Chomsky vs Norvig and the Competence/Performance Distinction

In 2012 Noam Chomsky raised concerns about the current approach to AI at the "Brains, Minds, and Machines" symposium at MIT. His skepticism about the explanatory power of the widespread statistical methods used in modeling intelligence sparked an elaborate rebuttal by Peter Norvig, Google's director of research. While Chomsky and Norvig differ in their views on the practical applications of capturing a cognitive system, this discord seems to be informed by a fundamental divide that lies in their theoretical approaches: the competence/performance distinction. In this paper, I will present Chomsky's arguments for a theory-driven competence-based model to capture language, and Norvig's arguments against this model in favor of a data-driven performance-based model. Finally, I will present an adaptation of Searle's Chinese Room scenario to argue that current probabilistic models are insufficient in creating a theory of language that accurately represents natural human language.

Chomsky has famously distinguished between competence, a speaker's internal knowledge, and performance, an instance of that knowledge. While one's linguistic competence is an idealized capacity that is universal to all language-speakers' *psychology*, performance is external and reflects the individual's *behavior*. This behavior (utterance) is produced, in part, by the linguistic rules that one's competence underlies. As such, a competence-based approach to a linguistic system would emphasize the internal mechanisms that drive language, while a performance-based approach would seek to classify the external language data to make generalizations about the system.

Chomsky is a marked advocate of competence-based models to study, interpret, and generate language. He argues that in order to develop a theory of language, we need to understand and explain the language system itself. The first step is to determine what tasks the system is performing. In the brain, these tend to be computational tasks that receive input, transform it, and output the result. This seems to be the case for language as well: a speaker receives input, which may be his own thoughts/linguistic intent, processes that input, and outputs it as an instance of speech. Another component to understanding the language system is identifying fundamental principles that underlie its core properties. According to Chomsky, the fundamental principle of language is Universal Grammar, the theory that all children have genetically endowed language faculty that allow them to acquire any natural language, and that this faculty contains an innate set of internal rules that govern

the structure of language. Two core properties of language are productivity, the ability to generate novel sentences, and discrete infinity, or unlimited productivity by finite means (finite number of words). As such, to propose a theory of language, we must identify and understand the structural linguistic rules that afford humans the ability to generate and comprehend infinite novel sentences. Once we have ascertained these rules (competence), we may then turn to language data to confirm the theory, and start addressing the thousands of intervening variables that may affect the process of externalizing linguistic output. Since we can now understand the mechanism of language, we'll use these variables, which are secondary to competence, to generate better predictions of the performance output, which have instantiated the rules of competence.

In studying competence, Chomsky underscores the importance of a theory's explanatory power over its predictive/simulatory ability. He argues that current AI models that are performance-driven and "analyze masses of data" are not conducive to gleaning insight about the computational systems themselves. Rather, they will lead to better approximations of future performance outcomes without understanding the underlying principles which determined how those outcomes came to be. In essence, Chomsky has been an outspoken opponent against statistical modeling culture which prioritize simulation at the expense of explanation, and is "unlikely to yield general principles about the nature of cognition".

Conversely, Norvig is an advocate for performance-based models of language in which the theory is driven by data. He argues that while the types of intuition outlined in Chomsky's model have been an important component of scientific methods, observation - namely the "accumulation of facts" - is the dominant model of science today. As a researcher intent on developing a Natural Language Processor, Norvig's experience with linguistic data has revealed to him that language is "messier" than an idealized set of rules would predict. When we perceive language to be interpreted, the input data is accompanied by noise that the language system must parse in order to receive the true signal. He argues that language processing is a stochastic phenomena as opposed to deterministic, in that it is subject to random change and variation that is not accounted for by stringent rules. As such, he believes that a probabilistic model, which does not make categorical judgements but rather predicts likely patterns from the noisy data best represents linguistic facts, and statistical models that are trained based on input from prior data, best make sense of those facts.

Norvig then provides a critique of Chomsky's competence-based model in three domains where a performance-based model may provide a better analysis. First, within sentences that require both grammaticality judgements as well as probabilistic judgments. In *Syntactic Structures*, Chomsky introduced a pair of paradigm sentences to illustrate the productivity of language. He argued that in cases where a system had to make a grammaticality judgment based on two novel sentences (a) and (b), one which used internal structural relations would be able to differentiate between the two, but a finite-state model could not make a judgement. If the model was trained on the linear order of words in a sentence, and that sentence or its constituent pieces did not appear in the corpus, then the model would be unable to make a grammaticality judgment that was dependent on structure.

