Arrows of Time for Large Language Models

Vassilis Papadopoulos* Jérémie Wenger† Clément Hongler*

*EPFL †Goldsmiths, University of London

Next-Token Prediction:

Once upon a time, there was a $\boxed{?}$ …

$\rightarrow$ Estimate $\mathbb{P}\{X_k = x_k \mid X_1 = x_1, \dots, X_{k-1} = x_{k-1}\}\ \forall k = 1, \dots, n$

Forward Model $\overrightarrow{\mathcal{M}}$:

$$\overrightarrow{\mathbb{P}}\{X_1 = x_1, \dots, X_n = x_n\} = \prod_{k=1}^{n} \overrightarrow{\mathbb{P}}\{X_k = x_k \mid X_1 = x_1, \dots, X_{k-1} = x_{k-1}\}$$

Previous-Token Prediction:

… $\boxed{?}$ and they lived happily ever after.

$\rightarrow$ Estimate $\mathbb{P}\{X_k = x_k \mid X_{k+1} = x_{k+1}, \dots, X_n = x_n\}$

Backward Model $\overleftarrow{\mathcal{M}}$:

$$\overleftarrow{\mathbb{P}}\{X_1 = x_1, \dots, X_n = x_n\} = \prod_{k=1}^{n} \overleftarrow{\mathbb{P}}\{X_k = x_k \mid X_{k+1} = x_{k+1}, \dots, X_n = x_n\}$$

Two $\mathbb{P}$ Factorizations:

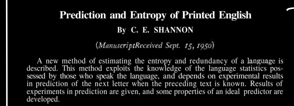
$$\mathbb{P}\{X_1 = x_1\}\mathbb{P}\{X_2 = x_2 \mid X_1 = x_1\} \cdots \mathbb{P}\{X_n = x_n \mid X_1 = x_1, \dots, X_{n-1} = x_{n-1}\} = \mathbb{P}\{X_n = x_n\}\mathbb{P}\{X_{n-1} = x_{n-1} \mid X_n = x_n\} \cdots \mathbb{P}\{X_1 = x_1 \mid X_2 = x_2, \dots, X_n = x_n\}$$

Train a FW & BW copies of the same LLM (BW $\Leftrightarrow$ FW on time-reversed dataset) $\rightsquigarrow$ Do we have $\overrightarrow{\mathbb{P}} = \overleftarrow{\mathbb{P}}$?

Information-Theoretically: No Difference Between $\overrightarrow{\mathbb{P}}$ and $\overleftarrow{\mathbb{P}}$

$\hookrightarrow$ Do we see $\overrightarrow{\ell}_{CE} = \overleftarrow{\ell}_{CE}$?

Shannon's Experiment: Next- and Previous-Letter Prediction


Prediction and Entropy of Printed English

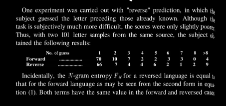
By C. E. SHANNON

(Manuscript Received Sept. 15, 1950)

A new method of estimating the entropy and redundancy of a language is described. This method exploits the knowledge of the language statistics possessed by those who speak the language, and depends on experimental results in prediction of the next letter when the preceding text is known. Results of experiments in prediction are given, and some properties of an ideal predictor are developed.

The idea of measuring the sum of cross-entropies in natural languages was pioneered by Shannon. He noted that this could also be done backwards…

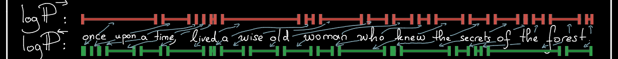
Experiments were performed on human subjects; Shannon noted that to his surprise, they would perform worse predicting backwards, but only slightly so.


One experiment was carried out with "reverse" prediction, in which the subject guessed the letter preceding those already known. Although the task is subjectively much more difficult, the scores were only slightly poorer. Thus, with two 101 letter samples from the same source, the subject obtained the following results:

Incidentally, the N-gram entropy F_N for a reversed language is equal to that for the forward language as may be seen from the second form in equation (1). Both terms have the same value in the forward and reversed cases.

Training: Minimize Cross-Entropy Losses

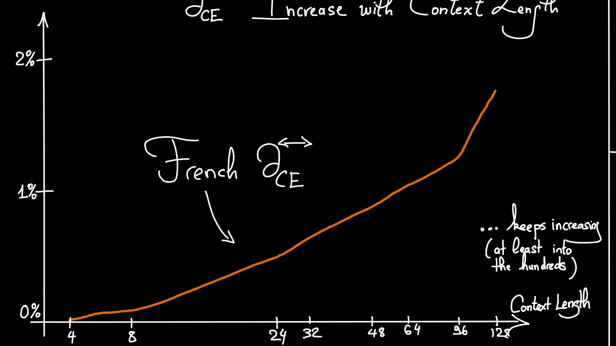
$$\overrightarrow{\ell}_{CE} = \sum_{k=1}^{n} -\log \overrightarrow{\mathbb{P}}\{X_k = x_k \mid X_1 = x_1, \dots, X_{k-1} = x_{k-1}\}$$
$$= -\log \overrightarrow{\mathbb{P}}\{X_1 = x_1, \dots, X_n = x_n\}$$



$$\overleftarrow{\ell}_{CE} = \sum_{k=1}^{n} -\log \overleftarrow{\mathbb{P}}\{X_k = x_k \mid X_{k+1} = x_{k+1}, \dots, X_n = x_n\}$$
$$= -\log \overleftarrow{\mathbb{P}}\{X_1 = x_1, \dots, X_n = x_n\}$$

