Profiling Neural Language Models

Alessio Miaschi
ItaliaNLP Lab, Istituto di Linguistica Computazionale (ILC-CNR), Pisa
alessio.miaschi@ilc.cnr.it
https://alemiaschi.github.io/
http://www.italianlp.it/alessio-miaschi/
I am a PostDoc at the ItaliaNLP Lab, Institute for Computational Linguistics “A. Zampolli” (ILC-CNR, 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.
I am a PostDoc at the ItaliaNLP Lab, Institute for Computational Linguistics “A. Zampolli” (ILC-CNR, 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 (ILC-CNR) 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

Research Fellows:
- Irene Dini

PhD Students:
- Luca Bacco
- Benedetta Iavarone
- Giovanni Puccetti

Master/Undergraduate/Visiting Students

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

- The rise of Neural Language Models
- Interpretability of Neural Language Models
- Case Study: Profiling Neural Language Model
- Conclusion and Future Directions
The rise of Neural Language Models
The field of NLP has seen an unprecedented progress in the last years. Much of this progress is due to the replacement of traditional systems with newer and more powerful Deep Learning (DL) models.
Introduction

- The field of NLP has seen an unprecedented progress in the last years
- Much of this progress is due to the replacement of traditional systems with newer and more powerful Deep Learning (DL) models
Introduction

- The field of NLP has seen an unprecedented progress in the last years
- Much of this progress is due to the replacement of traditional systems with newer and more powerful Deep Learning (DL) models
Neural Language Models

- Neural Network (NN) model trained to approximate the **language modeling** function

- A probabilistic language model (LM) defines the 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 Models

- Neural Network (NN) model trained to approximate the **language modeling** function

- A probabilistic language model (**LM**) defines the probability of a sentence $s = [w_1, w_2, \ldots, w_n]$ as:

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

- Bengio et al. (2003) proposed a model that assigns a distributed vector for each word and then uses a NN architecture to predict the next word → **Neural Probabilistic Language Model**
Transformer Models

- Nowadays, the Transformer architecture has become the preferred solution for the development of state-of-the-art NLMs.

- Transformers (Vaswani et al., 2017) use only attention and fully connected layers to create highly scalable networks capturing distant patterns.
Transformer Models

- Nowadays, the Transformer architecture has become the preferred solution for the development of state-of-the-art NLMs.

- Transformers (Vaswani et al., 2017) use only attention and fully connected layers to create highly scalable networks capturing distant patterns.

- Attention is the method that allows the model to "attend" to different positions of the input sequence to compute a representation of that sequence.

\[
Attention(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V
\]
Transfer Learning

Pretraining:
- word2vec
- GloVe
- skip-thought
- InferSent
- ELMo
- ULMFiT
- GPT
- BERT

Adaptation:
- classification
- sequence labeling
- Q&A
- ....

Transfer Learning

Transfer Learning

Pretraining

word2vec
GloVe
skip-thought
InferSent
ELMo
ULMFiT
GPT
BERT

Adaptation / Fine-tuning

classification
sequence labeling
Q&A
....

Transfer Learning

Pretraining:
- word2vec
- GloVe
- skip-thought
- InferSent
- ELMo
- ULMFiT
- GPT
- BERT

Adaptation:
- classification
- sequence labeling
- Q&A
- ....

BERT (Devlin et al., 2019)

- Encoder model (12/24 layers)
- Trained to approximate the Masked Language Modeling (MLM) function
**BERT (Devlin et al., 2019)**

- Encoder model (12/24 layers)
- Trained to approximate the **Masked Language Modeling (MLM)** function
- The model can be fine-tuned in order to solve several NLP tasks:
  - Sentiment analysis;
  - Question answering;
  - Textual entailment;
  - etc.
Parameters Are All You Need
Interpreting Neural Language Models
The Case for Interpretability

- The development of powerful state-of-the-art NLMs comes at the cost of **interpretability**, since complex NN models offer little transparency about their inner workings and their abilities.
The Case for Interpretability

- The development of powerful state-of-the-art NLMs comes at the cost of interpretability, since complex NN models offer little transparency about their inner workings and their abilities

Objectives:

- **Understand the nature of AI systems** → be faithful to what influences the AI decisional process
- **Empower AI system users** → derive actionable useful insights from AI choices
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.”

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.”


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:
  ○ Probing tasks (e.g. Hewitt and Manning, 2019; Pimentel et al., 2020);
  ○ Analysis of attention mechanisms (e.g. Clark et al., 2019);
  ○ Definition of diagnostic tests (e.g. Goldberg, 2019);
  ○ etc.
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 BERT’s Syntactic Abilities \cite{Goldberg2019}

