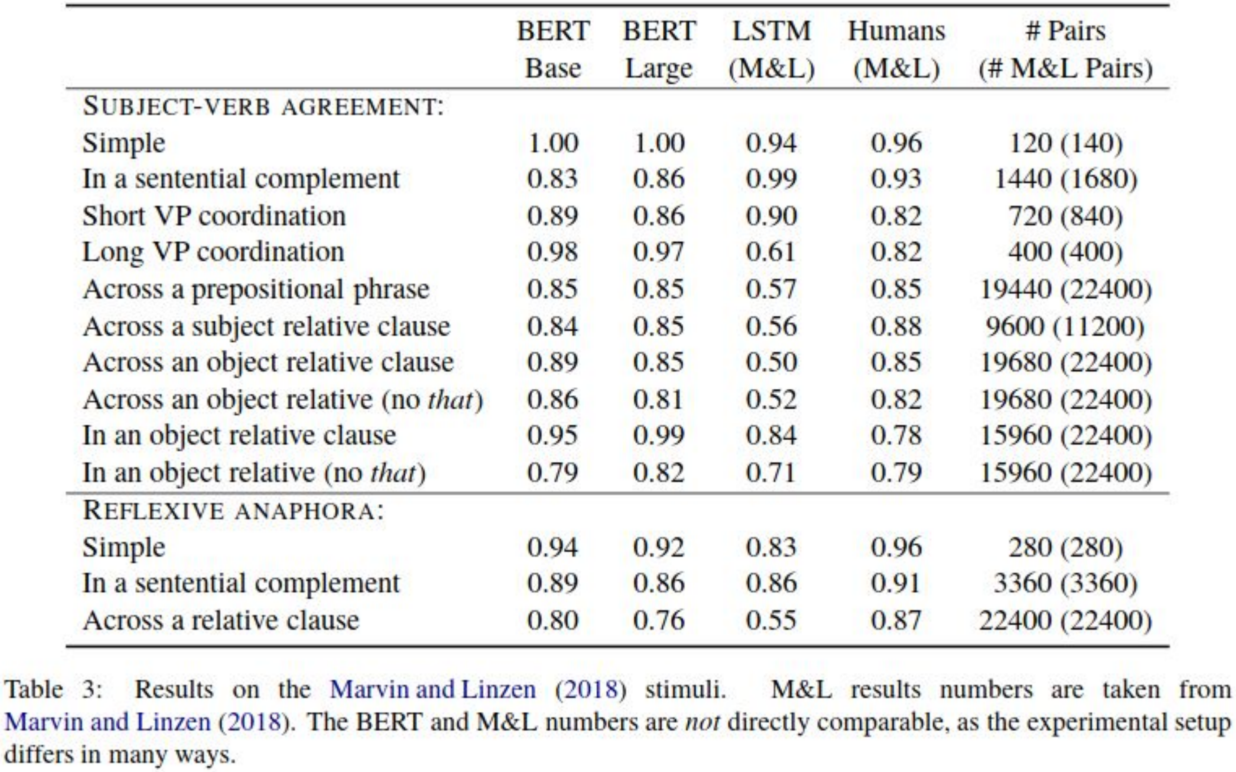
the game that the guard hates **[MASK]** bad

Assessing BERT’s Syntactic Abilities (Goldberg, 2019)

