What is the landscape of natural language?

Eric De Giuli (degiuli@lpt.ens.fr)

Institut de Physique Théorique Philippe Meyer, École Normale Supérieure, Paris

Note: for a brief summary, read only Motivation and Discussion sections.

Motivation

Successful deep learning architectures employ compositionality (e.g. [14]) and hierarchical representation. These features also underlie natural language: sentences are composed of words, and sentences have a hierarchical 'deep' structure (e.g. Fig 1a). So why do current ML systems struggle with natural language? Here we use a statistical physics perspective to characterize the emergence of deep structure in a model of natural language, the Random Language Model (RLM) [8].

Deep structure was formalized by Post and Chomsky with the notion of generative grammar [10, 4]. A generative grammar is a set of rules, operating by replacement, such that from an initial start symbol S, one can produce a set of 'grammatical' strings, called sentences. The sequence of rule applications is called a derivation. For example, the grammar $\{S \to SS, S \to (S), S \to ()\}$ produces all strings of well-formed parentheses, which constitute the language of the grammar. A simple derivation in this grammar is $S \to SS \to (S)S \to (())S \to (())()$. In linguistics, the observable symbols are typically taken to be words, and grammars produce sentences (Fig 1a) [4, 5].

The complexity of a language is limited by conditions imposed on its grammar, as described by the Chomsky hierarchy. Here we focus on context-free grammars (CFGs), whose derivations can be drawn as trees (Fig 1a). CFGs are the lowest order of the Chomsky hierarchy that support hierarchical structure. We find that CFGs possess two natural 'temperature' scales that control grammar complexity, one at the surface interface (ϵ_s), and another in the tree interior (ϵ_d). As either of these temperatures is lowered, there is a phase transition, which corresponds to the emergence of nontrivial information propagation. We characterize this phase transition using results from simulations, and understand its location by a balance between energy and entropy.

Generative grammars

A generative grammar is defined by an alphabet χ and a set of rules $\mathcal{R}$. The alphabet has N hidden, 'non-terminal' symbols χ_N, and T observable, 'terminal' symbols χ_T. The most general rule is of the form $a_1 a_2 \dots a_n \to b_1 b_2 \dots b_m$, where $a_i \in \chi_N, b_i \in \chi = \chi_N \cup \chi_T$. In a CFG, the rules are specialized to the form $a_1 \to b_1 b_2 \dots b_m$, and we will insist that $m \geq 1$, so that there is no 'empty' string. Without loss of generality, any such CFG can be put into Chomsky normal form, in which case all rules are of the form [9]

$$a \to b\, c \qquad \text{or} \qquad a \to A, \tag{1}$$

where $a, b, c \in \chi_N$ and $A \in \chi_T$. Note that we may have $b = a$, or $b = c$, or $a = b = c$. Any derivation in Chomsky reduced form can be drawn on a binary tree. We consider CFGs in this form. Beginning from the start symbol $S \in \chi_N$, rules are applied until the string contains only observable symbols. Such a string is called a sentence. The set of all sentences is the language of the grammar. Given a string of observables $\mathcal{S} = A_1 \dots A_\ell$ and a grammar $\mathcal{G}$, one can ask whether there exists a derivation that produces $\mathcal{S}$ from the start symbol S; if so, $\mathcal{S}$ is said to be grammatical. To enable continuous learning, we give each rule a non-negative real valued weight. For CFGs, to every rule of the form $a \to bc$ we assign a weight M_{abc}, and to every rule of the form $a \to A$ we assign a weight O_{aA}.

Each candidate derivation of a sentence has two different types of degrees of freedom. There is the topology $\mathcal{T}$ of the tree, namely the identity (terminal or non-terminal) of each node, as well as the variables, both terminal and non-terminal, on the nodes. We write $\Omega_\mathcal{T}$ for the set of internal factors, i.e. factors of the form $a \to bc$, and $\partial\Omega_\mathcal{T}$ for the boundary factors, i.e. those associated to $a \to A$ rules. The number of boundary factors is written $\ell_\mathcal{T}$, which is also the number of leaves. Since derivations are trees, the number of internal factors is $\ell_\mathcal{T} - 1$. We will write σ for non-terminal symbols, and o for terminals; these can be enumerated in an arbitrary way $1, \dots, N$ and $1, \dots, T$, respectively. Given $\mathcal{T}$, we can write σ_i for the value of the non-terminal on site i, and similarly o_j for the terminal on site j. The number of σ_i is $2\ell_\mathcal{T} - 1$, while the number of o_j is $\ell_\mathcal{T}$. We write $\mathcal{G}$ for the pair M, O.