<table>
<thead>
<tr>
<th>SUBJECT-VERB AGREEMENT</th>
<th>BERT Base</th>
<th>BERT Large</th>
<th>LSTM (M&amp;L)</th>
<th>Humans (M&amp;L)</th>
<th># Pairs (# M&amp;L Pairs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple</td>
<td>1.00</td>
<td>1.00</td>
<td>0.94</td>
<td>0.96</td>
<td>120 (140)</td>
</tr>
<tr>
<td>In a sentential complement</td>
<td>0.83</td>
<td>0.86</td>
<td>0.99</td>
<td>0.93</td>
<td>1440 (1680)</td>
</tr>
<tr>
<td>Short VP coordination</td>
<td>0.89</td>
<td>0.86</td>
<td>0.90</td>
<td>0.82</td>
<td>720 (840)</td>
</tr>
<tr>
<td>Long VP coordination</td>
<td>0.98</td>
<td>0.97</td>
<td>0.61</td>
<td>0.82</td>
<td>400 (400)</td>
</tr>
<tr>
<td>Across a prepositional phrase</td>
<td>0.85</td>
<td>0.85</td>
<td>0.57</td>
<td>0.85</td>
<td>19440 (22400)</td>
</tr>
<tr>
<td>Across a subject relative clause</td>
<td>0.84</td>
<td>0.85</td>
<td>0.56</td>
<td>0.88</td>
<td>9600 (11200)</td>
</tr>
<tr>
<td>Across an object relative clause</td>
<td>0.89</td>
<td>0.85</td>
<td>0.50</td>
<td>0.85</td>
<td>19680 (22400)</td>
</tr>
<tr>
<td>Across an object relative (no that)</td>
<td>0.86</td>
<td>0.81</td>
<td>0.52</td>
<td>0.82</td>
<td>19680 (22400)</td>
</tr>
<tr>
<td>In an object relative clause</td>
<td>0.95</td>
<td>0.99</td>
<td>0.84</td>
<td>0.78</td>
<td>15960 (22400)</td>
</tr>
<tr>
<td>In an object relative (no that)</td>
<td>0.79</td>
<td>0.82</td>
<td>0.71</td>
<td>0.79</td>
<td>15960 (22400)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>REFLEXIVE ANAPHORA</th>
<th>BERT</th>
<th>BERT</th>
<th>LSTM</th>
<th>Humans</th>
<th># Pairs (# M&amp;L Pairs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple</td>
<td>0.94</td>
<td>0.92</td>
<td>0.83</td>
<td>0.96</td>
<td>280 (280)</td>
</tr>
<tr>
<td>In a sentential complement</td>
<td>0.89</td>
<td>0.86</td>
<td>0.86</td>
<td>0.91</td>
<td>3360 (3360)</td>
</tr>
<tr>
<td>Across a relative clause</td>
<td>0.80</td>
<td>0.76</td>
<td>0.55</td>
<td>0.87</td>
<td>22400 (22400)</td>
</tr>
</tbody>
</table>

Table 3: Results on the Marvin and Linzen \cite{Marvin2018} stimuli. M&L results numbers are taken from Marvin and Linzen \cite{Marвин2018}. 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

no further fine-tuning

the encoder's weights are fixed

Encoder Layer

N x

input text

Slide from: https://people.cs.umass.edu/~miyyer/cs685_f20/slides/19-probes.pdf
Probing Task Approach

**Findings:**

- BERT encodes linguistic information in a hierarchical manner \((\text{Tenney et al., } 2019)\)

- BERT encodes information about the structure of a syntax tree \((\text{Hewitt and Manning, } 2019)\)

- BERT contains relational knowledge competitive with symbolic knowledge bases \((\text{Petroni et al., } 2019)\)
Case Study:
Profiling Neural Language Models
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

- 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-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/](http://linguistic-profiling.italianlp.it/)

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

[CLS] Why would you want a chihuahua? [SEP]
Profiling Neural Language Models

[CLS] Why would you want a chihuahua? [SEP]
Profiling Neural Language Models

[Diagram showing a linear model with tokens: [CLS] Why would you want a chihuahua? [SEP]]

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

- Fine-tuning of BERT on the Native Language Identification (NLI)

“No breakfast, coz you still have enough alcohol in your stomach.”
Linguistic Profiling of a Neural Language Model (Miaschi et al., 2020)

- Fine-tuning of BERT on the Native Language Identification (NLI)

“No breakfast, coz you still have enough alcohol in your stomach.”
Linguistic Profiling of a Neural Language Model (Miaschi et al., 2020)

- Fine-tuning of BERT on the *Native Language Identification* (NLI)

  “No breakfast, coz you still have enough alcohol in your stomach.”

- Probing tasks on the fine-tuned model
Linguistic Profiling of a Neural Language Model (Miaschi et al., 2020)
Linguistic Profiling of a Neural Language Model (Miaschi et al., 2020)

- We have split each NLI dataset in sentences correctly and incorrectly classified by BERT
- We computed the MSE for each subset and each probing feature
Linguistic Profiling of a Neural Language Model (Miaschi et al., 2020)

- We have split each NLI dataset in sentences correctly and incorrectly classified by BERT
- We computed the MSE for each subset and each probing feature
Probing Tasks Under Pressure *(Miaschi et al., 2021)*

**Open Issue:**

- Are probing classification tasks really suited for performing such investigation or they simply hint for surface patterns in the data?
Probing Tasks Under Pressure (Miaschi et al., 2021)

Open Issue:

- Are probing classification tasks really suited for performing such investigation or they simply hint for surface patterns in the data?

Control Tasks:

- Hewitt and Liang (2019) introduced control tasks, i.e. a set of tasks that associate word types with random outputs that can be solved by simply learning regularities.
Open Issue:

- Are probing classification tasks really suited for performing such investigation or they simply hint for surface patterns in the data?

Control Tasks:

- Hewitt and Liang (2019) introduced control tasks, i.e. a set of tasks that associate word types with random outputs that can be solved by simply learning regularities

Our Contribution:

- We put increasingly under pressure the effectiveness of a suite of probing tasks to test the linguistic knowledge implicitly encoded by BERT on Italian sentences.
Probing Tasks Under Pressure (Miaschi et al., 2021)

Hypothesis: If the predictions using control datasets progressively diverge from the predictions on the gold dataset, this possibly suggests that probing tasks are effective to test the linguistic knowledge embedded in BERT representations.
Hypothesis:

If the predictions using control datasets progressively diverge from the predictions on the gold dataset, this possibly suggest that probing tasks are effective to test the linguistic knowledge embedded in BERT representations.
Probing Tasks Under Pressure (Miaschi et al., 2021)
Conclusion and Future Directions
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
Thanks for the $\text{softmax} \left( \frac{Q K^T}{\sqrt{D_k}} \right) V$!
References

References