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

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

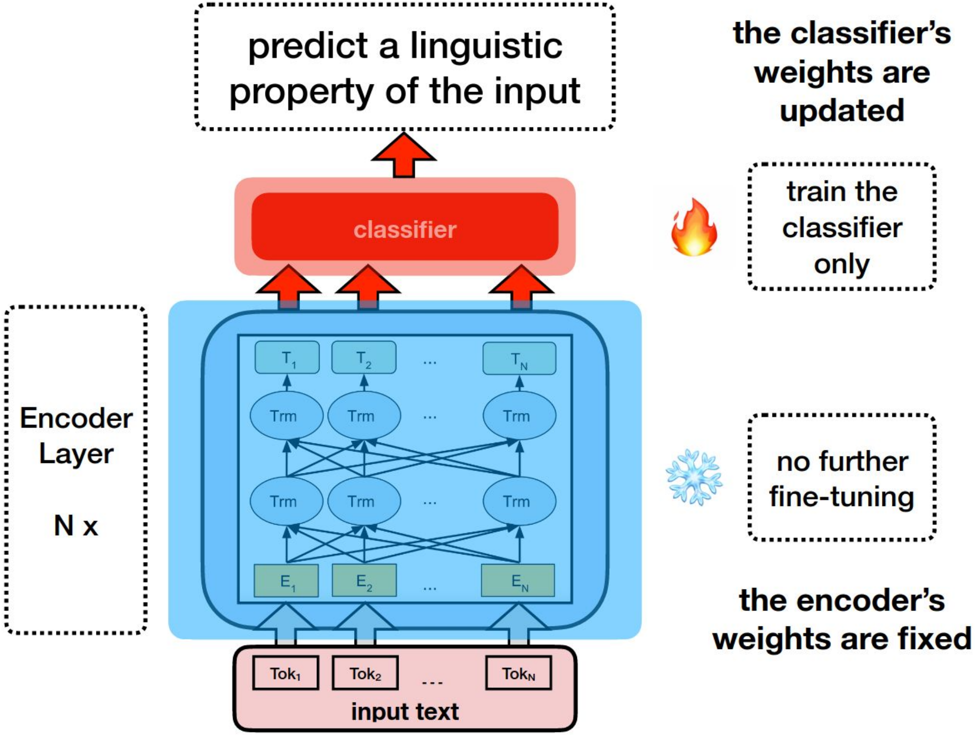
Assessing BERT's Syntactic Abilities (Goldberg, 2019)



	BERT Base	BERT Large	LSTM (M&L)	Humans (M&L)	# Pairs (# M&L Pairs)
SUBJECT-VERB AGREEMENT:					
Simple	1.00	1.00	0.94	0.96	120 (140)
In a sentential complement	0.83	0.86	0.99	0.93	1440 (1680)
Short VP coordination	0.89	0.86	0.90	0.82	720 (840)
Long VP coordination	0.98	0.97	0.61	0.82	400 (400)
Across a prepositional phrase	0.85	0.85	0.57	0.85	19440 (22400)
Across a subject relative clause	0.84	0.85	0.56	0.88	9600 (11200)
Across an object relative clause	0.89	0.85	0.50	0.85	19680 (22400)
Across an object relative (no *that*)	0.86	0.81	0.52	0.82	19680 (22400)
In an object relative clause	0.95	0.99	0.84	0.78	15960 (22400)
In an object relative (no *that*)	0.79	0.82	0.71	0.79	15960 (22400)
REFLEXIVE ANAPHORA:					
Simple	0.94	0.92	0.83	0.96	280 (280)
In a sentential complement	0.89	0.86	0.86	0.91	3360 (3360)
Across a relative clause	0.80	0.76	0.55	0.87	22400 (22400)

Table 3: Results on the Marvin and Linzen (2018) stimuli. M&L results numbers are taken from Marvin and Linzen (2018). The BERT and M&L numbers are *not* directly comparable, as the experimental setup differs in many ways.

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

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)

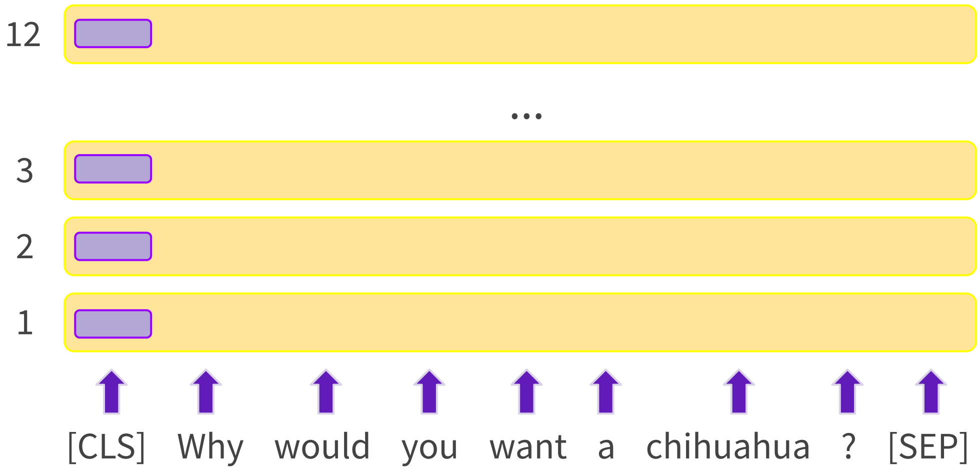
Profiling Neural Language Models

- 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)

