Created By
Thanh Vo (ttv170230)
Alejo Vinluan (abv210001)

ACL Paper Summary
The Grammar-Learning Trajectories of Neural Language Models

Leshem Chohsen, Guy Hacohen, Daphna Weinshall, Omri Abend
Hebrew University of Jerusalem

The paper *The Grammar-Learning Trajectories of Neural Language Models* discusses the phenomenon (or problem) that seemingly appears throughout various Natural Language Models (NLMs). The Natural Language Models (NLMs) may have fundamentally different initial parameters that makes them distinctive yet they still traverse the 'learning space' with highly predictable pathing, always reaching notable **morphological phenomena clusters**. Simply, the NLMs have been observed to learn in distinct phases/checkpoints. Therefore, this concept suggests that the NLMs all share a **mutual inductive bias —** they make the same assumptions as they 'learn'.

The paper then discusses experimental observations amongst several different Natural Language Models (NLMs) such as **GTP2_small, GTP2_large**, **TransformerXL, and etc.** To verify and authenticate the claim that the phenomena lays highly distinct regardless of various external factors, the authors address various factors that would refute their claims for such a property by peering at each model's architecture, initial training data, and metric based performance. The authors then experiment to find the same correlation given that several variables remain static while having the architecture and initial training data be variable. As the paper concludes, after doing several experiments they found that the correlation that they initially hypothesized was highly consistent.

The paper seeks to discuss a discovery from their research into a topic that the field of machine learning has a lack of focus on a more broad analysis of the interactions between language model **learning dynamics** and **learned representations**. Basically, the **learning dynamics** describe the methodology of language models used to learn. **Learn representations** is the end-resultant, predictive conclusion, or target that the language model produces. The paper notes that while there is significant focus on **learning dynamics** for the *improvement* of the language model, there is very little analysis on the relationship between **learning dynamics** and **learned representations**. The paper authors call this taking a more **behavioral approach**: only utilizing observable and measurable aspects of human behavior.

The authors utilized BLiMP to evaluate how far Natural Language Processors make generalizations. BLiMP is The Benchmark of Linguistic Minimal Pairs for English, which is a challenge set isolating major grammatical phenomena in English including specific contrasts in syntax, morphology, and semantics. First, the authors looked to evaluate how similar different Natural Language Model’s networks were. The authors utilized a formula that checked the performance of a certain model against BLiMP at certain checkpoints. The authors were able to find that different language models had high correlation after 5000 steps in BLiMP, even when utilizing different initializations or training data. The correlations are considered extremely high after 10,000 steps, which shows a clear correlation between different Natural Language Models.

The authors also wanted to evaluate the performance of different architectures. They used TransformerXL and GPT2 to test against each other. The checkpoints utilized above were not equivalent, so the authors utilized the perplexity on the development step to mark their checkpoints. The findings concluded that different architectures also have similar correlations at the same checkpoints.

Finally, the authors wanted to rule out how similar training data may have been used to train the different models. They found that TransformerXL was trained on Wikipedia whereas GPT2 was trained on scraped web pages. In order to further flesh out the extent of the effect of training data, the authors trained 3 GPT2tiny instances on different datasets: openWebText, openSubtitles, and newsCrawl. The authors concluded that the initial dataset is important to correlation, but initialization was not affected by the initial dataset.

Utilizing the above points, the authors were able to determine that the different Natural Language Models follow a similar trajectory during training. **The training size, time, and efficiency may affect what a model has learned, but not the order it is learning.** Once the authors started comparing the Natural Language Models to non-neural Language Models, they concluded that the models are no longer similar, even if the performance was similar. The authors tested a GPT2tiny against 2 different 5-gram Language Models where one was trained on WikiBooks and the other trained on GigaWord. The authors were able to conclude that different models have different internal biases. This continues to prove that **different models maintain an order of learning despite different architectures, model sizes, and training sets.**

Overall, the authors use multiple tests in order to conclude that different Natural Language Models all have similar learning trajectories. They isolate the models based on factors such as architecture, training data, and initialization to find how each of these factors affect learning trajectory. Finally, they compare the performance of each model based on the BLiMP benchmark and confirm their observations.

The authors of the paper are students and alumni based around Professor Daphna Weinshall's lab at The Hebrew University of Jerusalem. Professor Weinshall graduated from Tel-Aviv University with a PhD in Mathematics and Statistics in 1986, with a focus on models of evolution and population genetics. She researches computer vision, biological vision, and machine learning. Her recent publications are based around Learning and Object Recognition, Computer Vision, Cognitive Neuroscience, and Population Genetics which have generated 9755 citations within Google Scholar, which continues to grow as she fleshes out her research.

Leshem Chohsen is currently an AI researcher for IBM with a focus on Natural Language Processing. He was able to get a PhD from The Hebrew University of

Jerusalem, where he specialized on evaluating and training text generation through linguistic knowledge. Before joining IBM, Dr. Chohsen's work revolved around language data analysis. He wrote several publications including "Automatic metric valuation for grammatical error correction" and "Classifying syntactic errors in learner language".

Guy Hacohen is a PhD student at The Hebrew University of Jerusalem. He is focused on understanding how deep models interact with data. His recent publications are also written in association with Dr. Weinshall. These include "Let's Agree to Agree: Neural Networks Share Classification Order on Real Datasets" and "Principal Components Bias in Over-parameterized Linear Models, and its Manifestation in Deep Neural Networks".

Omri Abend is a research member at The Hebrew University of Jerusalem with a specialty in Computational Linguistics and Natural Language Processing. His research is based around statistical learning, language technology, and computational modeling of child language acquisition. His recent publications include "Semantics-aware Attention Improves Neural Machine Translation" and "Paths to Relation Extraction through Semantic Structures".

The authors of this paper overall have a focus and speciality on different aspects of Natural Language Processing. Dr. Chohsen and Omri Abend specialize in language itself and the statistics within linguistics. Dr. Weinshall and Guy specialize in the broad-level knowledge of Machine Learning and Natural Language Processing.

With all that in mind, their work is as important as the discovery of a new physics principle. It, in the field of Natural Language Processing (NLP) and Natural Language Models (NLM), provides a conceptual anchor that several scientists and researchers are able to further evaluate and analyze on. It gives contextual power to predictive tools and makes easily marked classifications on model learning conceptions which will potentially provide further understanding of language models. This will propagate further advances in the field and, one day, be notable for its significant upheaval towards getting Artificial Intelligence closer to human speech. This paper has been cited 10 times as indicated by Google Scholar.