Understanding Semantics in Neural Language Models via Learning Trajectories

Ejdis Gjinika[a,*]

[a]Università degli Studi di Brescia, Via Branze 38, Brescia, Italy
ORCID (Ejdis Gjinika): https://orcid.org/0009-0006-9817-5846

Abstract. Although Neural Language Models (NLMs) exhibit powerful performance on many NLP tasks, identifying what knowledge is contained in these models remains an open challenge. This work presents the application of four probing tasks based on figures of speech (hyperbole, metaphor, pleonasm, and oxymoron). The evaluations are based on the state-of-the-art Minimum Description Length (MDL) method and are conducted analysing learning trajectories among the model layers. The aim of this work is to understand whether NLMs have the semantic knowledge needed to identify these figures of speech, where such is located and when it is acquired during training. The temporal localization is conducted by analysing the model checkpoints, whereas the positional localization is about the model layers. The preliminary results show that this semantic knowledge is acquired in the initial phase of the training and is located in the middle layers of the models analyzed.

1 Introduction

Since the rise of Neural Language Models (NLMs), they showed promising abilities in different fields, such as machine translation, sentiment analysis and summarization [5, 7]. To complete these tasks, linguistic and syntactic knowledge has to be contained in these models, and many researchers focused their work on understanding and studying how this knowledge is learned and stored within these models [10, 3, 6].

One stable and widely used methodology for conducting these types of analysis is probing. This technique is used to understand what kind of knowledge is contained whitin a model. This is achieved by training a simple neural network, called *probe*, on some embedding representations produced by a frozen NLM. The idea is that if a probe successfully learns to complete the task, then the representations contain enough information relative to the task. Figure 1 shows schematically how a probe works. A growing line of research is the evaluation of how and which semantic knowledge is acquired by NLMs. In this direction, some work has been done: the authors of [1] design a probing task using metaphor to check if a models recognise their presence in a sentence instead the authors of [11] focus on hyperboles with the aim of understanding if NLMs can recognize this figure of speech in the English language. Other works focus on grammatical and syntactic knowledge and investigate if these proprieties are contained in BERT [9].

In general, the ability of NLMs to understand semantic aspects is a key component of a broader and more comprehensive understanding of language in a more general sense, and is still not fully explored, so more studies are needed in this direction. This work tries to fill this gap by applying methods to evaluate models' understanding of specific aspects of the language. In particular, the aspects analyzed are from the semantic part of the language, in particular, figurative language. These semantic concepts are systematically evaluated through probing of the available checkpoints and across the layers that compose the model. This allows us to conduct a more in-depth analysis and investigate this type of knowledge.

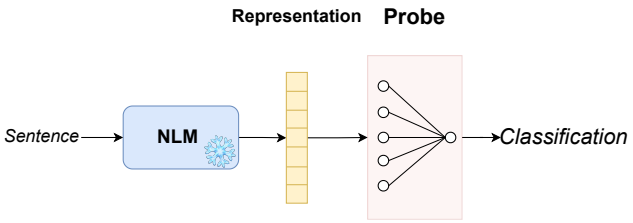

Figure 1. General scheme of the probing technique. A sentence is fed into a NLM (which is frozen) that produces a representation of the sentence. These representations are used to train the probe, i.e. a simple neural network, to complete a classification task.

In summary, this work addresses three research questions:

RQ1. When is the knowledge acquired during training?

RQ2. Where is this knowledge located, once learned?

RQ3. Does the model contain enough knowledge to identify these figures of speech?

2 Methodology

This work is based on the probing methodology and the preliminary results shown are based on the GPT-NeoX Pythia models [4]. The probes are designed as a neural network with two hidden layers of a number of neurons obtained by dividing the size of the embeddings of the models used by 4. These probes are trained using the representations produced by the model, as in the standard probing method. Subsequently, the probe is evaluated to finally understand whether the task is achieved by the NLM.

Probe evaluations The probe evaluations are conducted using the Minimum Description Length (MDL) approach, in particular in terms of the compression metric [13]. The accuracy metric could be the first choice for evaluating probes but this can lead to high values also for random initialized probes. So, accuracy is not the best

* Email: ejdis.gjinika@unibs.it

metric if we want a robust probe analysis. Instead, the compression metric quantifies the effort made by the probe to extract a specific concept together with the quality of the classification. These details make compression robust to random probes and more informative rather than accuracy.

