Dispatching “Poverty of the Stimulus”: Did Chatbot just learn Context Free Grammar from positive examples?

Stephen José Hanson

The New York Times recently published an interview with Noam Chomsky, who called GPT-3, the LLMs, or those cute chatbots: “the banality of evil”! First, this seems like an odd over-reaction to a neural network that makes lots of semantic errors. (the reference to the Arendt book on Eichman is even more puzzling in this context) Frankly this seems like something we need to lay out so we can understand the context and why this matters.

First, let’s set the table. Noam Chomsky appeared in the 1950s, arguing that language was complex (in fact, specifically not a finite state machine-FSM, George Miller and Chomsky had previously in 1956 –”Finitary models of language users”– laid out a hierarchy of grammar complexity), and that there was no way to learn it from data. Often called the Poverty of the Stimulus. These two “facts”, were not really testable, but various observations like children overgeneralizing past tense “I goed to the park” even though parents would correct their little “chatbot” -”no –dear, its I went to the park”, and the child would answer, yes I goed to the park and swinged on the swings!” It appeared that children were resistant to feedback, and that some ballistic process had been initiated developmentally.

These casual observations were buttressed by learning theorems (or anti-learning theorems) that indicated that language grammars could not be identified from positive examples only (Gold, 1961) but could from negative cases. Osherson and Weinstein, in the 90s (Systems that learn) further added to this emerging view, based on showing booleans of various complexity were also not learnable from positive examples. In effect, the creation of the idea that DATA from the language speaking community could not provide enough or complex enough or diverse enough data to learn human language was the origin of the Poverty of the Stimulus. Hence, certainly a neural network or any computational machine learning based system exposed to positive examples simply could not learn language or much of anything very interesting (Minsky and Papert 1969, of course, made a fine point on this!) . Gold’s theorem also had a ‘chilling effect’ on the psychological community when it

first became widely known and literally halted language learning research for decades (and was instrumental in the demise of behaviorism and Skinner's "Verbal Behavior" which Chomsky had dismantled in a review). The Poverty of the stimulus had become the de facto explanation for the lack of language learnability.

Linguists like Chomsky were now free to speculate on the design and structure of the language module as my colleague at Rutgers, Jerry Fodor, would refer to it. This functional modularity was first posited by Chomsky which initiated the speculation that ALL cognitive functions were modular (cognitive functionalism) which by the 80s Fodor and Chomsky had put forth a more principled program about modularity (including horizontal and vertical modularity!) and most importantly that modules were impenetrable, except from those specially trained linguists using a sentential contrast grammaticality test like a divining rod, hunting for the Universal Grammar..searching for the UG!

The Chomsky species of linguistic theory exploded, partly due to Chomsky himself, who was particularly good at constructing a linguistic argument but even better at destroying yours . He trained a generation of brilliant linguists all over the world which carried the UG far and wide, essentially influencing cognitive science, computational neuroscience, cognitive neuroscience and especially developmental psychology. Enter Steve Pinker.

But before we go there, we need to clear and set the table one more time. In 1987, 6000 copies of an edited book on Parallel Distributed Processing (PDP) sold in one hour at the 1987 AAAI conference, where Geoff Hinton in his usual clear, didactic manner layed out the details to 1000s of AI researchers, of a new algorithm (primarily due to Dave Rumelhart) that would be able to–tabula rasa– learn complex structure from data. Backpropagation, the implausible biological engine underlying the recent resurgence of neural networks, enabled a full frontal attack on Poverty of the Stimulus (POS). One early example of the broadside was NETALK. Charlie Rosenberg and Terry Sejnowski spent part of a summer in Baltimore failing to learn a relatively simple mapping of text to sound, using a Boltzmann machine, ironically in some ways more biologically plausible than Backpropagation. And later when Dave Rumelhart happened to drop by, he suggested they give it a try. They did, and Charlie, who was a graduate student at Princeton with George Miller, brought the simulation at the end of summer, and demonstrated what it could do. I was a visiting Faculty in the

Cognitive Science Center at Princeton and a MTS at Bell Labs–working with George on some language learning experiments. Charlie asked to borrow my DECALK (a 1970s text to speech system built by DEC) and simply ran the phonemic output of NETALK through one of the speech options of the DECTALK, a child's voice! Fait accompli! George in some ironic impulse completed the POS broadside by playing the DECTALK demo to the NBC Science Editor visiting with Mike Gazzaniga. It thus appeared on the TODAY show, Newsweek, and became the poster-child for learning language from nothing. No modularity, no language module, no UG.