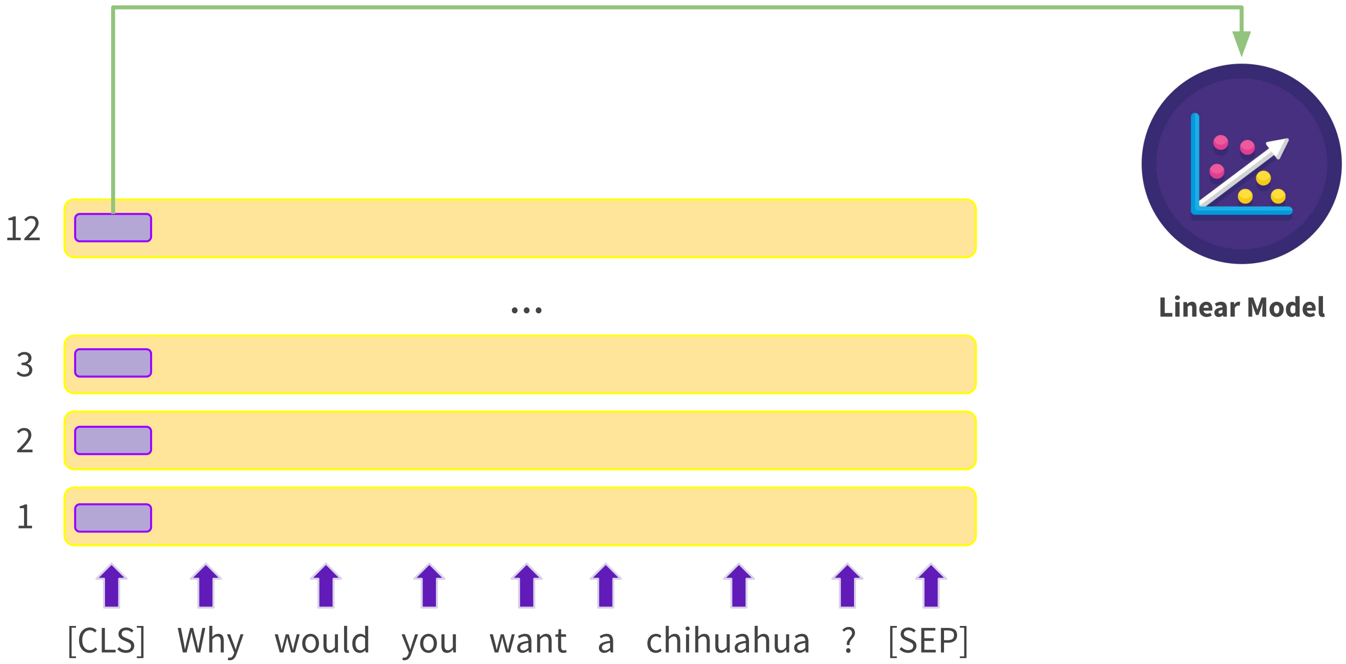
Research Question:

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

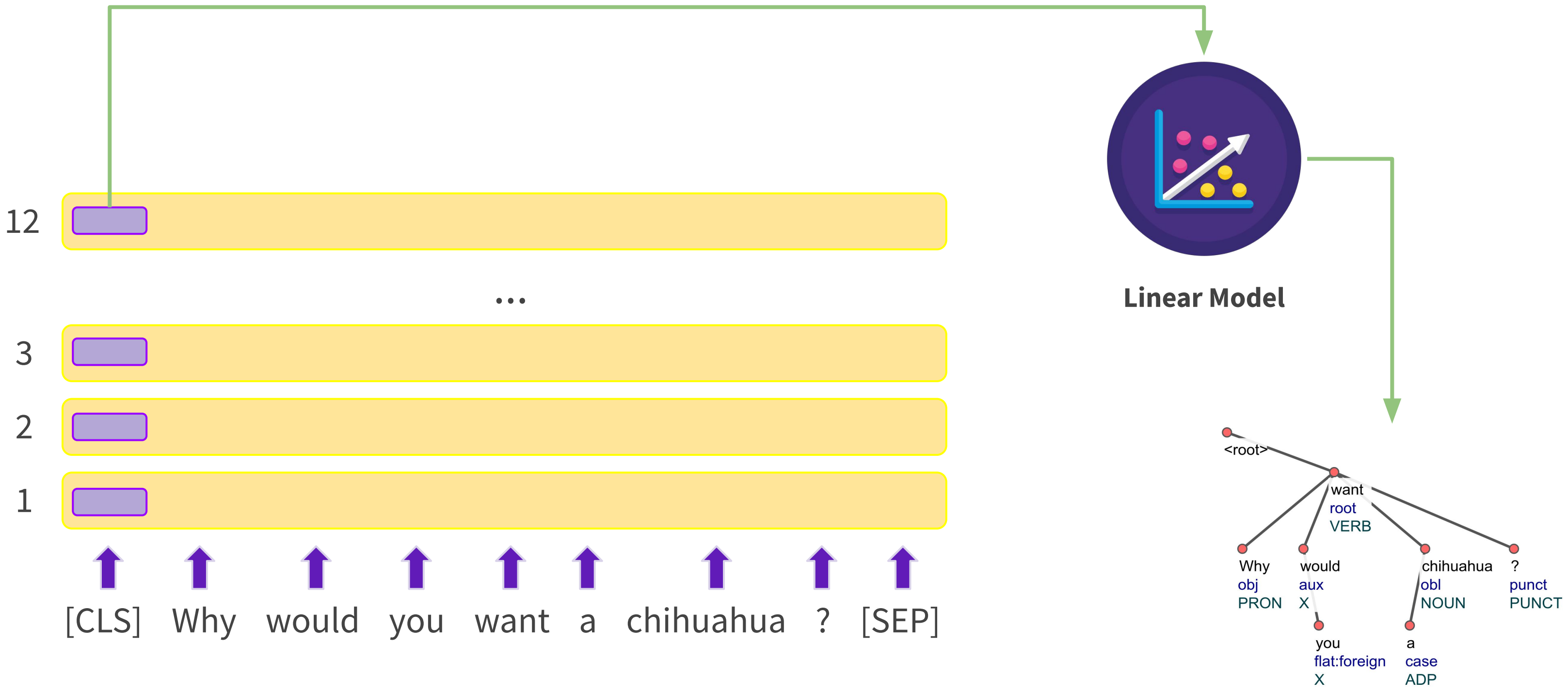
Profiling Neural Language Models



Profiling Neural Language Models



Profiling Neural Language Models



12

...

3

2

1

[CLS] Why would you want a chihuahua ? [SEP]

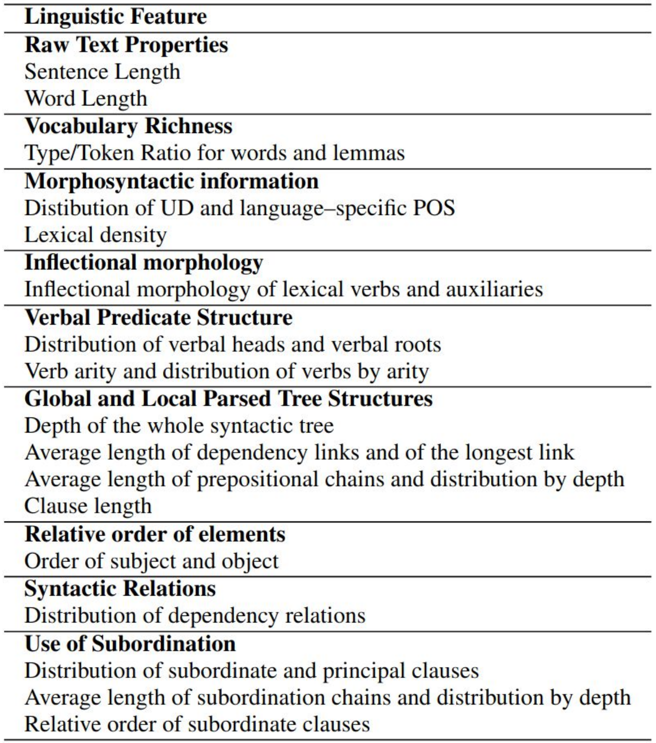
Linear Model

Profiling-UD: a tool for Linguistic Profiling of Texts

- ProfilingUD (Brunato et al., 2020) is a web–based application that performs linguistic profiling of a text, or a large collection of texts, for multiple languages