Probing tasks and data The tasks are designed based on four figures of speech: (i) hyperbole (a deliberate exaggeration of a concept), (ii) metaphor (the transfer of meaning from one conceptual domain to another by analogy), (iii) pleonasm (the use of redundant words or expressions) and (iv) oxymoron (the juxtaposition of contradictory terms). The figures of speech are merely related to the semantic field of language since they are focused on the meaning of a group of words. The data used are taken from different sources and the author pre-processes them in order to verify their correctness. In detail, the hyperbole data is taken from [12], the metaphor data is from [2], the pleonasm dataset is adapted from [8], and the oxymoron dataset is taken from [14].

Learning Trajectories Since model knowledge is strictly correlated with model training, it is interesting to understand when this knowledge is acquired. For this purpose, we have to probe the model in his intermediate states, called checkpoints, saved during the training process. By executing the probes they produce a compression value for each checkpoint, and these compression values represent the learning trajectory of a probing task. Calculating these trajectories makes it possible to observe when the knowledge required to successfully complete a task, is acquired during training. Moreover, comparing the learning trajectory of different models on the same task allows to understand the dynamic of the training process. For example, a certain task can reach high compression values later during training with respect to another task.

Layer analysis Since all model layers produce an embedding representation of fixed length, it is possible to analyse the learning trajectories for all the intermediate layers. This lets us break down the localization of the knowledge inside the model architecture. Intuitively, a concept is located in a specific layer if his compression is higher than that of the other layers.

3 Preliminary results and discussion

Our preliminary results are shown in Figure 2 that highlight some key differences in the four evaluated probing tasks: the oxymoron trajectory reaches the highest compression value (about 6), far exceeding the other trajectories. The pleonasm and metaphor trajectories reach values near 1.5: this suggests that the model struggle to recognize these rhetorical figures. Regarding the hyperbole trajectory, it settles near 2 after $10B$ training steps. The key claim that we can extract from these results is that the **trajectories reach high values in the initial stages of training**, i.e. before the first $60B$ tokens. Therefore, as we can see in Figure 2, the trajectories at some point stabilize—around $73B$ tokens for oxymoron and around $10B$ for the others—and no longer increase in value. This behavior indicates that the models, even when trained on more tokens, do not improve their ability to understand this figure of speech. This finding is consistent with the key claim that such linguistic aspects are learned early during training. Another takeaway from Figure 2 is the consistently low values of the pleonasm trajectory, which remain between 1.46 and 1.49 throughout all training steps. This indicates that the model's ability to recognize and handle this figure of speech is quite limited. This could be due to the fact that the model may treat superfluous words as "perturbations", undermining performance on this task.

Table 1. Compression values of the *best* and the *last* layer at fixed checkpoints for the oxymoron task on the $70M$ model. The portion of tokens denote the percentage of the total token ($300B$) a checkpoint is trained on.

Layer	2%	4%	6%	8%	10%	20%	50%	100%
Best (2)	3.30	3.54	3.50	3.61	3.59	3.79	3.61	3.74
Last (6)	3.25	3.59	3.50	3.65	3.57	3.55	3.38	1.99

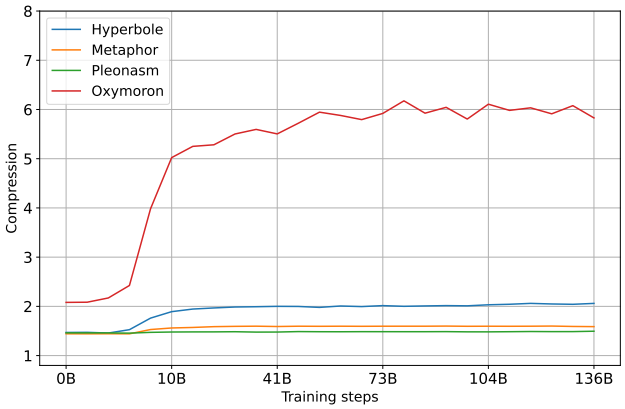

Figure 2. Best layer trajectories of the four figures of speech on the Pythia GPT-NeoX $410M$ model on $136B$ training steps.

From a preliminary probing on all the layers of the models, it emerges that the **layers that achieve a higher compression in their trajectory are usually located in the middle of the model**. Indeed, in the $410M$ model (which is composed of 24 layers) the best layer for each task are the layer 14 for hyperbole, 15 for metaphor and 12 for both oxymoron and pleonasm. A more detailed analysis, that we conduct on oxymoron, reveals that there are significant differences in the value of the best and the last layer. As summarized in Table 1, the compression values of the best layer are similar to the last layer on the first 10% portion of the $70M$ model on the oxymoron task, but the trajectory of the last layer begins to decrease considerably after 10%. In fact, the last layer has a compression at 10% of 3.57, at 20% of 3.55, at 50% of 3.38 and at the end of the training of 1.99, a decrease of $87,9\%$ with respect to the best layer. This is an interesting aspect of this type of analysis that is not yet fully explained by researchers.