$\hookrightarrow$ if $\overrightarrow{\mathbb{P}} = \overleftarrow{\mathbb{P}}$ then we should have $\overrightarrow{\ell}_{CE} = \overleftarrow{\ell}_{CE}$

$\overleftrightarrow{\partial}_{CE}$ Increase with Context Length

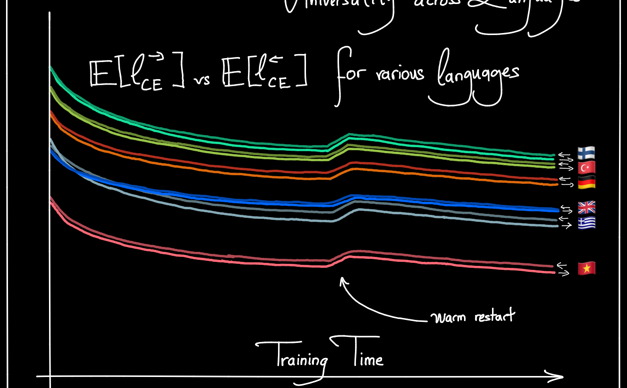


Arrow of Time: $\overleftrightarrow{\partial}_{CE} := \dfrac{\mathbb{E}[\overleftarrow{\ell}_{CE} - \overrightarrow{\ell}_{CE}]}{\frac{1}{2}\mathbb{E}[\overleftarrow{\ell}_{CE} + \overrightarrow{\ell}_{CE}]}$

measures the asymmetry in the estimability of languages by LLMs

Universality across Languages

$\mathbb{E}[\overrightarrow{\ell}_{CE}]$ vs $\mathbb{E}[\overleftarrow{\ell}_{CE}]$ for various languages

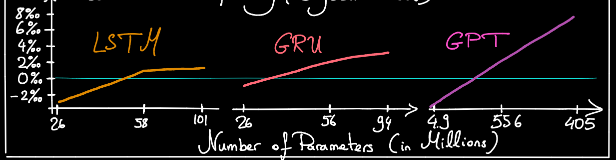


Key Takeaways

- Arrows of Time are universal across languages, model architectures, model sizes, and training times, provided the datasets are large enough and the models have enough parameters and training time
- While the effect is impressively robust, $\overleftrightarrow{\partial}_{CE}$ is never very large (at most a few percent)
- The effect size increases with the context length
- Artifacts due to the tokenization can be ruled out

Universality across Architectures: $\overleftrightarrow{\partial}_{CE} > 0$ as soon as models get large enough

$\overleftrightarrow{\partial}_{CE}$ at the end of training (English dataset)



Open Questions & Perspectives

- Arrows of Time in Code?
- Continuous Setting?
- Arrows of Time in DNA?
- Link with Thermodynamics?
- Arrows of Time $\leftrightarrow$ Life?
- Link with Causality?
- Quadratic Languages and Complexity?
- Stock Market Prices?
- Small-Data Arrows of Time?
- Scaling Laws?

Why Arrows of Time? $\rightsquigarrow$ Mathematical Models

Simple Model: Prime Multiplication

Language $\boxed{3 \times 19 = 57}$ $\boxed{11 \times 13 = 143}$ $\boxed{11 \times 17 = 187}$ … $\boxed{p \times q = pq}$ …

Easier to learn to multiply than to factor $\rightsquigarrow \overleftrightarrow{\partial}_{CE} > 0$

How can $\overleftrightarrow{\partial}_{CE}$ become spontaneously > 0?

Learning and Sparsity: Linear Languages

$\mathbb{P}_A: [\underbrace{x_1 x_2 \cdots x_m}_{m \text{ iid bits}} : \underbrace{y_1 y_2 \cdots y_m}_{m \text{ iid bits}}] \in \mathbb{F}_2^m$

$y = \overrightarrow{A}x \qquad x = \overleftarrow{A}y \qquad (\overrightarrow{A})^{-1} = \overleftarrow{A}$

$\overleftarrow{\mathcal{M}}$ learns $\mathbb{P}_A$ easily starting from $\mathbb{P}_B$

$\longleftrightarrow$ $\overleftarrow{A} - \overleftarrow{B}$ sparse

Sparsity Symmetry Breaking

$\overrightarrow{A} - \overrightarrow{B}$ sparse $\Rightarrow$ Typically $\overleftarrow{A} - \overleftarrow{B}$ less sparse

Communication Model: Alice, Bob, and Carol

Suppose Alice, Bob, and Carol share a common language $\mathbb{P}_B$, with Alice and Bob using forward models and Carol using a backward model.

Now if Alice manages to learn $\mathbb{P}_A$ easily from $\mathbb{P}_B$, $\overrightarrow{A} - \overrightarrow{B}$ should be sparse, so if she sends $\mathbb{P}_A$ samples Bob will learn $\mathbb{P}_A$ easily; for Carol it will be harder

$\hookrightarrow$ Language is "selected" to be easy forward, so backward is harder

Acknowledgements: Special thanks to Stéphane d'Ascoli, Samy Bengio, Gloria Capano, Diego Dorn, Franck Gabriel, Alex Graves, Ron Maimon, Jacob Menick, Christos Papadimitriou, João Penedones, Matthieu Wyart, and Nicolas Zlatoff for interesting discussions and suggestions. Thanks to the Quine / Futurarium Association for initial support. Financial support from an EPFL FSB Seed Grant for Alife is acknowledged. Thanks to Franck Gabriel and Pauline Lamotte for design advice. Thanks to MA Travel for organizing the trip during which this poster was designed. Timestamp: 2024-07-18