To define a probability measure on parses, it is convenient to factorize it into the part specifying $\mathcal{T}$, and the remainder. In this way we separate the the tree shape from the influence of the grammar on variables. For a fixed $\mathcal{T}$ the weight of a parse $\{\sigma_i, o_t\}$ is

$$W(\{\sigma_i, o_t\}|\mathcal{T}, \mathcal{G}) = \prod_{\alpha \in \Omega_\mathcal{T}} M_{\sigma_{\alpha_1}\sigma_{\alpha_2}\sigma_{\alpha_3}} \prod_{\alpha \in \partial\Omega_\mathcal{T}} O_{\sigma_{\alpha_1} o_{\alpha_2}}, \tag{2}$$

where each $\alpha = (\alpha_1, \alpha_2, \alpha_3)$ is a factor in the order $\sigma_{\alpha_1} \to \sigma_{\alpha_2}\sigma_{\alpha_3}$.

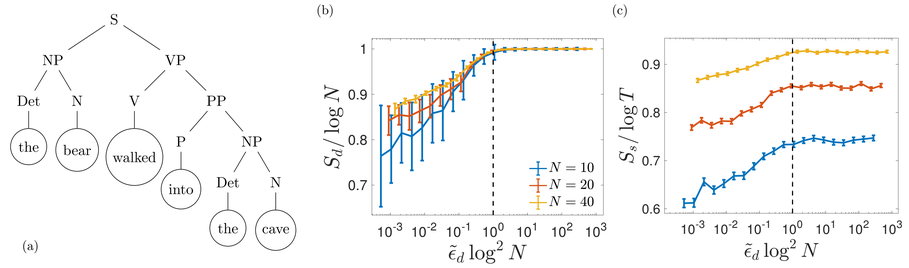

Fig. 1. (a) Illustrative derivation tree. (b,c) Shannon entropy of random CFGs as functions of $\tilde{\epsilon}_d = \epsilon_d/N^3$. (b) Entropy of hidden configurations. (c) Entropy of observed strings. The constant value for $\epsilon_d > \epsilon_*$ depends on the surface temperature ϵ_s. Bars indicate 20^{th} and 80^{th} percentiles, indicating the variation over grammars at each parameter value.

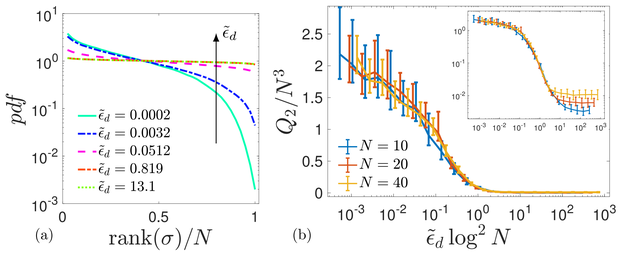

Fig. 2.(a) Zipf plot of hidden symbols for $N = 40$: the frequency of each hidden symbol, arranged in decreasing order. Here $\tilde{\epsilon}_d = \epsilon_d/N^3$. (b) Order parameter Q_2, with bars indicating 20^{th} and 80^{th} percentile ranges over grammars at each parameter value. Inset: same plot in log-log axes.