4 Conclusions and Future work

In conclusion, this work addresses some concerns about the positional and temporal localization of semantic knowledge. By using the state-of-the-art Minimum Description Length (MDL) method and probing techniques applied to several checkpoints and layers of the models, we can locate this type of knowledge. The preliminary results shown are produced systematically using probing tasks and evaluating learning trajectories. Furthermore, to investigate the position of the acquired knowledge, the probing technique is extended to all layers of the model, identifying different performance across the layers. The results highlight that semantic knowledge is acquired early during training (RQ1) and is located in the middle layers of the model (RQ2). In addition, the different figures of speech have different degrees of detectability by the model (RQ3). To extend this work, it may be interesting to expand the probing tasks to different figures of speech or different types of knowledge. Another engaging future direction is to analyze other architectures and bigger models.

References

[1] E. Aghazadeh, M. Fayyaz, and Y. Yaghoobzadeh. Metaphors in pre-trained language models: Probing and generalization across datasets and languages. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2037–2050. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.ACL-LONG.144. URL https://doi.org/10.18653/v1/2022.acl-long.144.

[2] E. Aghazadeh, M. Fayyaz, and Y. Yaghoobzadeh. Metaphors in pre-trained language models: Probing and generalization across datasets and languages. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2037–2050, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.144. URL https://aclanthology.org/2022.acl-long.144/.

[3] Y. Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Comput. Linguistics*, 48(1):207–219, 2022. doi: 10.1162/COLI_A_00422. URL https://doi.org/10.1162/coli_a_00422.

[4] S. Biderman, H. Schoelkopf, Q. G. Anthony, H. Bradley, K. O'Brien, E. Hallahan, M. A. Khan, S. Purohit, U. S. Prashanth, E. Raff, A. Skowron, L. Sutawika, and O. van der Wal. Pythia: A suite for analyzing large language models across training and scaling. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 2397–2430. PMLR, 2023. URL https://proceedings.mlr.press/v202/biderman23a.html.

[5] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[6] Z. Jiang, F. F. Xu, J. Araki, and G. Neubig. How can we know what language models know. *Trans. Assoc. Comput. Linguistics*, 8:423–438, 2020. doi: 10.1162/TACL_A_00324. URL https://doi.org/10.1162/tacl_a_00324.

[7] K. S. Kalyan. A survey of GPT-3 family large language models including chatgpt and GPT-4. *Nat. Lang. Process. J.*, 6:100048, 2024. doi: 10.1016/J.NLP.2023.100048. URL https://doi.org/10.1016/j.nlp.2023.100048.

[8] O. Kashefi, A. T. Lucas, and R. Hwa. Semantic pleonasm detection. In M. Walker, H. Ji, and A. Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 225–230, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2036. URL https://aclanthology.org/N18-2036/.

[9] A. Miaschi, D. Brunato, F. Dell'Orletta, and G. Venturi. Linguistic profiling of a neural language model. In D. Scott, N. Bel, and C. Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 745–756. International Committee on Computational Linguistics, 2020. doi: 10.18653/V1/2020.COLING-MAIN.65. URL https://doi.org/10.18653/v1/2020.coling-main.65.

[10] F. Petroni, T. Rocktäschel, S. Riedel, P. Lewis, A. Bakhtin, Y. Wu, and A. H. Miller. Language models as knowledge bases? In K. Inui, J. Jiang, V. Ng, and X. Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2463–2473. Association for Computational Linguistics, 2019. doi: 10.18653/V1/D19-1250. URL https://doi.org/10.18653/v1/D19-1250.

[11] N. Schneidermann, D. Hershcovich, and B. S. Pedersen. Probing for hyperbole in pre-trained language models. In V. Padmakumar, G. Vallejo, and Y. Fu, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 200–211. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.ACL-SRW.30. URL https://doi.org/10.18653/v1/2023.acl-srw.30.

[12] E. Troiano, C. Strapparava, G. Özbal, and S. S. Tekiroğlu. A computational exploration of exaggeration. In E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3296–3304, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1367. URL https://aclanthology.org/D18-1367/.

[13] E. Voita and I. Titov. Information-theoretic probing with minimum description length. In B. Webber, T. Cohn, Y. He, and Y. Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.14. URL https://aclanthology.org/2020.emnlp-main.14/.

[14] F. Xu, Z. Zhu, and X. Wan. Creative destruction: Can language models interpret oxymorons? In F. Liu, N. Duan, Q. Xu, and Y. Hong, editors, *Natural Language Processing and Chinese Computing*, pages 645–656, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-44693-1.