(a) Colorless green ideas sleep furiously

(b) *Furiously sleep ideas green colorless

However, Norvig challenges this claim, arguing that not only can a probabilistic model detect which sentence is grammatical, but it can also determine the likelihood of the sentence. In fact, when trained over the Google Books corpus, a probabilistic model found that sentence (a) was 10,000 times more probable than (b), and that it was improbable compared to other sentences that appear more regularly in the corpus (ex. "Effective green products sell well"). Because a performance-based model approaches grammaticality as a graded continuum given a vast input of data, it can make these probabilistic judgements, whereas a competence-based model which approaches grammaticality categorically may only differentiate whether the sentence is grammatical, and can say nothing about its likelihood to be uttered.

Second, Chomsky's categorical model cannot account for the variation of language found in the linguistic data of natural languages. Chomsky's competence model posits that language learning is afforded by acquiring binary values for particular grammatical parameters. Languages adhere to the core rules of the grammar, and differentiate the grammatical from the ungrammatical based on whether these features are turned "on" or "off". For example, transitivity, or the ability for a verb to take a direct object. According to the grammar, a verb specified as intransitive, such as "quake", can never take a noun as a direct object following it, whereas a transitive verb such as "washed" must be followed by an object. However, Norvig counters this system by introducing examples such as (c) from the corpus that contradict this point, where intransitive verbs *do* take objects. A categorical model cannot account for these instances of corpus data that violate the grammar, so it must either

dismiss these examples without the theoretical ground to do so, or expand the model; both of which dilute the model, decreasing its deterministic and explanatory power.

(c) *It quaked her bowels

A probabilistic model, however, may account for both uses of the verb, both transitive and intransitive, and predict the likelihood of both being used in natural language.

Finally, Norvig argues that Chomsky's disregard for observational data in developing his theory leads to one that can neither predict nor explain what language is. He states that because Chomsky's simplistic, internalist theory which provides categorical judgements cannot account for the variation in natural language, it is an Platonistic rather than realistic model of language. By discarding data that doesn't fit into his formalism, Chomsky is moving further from the empirical science that Norvig believes linguistics to be, and closer to a mathematical idealism that can explain how language *should* behave, not how it actually does. Furthermore, Norvig believes that the goal of a language model is not to glean insight into the deep "whys" that underlie the system's processing capabilities, but the "hows" that underlie its mechanism for future predictive capabilities. As a result, Norvig adopts a position that forgoes a system with a form that models nature's functions. He explains that if the goal of science is to collect data, make sense of it, and use these findings to generalize to future predictions, then the mechanism by which a model completes this secondary the accuracy of its outputs. A model with predictive power need not match the true underlying nature of the system it's simulating, especially if the alternative is a competence-based model that does not accurately reflect reality.

These three criteria are all better suited by a probabilistic model, and further bolster Norvig's performance-based standpoint. By analyzing the corpus of observable language data, we can begin to understand its "messy," ambiguous nature. A language system is contingent on the outcomes of complex processes, and by virtue of its contingency, it cannot be accounted for by a categorical model that makes deterministic judgements in isolation, but by probabilistic analysis that may holistically account for the vast variation that co-occurs with it.

In the following section, I will present an adapted version of Searle's Chinese Room in order to concretize the theoretical and methodological divide between Chomsky and Norvig.

Imagine there is a room. Outside the room, there is a Chinese-speaker who slips written messages in Chinese underneath the door to the room. Inside the room, there is an English-speaker

who does not understand a word of Chinese. However, in the room with him, he has a manual containing everything he needs to interpret the message, and produce a response, which he then slips out the door. The goal of the speaker is to receive the input in Chinese, and output a proper response that is comprehensible by the speaker outside the room.

According to Chomsky's model, the manual that allows the speaker to interpret and transform the language is their **competence**. It contains the structural principles of UG that afford generativity. The English-speaker would apply these rules to break down the Chinese input into its respective constituents, and then build up a structure that adheres to these rules and the parameters specified by the language. The response that they subsequently send through the door is an instance of this competence being used, and is considered a performance. It is important to note that the manual of rules is not a look-up table of every possible value of the language, but a set of constraints that guide interpretation and generation. The goal is to understand how these constraints and parameters apply, and why they produce the performances that ensue.