- It allows the extraction of more than 130 features, spanning across different levels of linguistic description

- Link: http://linguistic-profiling.italianlp.it/



Linguistic Feature
Raw Text Properties
Sentence Length
Word Length
Vocabulary Richness
Type/Token Ratio for words and lemmas
Morphosyntactic information
Distibution of UD and language–specific POS
Lexical density
Inflectional morphology
Inflectional morphology of lexical verbs and auxiliaries
Verbal Predicate Structure
Distribution of verbal heads and verbal roots
Verb arity and distribution of verbs by arity
Global and Local Parsed Tree Structures
Depth of the whole syntactic tree
Average length of dependency links and of the longest link
Average length of prepositional chains and distribution by depth
Clause length
Relative order of elements
Order of subject and object
Syntactic Relations
Distribution of dependency relations
Use of Subordination
Distribution of subordinate and principal clauses
Average length of subordination chains and distribution by depth
Relative order of subordinate clauses

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

- We investigated the linguistic knowledge implicitly encoded by BERT

Research questions:

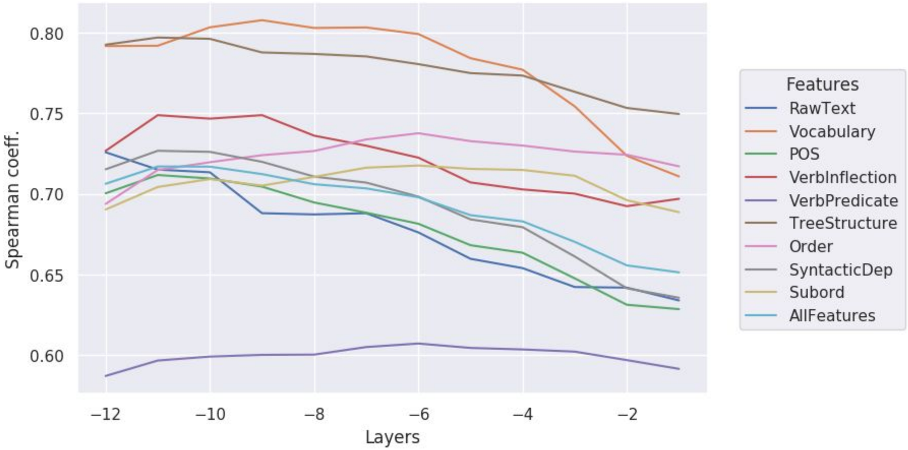
1. What kind of linguistic properties are encoded in a pre-trained version of BERT?

2. How this knowledge is modified after a fine-tuning (i.e. training of the model on a specific task) process

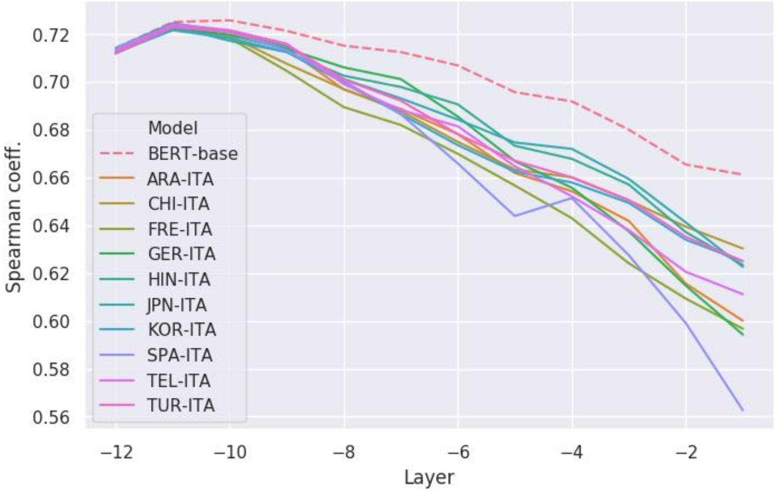
3. Whether this implicit knowledge affects the ability of the model to solve a specific downstream task