This defines a conditional probability measure on parses

$$\mathbb{P}(\{\sigma_i, o_t\}|\mathcal{T}, \mathcal{G}) = \frac{W(\{\sigma_i, o_t\}|\mathcal{T}, \mathcal{G})}{Z(\mathcal{T}, \mathcal{G})} \tag{3}$$

where

$$Z(\mathcal{T}, \mathcal{G}) = \sum_{\{\sigma_i, o_t\}} W(\{\sigma_i, o_t\}|\mathcal{T}, \mathcal{G}) \tag{4}$$

All parses have S at the root node. For simplicity, in this work we consider as a model for the tree topology probability $\mathbb{P}(\mathcal{T}|\mathcal{G}) = W_{tree}/Z_{tree}$ with $W_{tree}(\mathcal{T}) = p^{|\partial\Omega_\mathcal{T}|}(1-p)^{|\Omega_\mathcal{T}|}$, where p is the emission probability, the probability that a hidden node becomes an observable node. p controls the size of trees; we will choose it such that the tree size distribution is cutoff above a length $\xi = 1000$ (for details see [8]). A model with weights of the form (2) is called a weighted CFG (WCFG). We will scale M and O such that their median values are the corresponding uniform probabilities, $\overline{M} = 1/N^2$ and $\overline{O} = 1/T$. (In [8] we show that our results are robust in a model with strict normalization of weights).

Random Language Model

What is an appropriate measure on grammars? If grammar weights are the accumulation of many small, independent multiplicative effects, then they will lead to a lognormal distribution. Define deep and surface sparsities as, respectively, $s_d = \frac{1}{N^3}\sum_{a,b,c} \left|\log M_{abc}/\overline{M}\right|^2$, $s_s = \frac{1}{NT}\sum_{a,B} \left|\log O_{aB}/\overline{O}\right|^2$, A lognormal distribution of grammar weights is $\mathbb{P}_G(M, O) \equiv Z_G^{-1}\, J\, e^{-\epsilon_d s_d} e^{-\epsilon_s s_s}$ where $J = e^{-\sum_{a,b,c}\log M_{abc} - \sum_{a,B}\log O_{aB}}$. We define the RLM as the ensemble of grammars drawn from this distribution.

The Lagrange multipliers ϵ_d and ϵ_s satisfy $\overline{s_d} = N^3/(2\epsilon_d), \overline{s_s} = NT/(2\epsilon_s)$. When $\epsilon_d \to \infty$, $\overline{s_d} \to 0$, which is the value corresponding to a completely uniform deep grammar, that is, when for a non-terminal a, all rules $a \to bc$ have the same probability $1/N^2$. This is clearly the limit in which the grammar carries no information. As ϵ_d is lowered, the deep sparsity increases, and the grammar carries more information. Thus ϵ_d plays the role of temperature; we will refer to it as the deep temperature. Similarly, ϵ_s controls information transmission at the surface.

We sampled 7200 distinct grammars from the RLM at $T = 27, \epsilon_s/(NT) = 0.01$ and varying N and ϵ_d. Since ϵ_s is small, there is already simple structure at the surface; we explore how deep structure emerges as N and ϵ_d are varied (see [8]) [1].

The information content of a grammar $\mathcal{G}$ is naturally encoded by the Shannon entropy rate of observed strings, $S_s(\mathcal{G}) = \langle \log 1/\mathbb{P}(o|\mathcal{G})\rangle$. For CFGs we can also consider the entropy rate of deep configurations, $S_d(\mathcal{G}) = \langle \log 1/\mathbb{P}(\sigma|\mathcal{G})\rangle$. In both cases the ensemble average is taken with the actual probability of occurrence, $\mathbb{P}(o|\mathcal{G})$ for S_s, and $\mathbb{P}(\sigma|\mathcal{G})$ for S_d. The grammar averages $\overline{S}_s$ and $\overline{S}_d$ are shown in Fig. 1bc.

Random Language Model (cont'd)

The dependence on ϵ_d is striking: for $\epsilon_d \gtrsim N^3/\log^2 N$, both S_s and S_d are flat. In this regime, $S_d \approx \log N$, indicating that although configurations strictly follow the rules of a SCFG, deep configurations are nearly indistinguishable from completely random configurations. However, at $\epsilon_d = \epsilon_* \approx N^3/\log^2 N$ there is a pronounced transition, and both entropies begin to drop. This transition corresponds to the emergence of deep structure.

Fig. 2a shows the Zipf plot of deep structure; the Zipf plot for surface structure is similar, but less dramatic (see [8]). We see a sharp change at ϵ_*: for $\epsilon_d > \epsilon_*$, the frequencies of hidden symbols are nearly uniform, while below ϵ_*, the distribution is closer to exponential. The permutation symmetry among hidden symbols is thus spontaneously broken at ϵ_*.