Back to Pinker. So, the plot thickens. McClelland and Rumelhart provided a simulation of learning the past-tense using backpropagation aiming at the heart of POS with this blunt little dagger. Steve Pinker and Alan Prince (a Chomsky trained Linguist at Rutgers), deconstructed the M&R past-tense learning model showing it was making odd past-tense predictions–essentially errors and doing stupid things (sound familiar?). There were many such skirmishes, one notable attack on POS was a lovely book by Jeff Elman and Liz Bates on "RETHINKING INNATENESS". The lines were drawn, and debates were had, but neither side gave an inch. In the meantime, Backpropagation fell on hard times. There wasn't enough curated data and there wasn't enough complexity in a single-hidden layer network to scale up and learn complex behavior. Although many of us were trying to build what we hoped would end up being GTP_zero, unfortunately, the best we got were auto-encoders that would repeat what they saw in the input and sometimes even produce compelling grammatical behavior (Hanson & Kegl, 1987). In terms of recurrent networks, there have always been interesting results both mathematically (Sieglemen & Sontag, 1995) and simulation (Giles et al,1990) where an RNN could learn from examples, a FSM with 1000s of states. But no GPT in sight.

Another tactic that Chomsky took in buffering POS from critique was specifying the appropriate way to test the UG theory. Besides POS and grammar complexity, and special grammaticality methodologies to test the theory there was the distinction he drew between performance and competence. This was first proposed in "Aspects of Theory Syntax" (1965). So linguistic competence was defined as the system of unconscious knowledge that one knows when they know a language. So its implicit knowledge (can't be verbalized) and it is the UG. Performance on the other hand was all other factors that allow one to use one's language in practice. So it is filled with error and noise and irrelevant details to the underlying UG, but is based on the UG. This is convenient on some level

as it makes the theory fundamentally untestable unless you were equipped with one of the Chomsky trained linguists and their finely honed grammaticality judgments. In 1963 George Miller published a paper in the Quarterly Journal of Psychology entitled: "**A Chronometric Study of Some Relations between Sentences".** George did a clever study looking at the speed with which human subjects would transform a passive sentence to active and or a positive to a negative and also compared passive negative to active positive etc. There were 100s of ms differences between the various cases the AP→ PN took the longest. If a transfomational grammar was at play in human judgement, then this was a process/computational experimental analysis of these transformations! Miller and Chomsky had started the Cognitive Studies Center at Harvard years before and collaborated on the idea of psycholinguistics. To Miller's surprise, Chomsky was no longer interested in psycholinguistics or psychological experiments of any sort, and informed George thusly. The performance/competence distinction effectively neutralized any experimental work informing linguistics thereafter, and George told me this story with his disappointment about using psycholinguistics to inform linguistics about UG. Chomsky simply cut that path off. UG had basically with POS, C/P and complexity become dependent on its own internal consistency, since it would have no outside data. Oh, except for the specially trained diviners with their magical divining rods.

So Let's talk about GPT-3 and the recent release of GTP-4 this week. First, we don't know how they work! *Period*. No, they are not a smished together globs of averaged wiki pages that have been "plagerized" by GPT (as if a GPT would be too stupid to reason, but smart enough to steal text and use it in the correct context!). Much of the speculation on CONNECTIONIST LIST, is residue left over from the 1980s and 90s, when critics claimed NN were just look up tables–not look-up tables…NN generalize to unseen data. Or averaging data together to make prototypes. Not averaging– no protoypes. Or will experience catastrophic forgetting. Not true. Even 1 layer hidden networks didn't experience CF unless you trained them with overlapping tasks, and similar input codes, in effect making it impossible to learn without CF. Finally GPTs are not "pattern associators" — this is an incoherent claim, as it is not specific enough to make sense of. But to be emphatically clear… We don't know what is happening! There is no simple explanation at this point what they are doing. There is some interesting research recently showing that GPTs are creating local structures (circuits?) that have some functional import (OPENAI). But are these "contexts" or some other type

of referential structure that are being built by GTP, in order to “communicate”. I don’t know, but I bet the answers will be transformative. By the way I don’t know if they are reasoning, or conscious or would even make good psychiatrists. But I do believe we all buried the lead. Let’s dig it up:

Chatbots are passing human grammaticality tests every day.

JUST FROM LEARNING on LANGUAGE DATA (albeit a huge amount) WITH AN UNREASONABLY AMOUNT of WEIGHTS (billions),GTP-3.5 is learning grammar. But what grammar? It is clearly more complex than a 100M state *FSM* (maybe) but maybe its *context sensitive* in the Chomsky/Miller hierarchy, but certainly must be *context free*.

GPTs are clearly learning grammar. A good time for linguists to pull out their divining rods and test it! At the very least it now appears POS is false. The UG story is badly damaged and the cult of Chomsky may be slipping away in history.