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

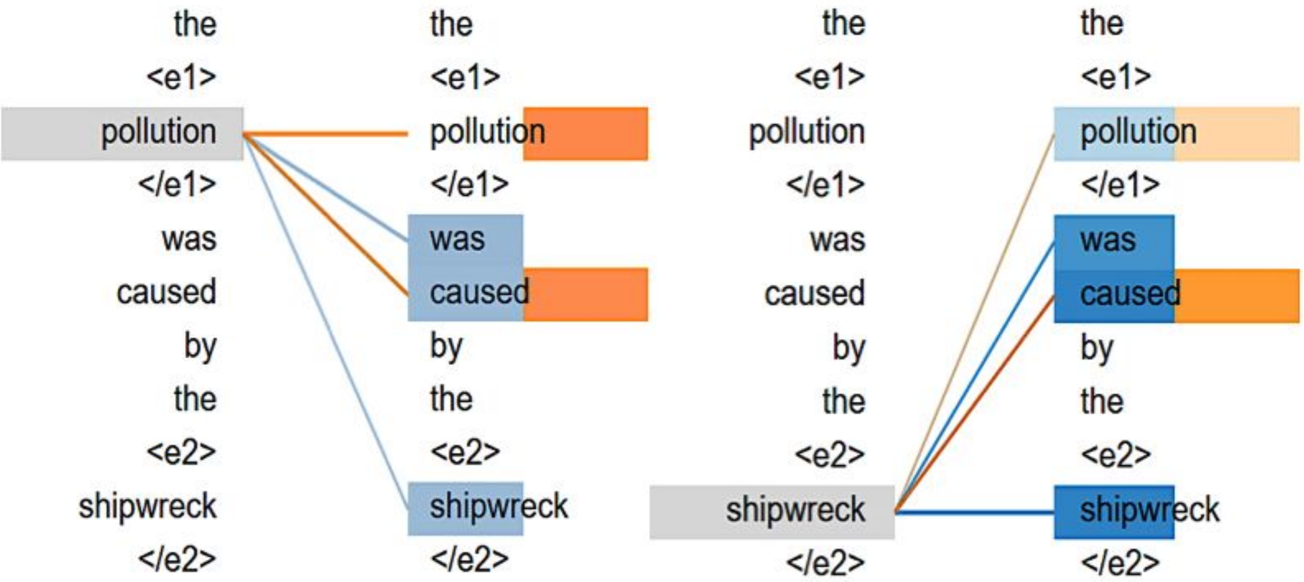
Pre-trained Model:



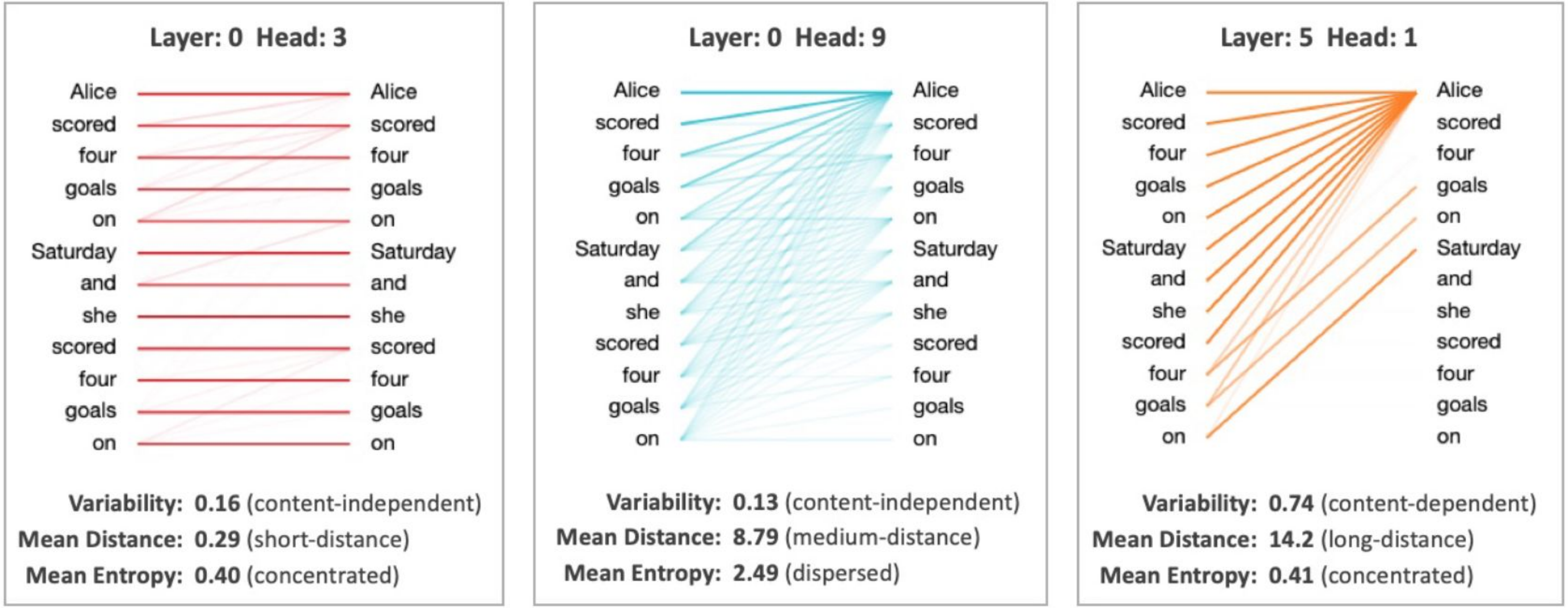
Fine-tuned Model:



Analysis of Attention Mechanisms



Analysis of Attention Mechanisms


Layer: 0 Head: 3
Alice scored four goals on Saturday and she scored four goals on
Alice scored four goals on Saturday and she scored four goals on
Variability: 0.16 (content-independent)
Mean Distance: 0.29 (short-distance)
Mean Entropy: 0.40 (concentrated)
Layer: 0 Head: 9
Alice scored four goals on Saturday and she scored four goals on
Alice scored four goals on Saturday and she scored four goals on
Variability: 0.13 (content-independent)
Mean Distance: 8.79 (medium-distance)
Mean Entropy: 2.49 (dispersed)
Layer: 5 Head: 1
Alice scored four goals on Saturday and she scored four goals on
Alice scored four goals on Saturday and she scored four goals on
Variability: 0.74 (content-dependent)
Mean Distance: 14.2 (long-distance)
Mean Entropy: 0.41 (concentrated)

Vig and Belinkov (2019)

Conclusion and Future Directions

- NLMs have reached astonishing performance in almost all NLP tasks
- However, this improvement comes at the cost of **interpretability**
- Several methods have been implemented to understand the inner mechanisms and decision-making processes of these models
  - and it is an ever-evolving and exciting area of research (e.g. Li et al., 2022, Bensemann et al., 2022)

Conclusion and Future Directions

- NLMs have reached astonishing performance in almost all NLP tasks
- However, this improvement comes at the cost of **interpretability**
- Several methods have been implemented to understand the inner mechanisms and decision-making processes of these models
  - and it is an ever-evolving and exciting area of research (e.g. Li et al., 2022, Bensemann et al., 2022)

Future Directions:

- Study how the linguistic knowledge arise during the pre-training phase of a NLM and how it changes when dealing with different training objectives
- Improve the robustness of NLMs by e.g. selecting input data appropriately during the pre-training phase and thus strengthening their implicit linguistic competence
- …Prompting for linguistic competence? (Liu et al., 2021)

Thanks for the attention!

🌐 https://alemiaschi.github.io/

@AlessioMiaschi

🌐 http://www.italianlp.it/

@ItaliaNLP_Lab

References

- Bengio, Yoshua, et al. (2003). "A neural probabilistic language model." *The journal of machine learning research* 3, pages 1137-1155.
- Vaswani, Ashish, et al. (2017). "Attention is all you need." *Advances in Neural Information Processing Systems* (NEURIPS)
- Goldberg, Yoav (2019). "Assessing BERT's syntactic abilities." *arXiv preprint arXiv:1901.05287*.
- 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.
- 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)*.

References

- 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*.
- Vig, Jesse and Belinkov, Yonatan (2019). “Analyzing the Structure of Attention in a Transformer Language Model”. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, Florence, Italy