What is the correct order parameter to describe this transition? For each interior rule $a \to bc$ we can define $Q_{abc}(\mathcal{G}) = \langle \delta_{\sigma_{\alpha_1},a}\left(N^2 \delta_{\sigma_{\alpha_2},b}\delta_{\sigma_{\alpha_3},c} - 1\right)\rangle$, averaged over all interior vertices α, and averaged over derivations. Here σ_{α_1} is the head symbol at vertex α, and $\sigma_{\alpha_2}, \sigma_{\alpha_3}$ are the left and right symbols, respectively. Q measures patterns in rule application at each branching of a derivation tree. It is thus an order parameter for deep structure. Upon averaging over grammars in the absence of any fields, the permutation symmetry must be restored: $\overline{Q}_{abc} = q_0 + \delta_{ab}\, q_l + \delta_{ac}\, q_r + \delta_{bc}\, q_h + \delta_{ab}\delta_{ac}\, q_s$. As shown in [8], these components show a transition, but there is significant noise below ϵ_*, despite there being 120 replicas at each point. Evidently, Q_{abc} has large fluctuations below ϵ_*. This suggests a definition $Q_2 \equiv \sum_{a,b,c} Q_{abc}^2$, plotted in Fig 2b. The signal is clear: on the large scale, Q_2 has a scaling form $Q_2 \approx N^3 f(\epsilon_N/\epsilon_*)$, and is small above ϵ_*. The scaling $Q_2 \sim N^3$ suggests that below the transition, all hidden symbols start to carry information in the deep structure.

Discussion

We have shown that the RLM has a transition to deep structure as ϵ_d is lowered. By a scaling analysis of Z (see [8]), we can understand the transition at $\epsilon_d \approx \epsilon_*$. Fix a sentence of length ℓ and define the energy of a parse as $-\log W - \log W_{tree}$, to be compared with its entropy S. We find that for $\epsilon_d \gg \epsilon_*$, the energy of a parse is unimportant, and the grammar is thus irrelevant: the language produced by the WCFG must then be indistinguishable from random sequences, as found empirically (Fig1bc). In contrast, for $\epsilon_d \ll \epsilon_*$, the language reflects those sequences with high intrinsic weight, and their entropy is less important.

Around 6000 languages are spoken around the world [1]; given fractured input, how does a child come to learn the precise syntax of one of these many languages? [2] One scenario for learning is the Principles and Parameters theory [3]. This posits that the child is biologically endowed with a general class of grammars, the 'principles,' and by exposure to one particular language, fixes its syntax by setting some number of parameters, assumed to be binary. For example, the head-directionality parameter controls whether a language is head-initial, like English, in which verbs come before objects, or head-final, like Japanese, in which verbs come after objects. A vast effort has been devoted to mapping out the possible parameters of human languages [1, 13]. The richness of the structure has been used as criticism of the approach [11]: if the child needs to set a huge number of parameters, then the theory appears at odds with 'poverty of the stimulus' arguments in favor of innate linguistic knowledge.

The RLM can shed some light on this debate[2]. Following experimental work [15], we picture the learning process as follows. Initially, the child does not know the rules of the grammar, so it begins with some small number of hidden symbols and assigns uniform values to the weights M and O. To learn is to increase the likelihood of the grammar by adjusting the weights. New hidden symbols are added when new data cannot be acceptably parsed. As weights are driven away from uniform values, the temperatures ϵ_d and ϵ_s decrease. Eventually the transition to deep structure is encountered, and the grammar begins to carry information. A crucial point is that the child's environment acts as a field on this likelihood-ascent. As temperature is lowered, the RLM is expected to spontaneously break any symmetries present: for example, a left-right symmetry breaking could correspond to setting the head directionality parameter.

Although this description is schematic, we insist that the various symmetry-breaking transitions that could give rise to parameters are already implicit in the definition of the model, without any detailed additional information needed to be supplied. If the RLM can be solved, by which we mean that the partition function Z can be computed, then the series of symmetry-breaking transitions that occur in the presence of a field can be inferred: thus the structure of the syntax of human languages could be deduced. This is a tantalizing goal for future work.