According to Norving's model, the instruction manual is the entire corpus of linguistic **performances** ever produced, and the English-speaker in the room is analogous to a supercomputer that must process this data. The goal of the speaker is to make generalizations from this massive accumulation of noisy data, and apply these generalizations to permute the input which creates a likely output. In this case, the externally supplied information fuels the predictions and hypotheses that are generalized from it.

Chomsky would argue that while his competence-based model is "doing" language by making use of its internal knowledge to generate novel outputs, Norvig's model is merely simulating language without addressing the underlying principles that presuppose its performances. As such, this model is limited to judgments based on linear strings rather than structure, and cannot account for situations in which it is presented structurally permissible sentences that do not appear in the corpus. For example, (d) below is a grammatical sentence in English, in that adheres to the structural rules of sentence formation. However, this sentence and its structure are highly improbable, and prior to it being published as a paradigm example, likely would not have appeared anywhere else. If a sentence of this type was slipped under the door, what kinds of information could the speaker elicit from the corpus data that would help interpret it?

(d) Tigers tigers tigers fight fight fight

If the model cannot resolve the meaning of this sentence, (which it can't without structural rules), then this undermines the interpretive goal of Norvig's model. Furthermore, if it cannot even interpret the sentence, then it cannot generate an output, novel or not. A flaw of this type then begs the question of whether these types of structural inputs ever make it into the room. In this case, Chomsky would believe the Chinese-speaker outside of the room to be AI researchers. When they ask questions to the machines, they are only asking ones within the scope of the machine's capabilities, namely questions that have appeared in the corpus. As a result, the output answers, while appearing to be accurate, serve as their own self-reinforcing concept of success. If you only ask a machine the kinds of questions you know it can answer, then you will never glean insight into the chinks in its armor - the structural analyses that it cannot compute without competence. As a result of this self-confirming evidence, the algorithms that performance-based systems developed in response to massive data sets have been optimized over time, but still do not yield explanatory insight into the mechanisms of the language system itself. Chomsky argues that technological change is inevitable, and without an understanding of the fundamental language principles which will endure throughout time, the progress made in the algorithmic dimension is temporal, and surely will not outlast true understanding.

Finally, Chomsky questions the necessity of the performance-based corpus manual itself. Norvig posits that a collection of the entirety of spoken language is the only way to account for the variation in natural language. Without accounting for this variation in our subsequent theories, we are missing the point, and limiting the interpretative and predictive capabilities of the system . However, Chomsky argues the opposite. He postulates whether there is any point in trying to make sense of noisy data. If language is constraint-based, and we are endowed with a manual of these constraints, then to interpret and generate language, we must pick out the properties of our input that adhere to these rules from the peripheral noise. For example, babies learning language are inundated with an overabundance of stimuli and massive noise. If they were conducting statistical analyses on their entire corpus of input, then it would not deduce the innate properties that distinguish signal from noise. However, in a competence-based system, if you feed the babies (or machines in the case above) "bad" input, they can still pick out the good rules.

To summarize, the goal of the system is not to perform statistical analyses on the entirety of data one is exposed to, but to pick out computational properties and units that combine to produce the

structure of the utterances. The problem with Norvig's model is that it's solely reflecting the output, not the idealized internal rules that Chomsky posits, and if it is incapable of handling the output it must reflect, then the system will malfunction. Conversely, a competence-based system can resolve ambiguous input through the applying innate structural knowledge to generate an externalized output. If a system seeks to represent the external language without understanding the underlying structure, then it is limited in both its explanatory and predictive capabilities. It cannot make grammaticality judgements necessary for interpretation, nor can it combine constituents to generate a response.

In conclusion, while a data performance-based system may hold predictive power, this model cannot capture or explain the productivity, structural organization, and constraints evidenced in natural languages. Conversely, while a theory-driven competence-based system may not be optimized for the predictive capabilities of the previous model, its explanatory power and an internal rules afford understanding of the underlying structures, after which the variation in observational data may serve as supplementary evidence, and can be optimized to bolster predictions of performance.

References

Katz, Yarden. “Noam Chomsky on Where Artificial Intelligence Went Wrong.” *The Atlantic*, Atlantic Media Company, 7 Dec. 2017, https://www.theatlantic.com/technology/archive/2012/11/noam-chomsky-on-where-artificial-intelligence-went-wrong/261637/#:~:targetText=Chomsky critiqued the field of,more modern, computationally sophisticated form.&targetText=But as far as a,more harshly, kind of shallow.

Norvig, Peter. “On Chomsky and the Two Cultures of Statistical Learning.” *Norvig.com*, http://norvig.com/chomsky.html.